Apunte de Regresión Lineal

María Eugenia Szretter Noste
Carrera de Especialización en Estadística
para Ciencias de la Salud
Facultad de Ciencias Exactas y Naturales,
Universidad de Buenos Aires

Agosto - Octubre de 2017

Índice

Prefacio

Las notas de regresión lineal que componen estas páginas fueron escritas como material teórico y práctico para el curso Regresión Lineal de la Carrera de Especialización en Estadística para Ciencias de la Salud, que se dicta en la Facultad de Ciencias Exactas y Naturales, de la Universidad de Buenos Aires que tuve la alegría de dar durante algo más de dos meses, en 2011 y luego cada dos años hasta 2017. Presuponen un conocimiento estadístico obtenido en un curso básico y hacen énfasis en un enfoque aplicado de la regresión lineal, para un público que viene, en general, de las ciencias médicas o biológicas. La información sigue un programa estándar en el tema: correlación, regresión lineal simple y regresión lineal múltiple y representa una primera introducción al tema. La idea es hacer un énfasis en los modelos y la interpretaciones, sin perder (del todo) el entusiasmo en el camino. En esa dirección, estas notas buscan presentar al modelo lineal como el primer modelo estadístico a estudiar en detalle, e intenta mostrar cuáles de las herramientas presentadas se generalizan a otros modelos estadísticos. En cada capítulo, además, se incluyen una serie de ejercicios que (espero) complementen el aprendizaje.

Los gráficos y las salidas que acompañan las notas fueron realizados usando el paquete `R`, R Core Team [2015]. El resto de las figuras fueron extraidas de varios buenos textos disponibles sobre el tema (y debidamente citados). Quizá la mejor hoja de estas notas sea la bibliografía. Esta versión 2017 de las notas incorpora además de varias correcciones, la introducción de los comandos de `R` en el texto, así como varios nuevos ejercicios y los scripts en `R` para resolverlos.

Finalmente quiero agradecer a varios colegas las conversaciones y opiniones sobre los temas que aparecen a continuación, que ayudaron a dar (esta) forma a estas notas, en especial a Liliana Orellana y a Andrés Farall.

Este material puede descargarse de la web de la siguiente dirección
http://mate.dm.uba.ar/~meszre/apunte_regresion_lineal_szretter.pdf
En la misma dirección, hay una carpeta con todos los archivos de datos mencionados en el texto, o necesarios para los ejercicios. Dicha carpeta contiene también scripts en `R` que resuelven los ejercicios. La dirección de la carpeta es
`http://mate.dm.uba.ar/~meszre/datos_regresion`

1. Correlación

La regresión lineal, de la que tratan estas notas, se ocupa de investigar la relación entre dos o más variables continuas. En esta sección, comenzaremos tratando de describir el vínculo observado y luego nos sofisticaremos resumiendo en un valor numérico nuestra conclusión.

¿Con qué datos contamos para llevar a cabo un análisis? Disponemos de n observaciones de dos variables aleatorias medidas en los mismos individuos, como describimos en la Tabla 1.

Tabla 1: Observaciones a nuestra disposición. Aquí X_1 quiere decir, la variable X medida en el individuo 1, etc.

Individuo	Variable X	Variable Y
1	X_1	Y_1
2	X_2	Y_2
$\vdots$	$\vdots$	$\vdots$
n	X_n	Y_n

En estas notas, estamos pensando en que medimos ambas variables en la misma unidad: puede tratarse de un individuo, un país, un animal, una escuela, etc. Comencemos con un ejemplo.

1.1. Gráficos de dispersión (o scatter plots)

Ejemplo 1.1 *Queremos investigar la relación entre el pocentaje de niños que ha sido vacunado contra tres enfermedades infecciosas: difteria, pertusis (tos convulsa) y tétanos (DPT, que se suele denominar, triple bacteriana) en un cierto país y la correspondiente tasa de mortalidad infantil para niños menores a cinco años. El Fondo de las Naciones Unidas para la Infancia considera a la tasa de mortalidad infantil para niños menores a cinco años como uno de los indicadores más importantes del nivel de bienestar de una población infantil. Datos publicados en United Nations Children's Fund,* ***The State of the World's Children 1994****, New York: Oxford University Press. Y tratados en el libro Pagano, Gauvreau, y Pagano [2000], Capítulo 17.*

Los datos para 20 países, del año 1992, se muestran en la Tabla 2. Si X representa el porcentaje de niños vacunados a la edad de un año, e Y representa la tasa de mortalidad infantil de niños menores de 5 años, tenemos una pareja de resultados (X_i, Y_i) para cada país en la muestra.

¿Cómo se lee la información desplegada en la Tabla 2? Por ejemplo, para Bolivia $X_1 = 77{,}0$, es decir, en el año 1992, un 77 % de los niños menores de un año

Tabla 2: Datos para 20 países en los que se midieron dos variables, X : porcentaje de niños vacunados a la edad de un año en cada país, Y : es la tasa de mortalidad infantil de niños menores de 5 años en cada país. Archivo: `paises.txt`.

País	Porcentaje vacunado	Tasa de mortalidad menor a 5 años
Bolivia	77,0	118,0
Brasil	69,0	65,0
Camboya	32,0	184,0
Canadá	85,0	8,0
China	94,0	43,0
República Checa	99,0	12,0
Egipto	89,0	55,0
Etiopía	13,0	208,0
Finlandia	95,0	7,0
Francia	95,0	9,0
Grecia	54,0	9,0
India	89,0	124,0
Italia	95,0	10,0
Japón	87,0	6,0
México	91,0	33,0
Polonia	98,0	16,0
Federación Rusa	73,0	32,0
Senegal	47,0	145,0
Turquía	76,0	87,0
Reino Unido	90,0	9,0

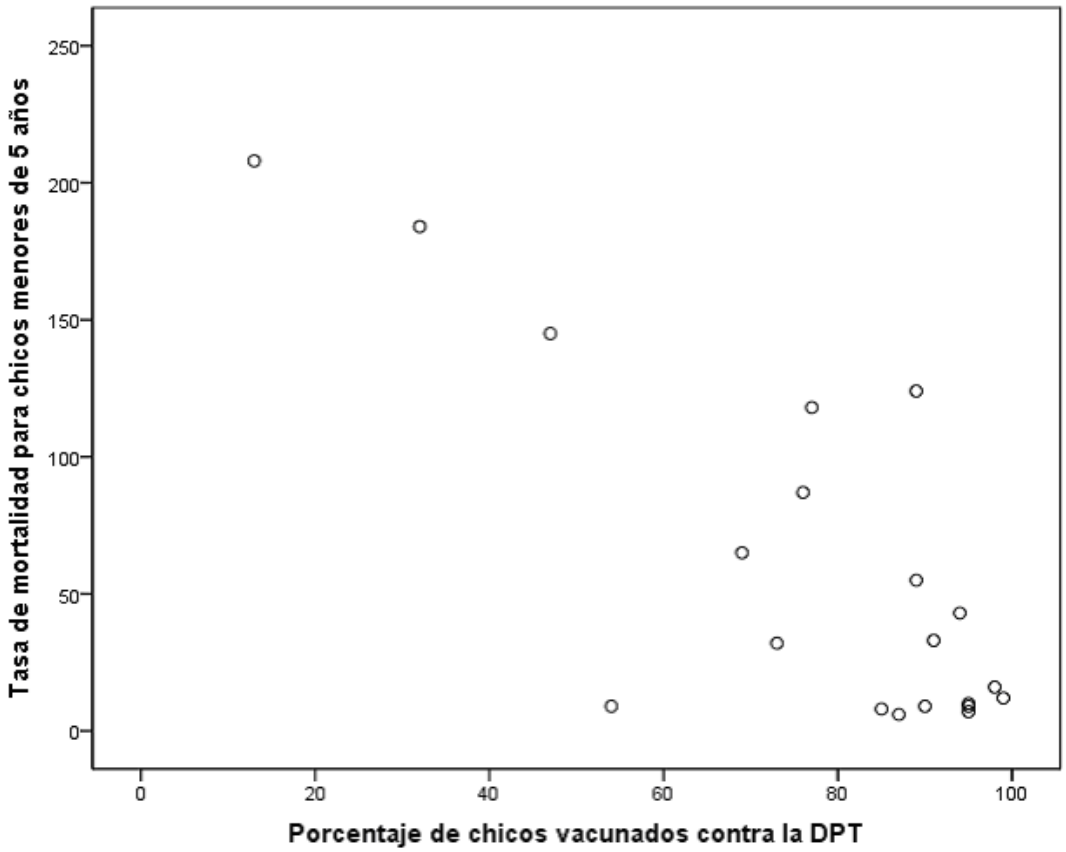
estaban vacunados contra la DPT y (en el año 1992) 118 niños menores de 5 años murieron por cada 1000 niños nacidos vivos.

¿Cómo puede visualizarse esta información? La forma más sencilla es mediante un gráfico de dispersión (o scatter plot). En un scatter plot se ubican los resultados de una variable (X) en el eje horizontal y los de la otra variable (Y) en el eje vertical. Cada punto en el gráfico representa una observación (X_i, Y_i).

En este tipo de gráfico se pierde la información del individuo (paciente o país), y aunque si hubiera pocos puntos se los podrían rotular, esencialmente esta información no suele estar disponible en un scatter plot. El gráfico de dispersión de los datos de la Tabla 2 puede verse en la Figura 1. Ahí vemos que, por ejemplo, Bolivia está representada por el punto $(77, 118)$.

Usualmente con este gráfico podemos determinar si existe algún tipo de relación

Figura 1: Scatter plot: tasa de mortalidad infantil (menor a 5 años) versus el porcentaje de chicos menores de un año vacunados contra la DPT.

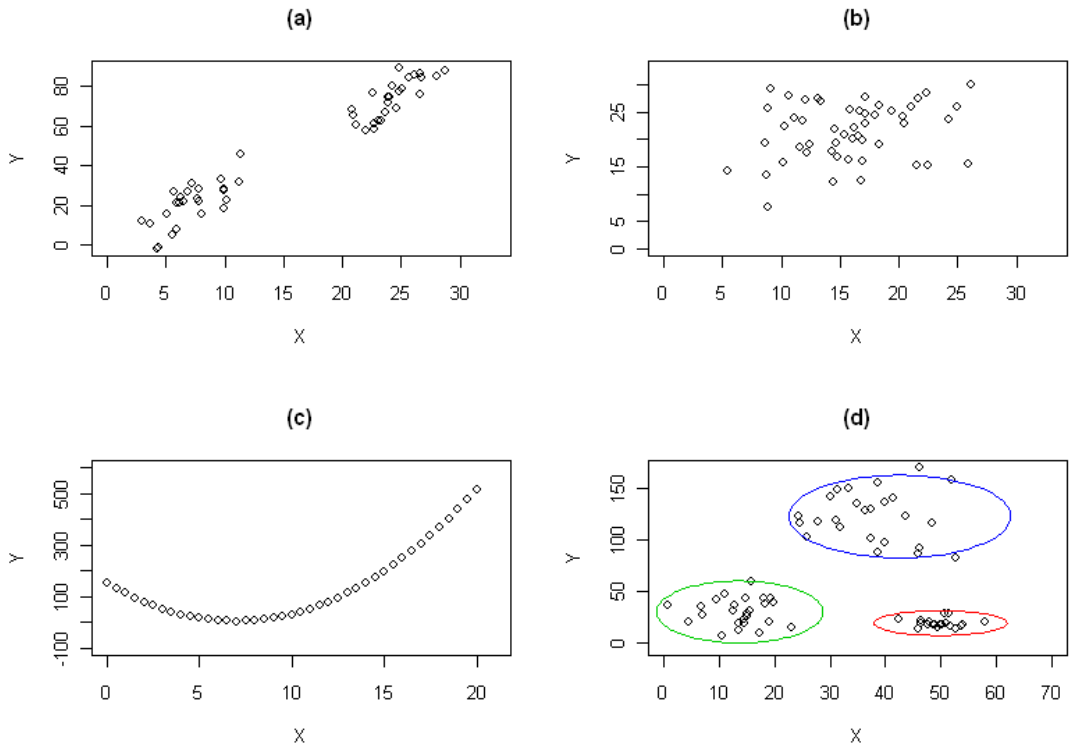


entre X e Y. Para este caso vemos que a medida que aumenta el porcentaje de niños inmunizados, decrece la tasa de mortalidad. ¿Qué otras cosas podríamos observar? En la Figura 2 ilustramos algunas posibilidades, que describimos a continuación.

- **Ausencia de datos.** Puede ser que no hayamos medido ninguna observación cuya variable X se encuentre entre cierto rango de valores (en la Figura 2 (a) por ejemplo, no hay observaciones con coordenada X entre los valores 13 y 21). O que esta combinación entre X e Y no exista, o no se dé biológicamente. Esto indica que la relación que observamos entre las variables graficadas es solamente válida para algunos valores de las variables.

- **No asociación**. ¿Cómo luce un gráfico de dispersión de dos variables que no están asociadas? En la Figura 2 (b) hay un ejemplo. Luce como una nube de

Figura 2: Gráficos de dispersión de cuatro conjuntos de datos diferentes: (a) ausencia de datos; (b) no asociación; (c) vínculo curvilíneo; (d) agrupamientos.



puntos: los valores bajos de X pueden aparecer asociados tanto con valores altos de Y como con valores bajos de Y. Lo mismo para los valores altos de X. Lo mismo para los valores intermedios de X.

- **Vínculo curvilíneo**. Esto aparece cuando los valores de Y se vinculan a los de X por medio de una función. Por ejemplo, si en el eje X graficáramos los valores del tiempo medidos con un cronómetro a intervalos regulares y en el eje Y la posición de un objeto en cada instante de tiempo medido, y si este objeto se moviera siguiendo un movimiento rectilíneo uniformemente variado, observaríamos en el gráfico una función cuadrática, como aparece en la Figura 2 (c). A veces la curva no describe la ubicación de los puntos en la gráfica de manera exacta, sino en forma aproximada (hay errores de medición, por ejemplo, o una relación sólo aproximadamente cuadrática entre las variables).

- **Agrupamientos.** Cuando en el gráfico de dispersión se ven las observaciones separadas en grupos esto puede indicar que hay variables que están faltando incluir en el análisis. Por ejemplo, la Figura 2 (d) puede tratarse de mediciones del largo de pétalo y del sépalo de una flor, de un grupo de flores para las cuales no se ha registrado la variedad. Si habláramos con el biólogo que llevó a cabo las mediciones podríamos encontrar que se trató de tres variedades distintas de flores, que dieron origen a los tres grupos indicados con elipses de colores en el gráfico.

Esencialmente, en el scatter plot nos interesa evaluar

- la *forma* de la relación entre las dos variables
  - lineal
  - no lineal: cuadrática, exponencial, etc.
  - ausencia de relación
- el *sentido* de la asociación, que puede ser
  - asociación creciente: ambas variables aumentan simultáneamente
  - asociación decreciente: cuando una variable aumente, la otra disminuye
  - no asociación
- la *fuerza* de la asociación, esto tiene que ver con la dispersión de los datos. Si el vínculo puede resumirse con una recta o una curva, cuán alejados de dicha recta (o curva) están los datos. Esto suele resumirse cualitativamente: diremos que la asociación es fuerte, moderada o débil, de acuerdo a si los puntos graficados presentan poca, moderada o mucha dispersión de la recta (o curva) que los describe.
- la identificación de unas pocas observaciones que no siguen el patrón general, o de otras características de los mismos como ausencia de datos en algunas regiones, agrupamientos, variabilidad de los datos que depende de una de las variables, etc.

1.1.1. Desventajas de los scatter plots

Los scatter plots son herramientas básicas del estudio de varias variables simultáneas. Sin embargo adolescen de dos problemas, esencialmente.

1. Si hay muchas observaciones todas iguales, en general no se las puede graficar a todas. En el gráfico de dispersión uno no puede notar si hay puntos repetidos en la muestra observada.

2. Sólo se pueden visualizar los vínculos entre dos variables. En gráficos tridimensionales se podrían graficar hasta tres variables, y luego habría que elegir con mucho cuidado el punto de vista del observador para exhibir las características más sobresalientes del gráfico. Cuando el interés está puesto en estudiar varias variables simultáneamente, pueden hacerse varios gráficos de dispersión simultáneos. Es decir, cuando tenemos las variables (X, Y, Z) haremos tres gráficos: Y versus X, Z versus X, y Z versus Y. Los haremos en la Sección 5.1.1.

1.2. Coeficiente de correlación de Pearson

Descriptivamente hablando, en estas notas estaremos interesados en las situaciones donde aparece una relación entre X e Y del estilo de las graficadas en la Figura 3, que pueden globalmente describirse bien a través de una relación lineal (puntos situados más o menos cerca del gráfico de una recta). Cuando los gráficos de dispersión son del estilo de los que aparecen en la Figura 2 (a), (c) ó (d) las técnicas estadísticas que mejor describen este tipo de vínculo entre las variables no se encuadran dentro de la regresión lineal. En la Figura 3 (a) se ve una *asociación positiva* entre las variables, esto quiere decir que a medida que crece X, esencialmente Y crece. En cambio, en la Figura 3 (b) las variables están *negativamente asociadas*: cuando X crece, Y decrece, en general.

1.2.1. Definición del coeficiente de correlación

Para cuantificar el grado de asociación entre X e Y se pueden describir coeficientes. Antes de hacerlo, repasemos los coeficientes poblacionales que asumimos conocidos, ya que se ven en cualquier curso introductorio de probabilidades y estadística

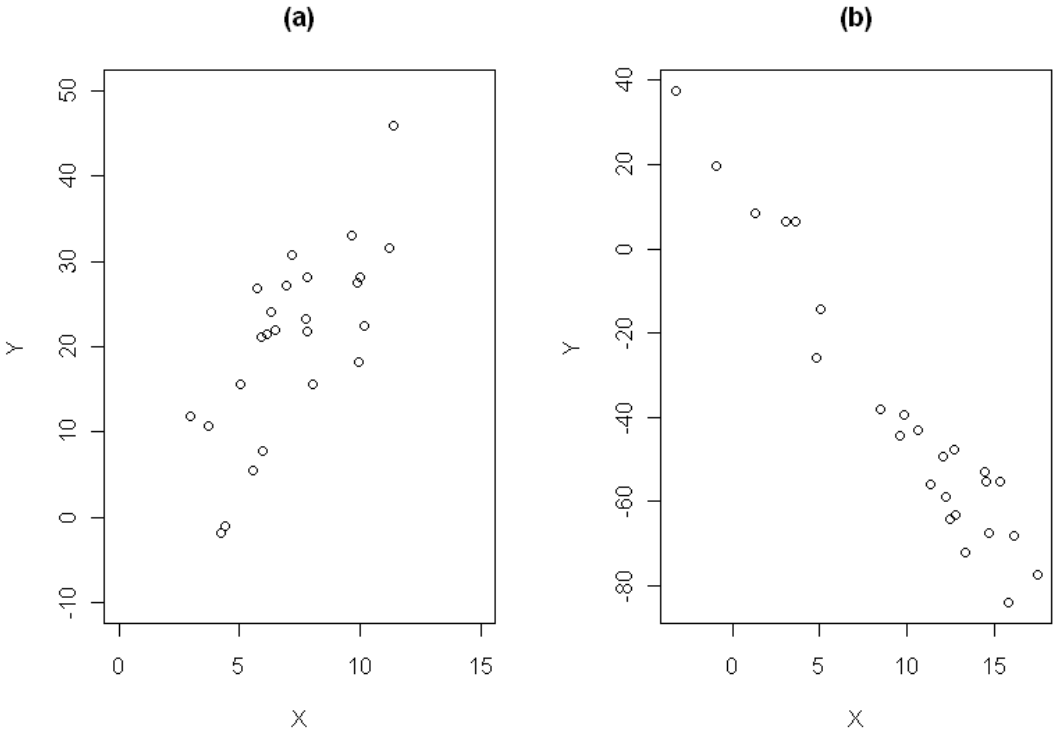
Para una sola variable numérica X podemos definir la **esperanza de X**

$$\mu_X = E(X)$$

como el valor poblacional que describe el centro de la variable. A su vez, tenemos también la **varianza poblacional de X** que es

$$\sigma_X^2 = E\left([X - E(X)]^2\right) = Var(X)$$

Figura 3: Dos conjuntos de datos con asociación lineal entre X e Y : el gráfico (a) muestra asociación lineal positiva, el (b) muestra asociación lineal negativa entre ambas.



que es una medida de la variación de la variable X respecto de su centro dado por $E(X)$. A partir de ella se define el **desvío estándar poblacional de X** por

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{Var(X)},$$

que es una medida de dispersión de la variable X.

¿Cómo estimamos estos valores poblacionales, en general desconocidos, a través de una muestra $X_1, X_2, \ldots, X_n$ de variables independientes con la misma distribución que la variable X? A la media poblacional, μ_X la estimamos por el promedio de las n observaciones disponibles. Llamaremos $\widehat{\mu}_X$ al estimador, es decir, a la función o cuenta que hacemos con las variables $X_1, X_2, \ldots, X_n$ observadas para estimar al número fijo μ_X (en este sentido, $\widehat{\mu}_X$ en realidad es un $\widehat{\mu}_X(X_1, X_2, \ldots, X_n)$), y

escribimos

$$\widehat{\mu}_X = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$\overline{X}_n$ o bien $\overline{X}$ es el *promedio o media muestral*. A la varianza poblacional la estimamos por

$$\widehat{\sigma}_X^2 = S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,$$

que es la *varianza muestral*. Entonces, el desvío estándar poblacional queda estimado por el *desvío estándar muestral*, es decir,

$$\widehat{\sigma}_X = S_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}.$$

Ahora estamos en condiciones de pensar en cómo definir un coeficiente que resuma el vínculo entre dos variables aleatorias X e Y medidas en el mismo individuo. El más utilizado de todos es el que se conoce como *coeficiente de correlación*, que se simboliza con una letra griega *rho*: ρ ó ρ_{XY} y se define por

$$\begin{aligned}\rho_{XY} &= E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] \\ &= \frac{cov\left(X,Y\right)}{\sigma_X\sigma_Y},\end{aligned}$$

o sea, el número promedio a nivel población del producto de X menos su media por Y menos su media divididos por el producto de los desvíos estándares. ¿Cómo estimamos a ρ? A través de r el *coeficiente de correlación de Pearson, o coeficiente de correlación muestral,* dado por

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{S_X \cdot S_Y}.$$

Al numerador, se lo denomina covarianza muestral entre X e Y,

$$\text{covarianza muestral } = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

y el denominador es el producto de los desvíos muestrales de cada muestra por

separado

$$\begin{aligned} S_X &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2} \\ S_Y &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}. \end{aligned}$$

Otra forma de escribir a r es la siguiente

$$r=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\left[\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right]\left[\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2\right]}}.$$

Observemos que el numerador $\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)$ puede ser positivo o negativo, pero el denominador $\sqrt{\left[\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right]\left[\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2\right]}$ siempre es positivo. Luego el signo de r está determinado por el del numerador. Veamos de qué depende.

$$\text{signo de } \left(X_i-\overline{X}\right)=\begin{cases} + & \text{si } X_i \text{ es más grande que } \overline{X} \\ - & \text{si } X_i \text{ es más chico que } \overline{X} \end{cases}$$

y también

$$\text{signo de } \left(Y_i-\overline{Y}\right)=\begin{cases} + & \text{si } Y_i \text{ es más grande que } \overline{Y} \\ - & \text{si } Y_i \text{ es más chico que } \overline{Y} \end{cases}$$

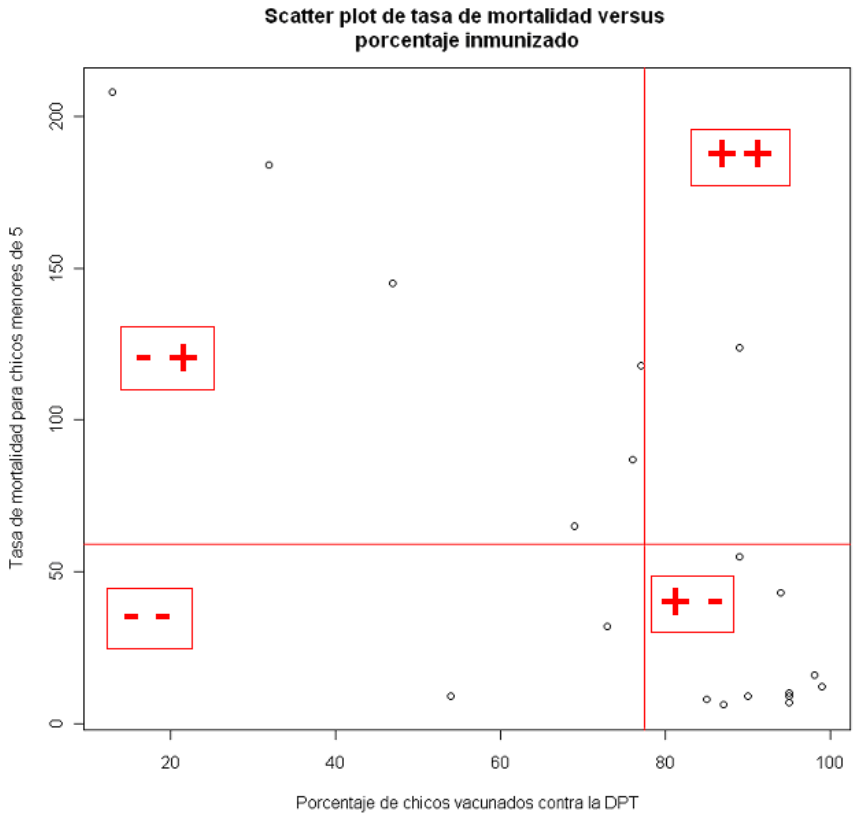
Luego, el

$$\text{signo de } \left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)=\begin{cases} + & \text{si } ++ \text{ ó } -- \\ - & \text{si } +- \text{ ó } -+ \end{cases}$$

Hacemos un scatter plot de las observaciones. Luego ubicamos en el plano el punto $\left(\overline{X},\overline{Y}\right)$. Trazamos una línea vertical que pase por $\overline{X}$ y otra línea horizontal que pase a la altura de $\overline{Y}$. Esto divide al gráfico en cuatro cuadrantes, como puede verse en la Figura 4. Luego, el signo del sumando iésimo de r será positivo, si para el individuo iésimo tanto X_i como Y_i son mayores que su respectivo promedio (es decir, la observación cae en el cuadrante noreste, al que hemos denotado por $++$)

o bien ambos valores son simultáneamente menores que su promedio, es decir, la observación cae en el cuadrante suroeste, que hemos denotado por $--$. En cambio, el sumando iésimo de r será negativo en el caso en el que la observación iésima tenga un valor X_i por encima de su promedio pero la Y_i sea menor que su promedio, o bien la X_i sea menor a su promedio y la Y_i sea mayor a su promedio.

Figura 4: Scatter plot de la tasa de mortalidad versus el porcentaje de niños menores a un año inmunizados, con la recta vertical y horizontal que pasan por $\left(\overline{X}, \overline{Y}\right)$, y los signos de cada sumando que interviene en el cálculo de r.



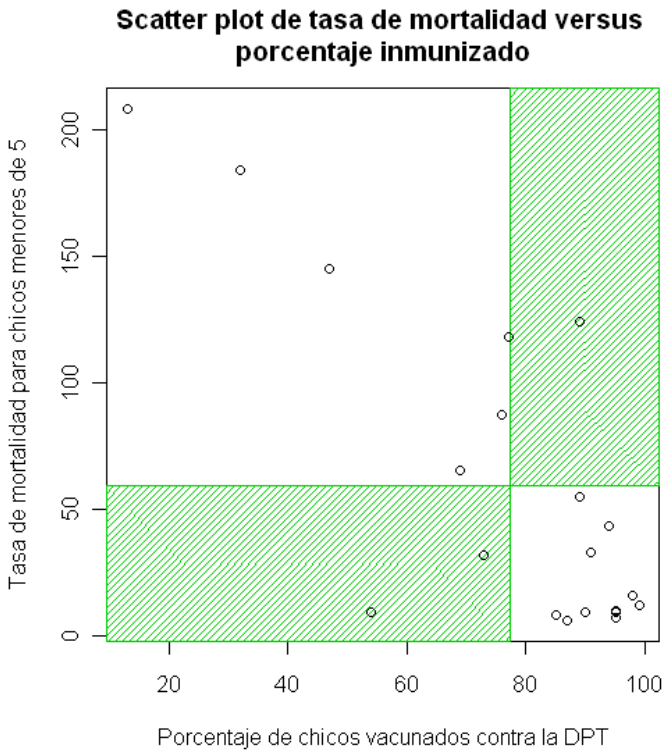
Esto en cuanto a cada sumando en particular. ¿Qué significará el signo de r?

Si r da positivo, será indicio de que la mayoría de las observaciones caen en los cuadrantes noreste (NE) y suroeste (SO), marcados con color en la Figura 5. Es decir, que cuando los valores de las X suelen estar por encima del promedio ocurre, simultáneamente, que los valores de Y también están sobre su promedio. Análogamente, cuando en un individuo el valor de X está por debajo del promedio, lo mismo ocurre con su valor de Y. En general, un valor positivo de r indica que

hay una asociación positiva entre las variables (cuando una crece, la otra también lo hace).

Si r da negativo, en cambio, tenemos una indicación de mayor número de observaciones en los otros cuadrantes marcados con fondo blanco en la Figura 5, y se invierten las situaciones descriptas anteriormente. Es decir, que cuando los valores de las X suelen estar por encima del promedio ocurre, simultáneamente, que los valores de Y están por debajo de su promedio. Análogamente, cuando en un individuo el valor de X está por debajo del promedio, ocurre lo inverso con su valor de Y, que superará a su promedio. En general, un valor negativo de r es indicador de asociación negativa entre las variables (cuando una crece, la otra decrece).

Figura 5: Scatter plot de la tasa de mortalidad versus el porcentaje de niños menores a un año inmunizados, con los cuatro cuadrantes delimitados por $(\overline{X}, \overline{Y})$. Las observaciones que caen en la región coloreada darán sumandos positivos del r.



Ejemplo 1.2 *Veamos qué ocurre en nuestro ejemplo. Calculamos los promedios*

de ambas variables, obtenemos

$$\begin{aligned} \overline{X} &= 77{,}4 \\ \overline{Y} &= 59 \end{aligned}$$

y le superponemos al scatter plot dos líneas rectas, una vertical que corta al eje x en 77,4 *y otra horizontal que corta al eje y en $Y = 59$. Las Figuras 4 y 5 muestran el gráfico de esta situación. Observamos que en los dos cuadrantes coloreados hay muy pocas observaciones (exactamente 3 de un total de 20).*

El coeficiente de correlación muestral en este caso da $-0{,}791$, *un valor negativo, lo cual hubiéramos podido anticipar ya que la mayoría de los términos involucrados en el cálculo de r (17 de los 20 sumandos) serán menores o iguales a cero.*

1.2.2. Propiedades del coeficiente de correlación muestral (y también de ρ)

A continuación damos las propiedades del coeficiente de correlación muestral r, pero estas también son válidas para el coeficiente de correlación poblacional ρ.

1. $-1 \leq r \leq 1$. El valor del coeficiente r está entre 1 y menos 1 porque puede probarse que el denominador es más grande (a lo sumo igual) que el numerador.

2. El valor absoluto de r, $|r|$ mide la fuerza de la asociación lineal entre X e Y, a mayor valor absoluto, hay una asociación lineal más fuerte entre X e Y.

3. El caso particular $r = 0$ indica que no hay asociación lineal entre X e Y.

4. El caso $r = 1$ indica asociación lineal perfecta. O sea que los puntos están ubicados sobre una recta de pendiente (o inclinación) positiva.

5. En el caso $r = -1$ tenemos a los puntos ubicados sobre una recta de pendiente negativa (o sea, decreciente).

6. El signo de r indica que hay asociación positiva entre las variables (si $r > 0$); o asociación negativa entre ellas (si $r < 0$).

7. $r = 0{,}90$ indica que los puntos están ubicados muy cerca de una recta creciente.

8. $r = 0{,}80$ indica que los puntos están cerca, pero no tanto, de una recta creciente. En la Figura 6 se pueden ver distintos grados de correlación, que están comentados más abajo.

9. r no depende de las unidades en que son medidas las variables (milímetros, centímetros, metros o kilómetros, por ejemplo) .

10. Los roles de X e Y son simétricos para el cálculo de r.

11. **Cuidado**: el coeficiente de correlación de Pearson es muy sensible a observaciones atípicas. Hay que hacer **siempre** un scatter plot de los datos antes de resumirlos con r. La sola presencia de una observación atípica (o de unas pocas observaciones que siguen un patrón raro) puede hacer que el valor de r resulte, por ejemplo positivo, cuando en verdad las variables X e Y están negativamente asociadas. Si vemos el scatterplot de los datos, esta (o estas) observación atípica debiera destacarse en su patrón alejado del resto y seremos capaces de detectarla. Si en cambio sólo nos limitamos a calcular el r esta situación podría escapársenos y podríamos terminar infiriendo un vínculo entre X e Y que es espúreo.

Un ejemplo de fuerte correlación positiva se da entre el volumen espiratorio esforzado (VEF), una medida de la función pulmonar, y la altura. En la Figura 6 (a) se muestra un gráfico de dispersión de observaciones de estas variables, que tienen correlación $\rho = 0{,}90$. En la Figura 6 (b) se puede observar una correlación positiva más débil entre niveles séricos de colesterol y la ingesta diaria de colesterol, aquí $\rho = 0{,}3$. Una fuerte correlación negativa ($\rho = -0{,}8$) se da entre la frecuencia del pulso en reposo (o la frecuencia cardíaca) y la edad, medidas en niños menores a diez años. Ahí vemos que a medida que un chico crece, la frecuencia de su pulso desciende. Una correlación negativa más débil $\rho = -0{,}2$ existe entre VEF y el número de cigarrillos fumados por día, como se ve en la Figura 6 (d).

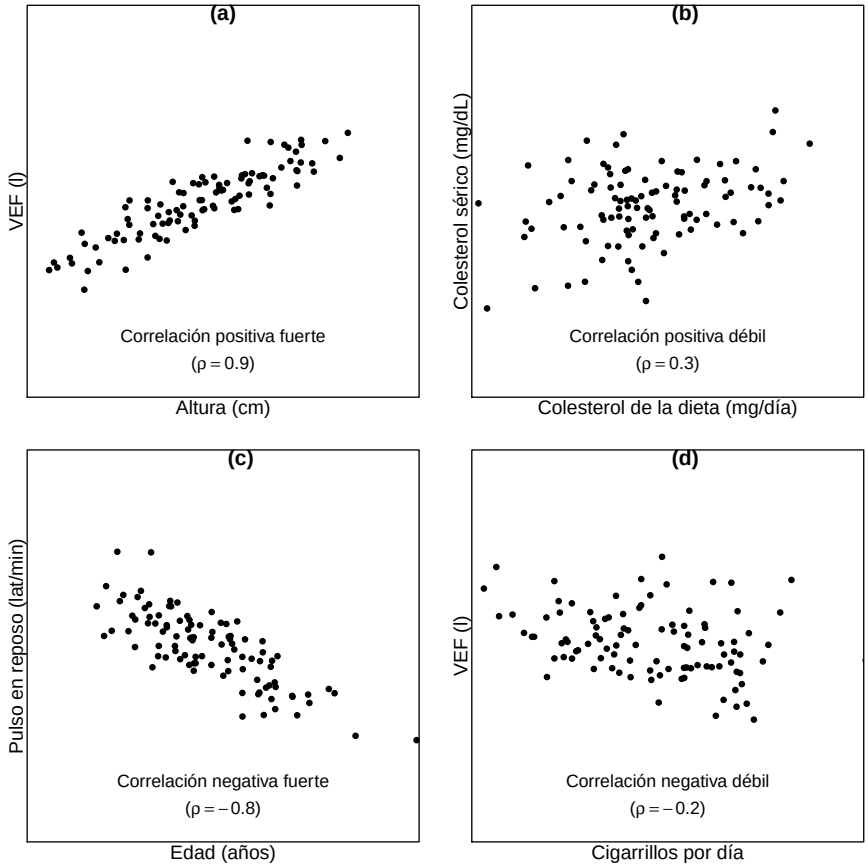
Cabe hacer un comentario respecto de la interpretación del coeficiente de correlación. Altos grados de asociación lineal entre X e Y no son señales de causalidad, es decir, una relación de causa y efecto entre ambas variables. Una alta correlación observada entre dos variables es compatible con la situación de que existan modelos que explican a Y por X, o bien a X por Y, o bien que exista una tercer variable que las determine a ambas simultáneamente.

1.2.3. Inferencia de ρ

La pregunta que nos hacemos en esta sección es la clásica pregunta de inferencia estadística, ¿qué podemos decir de ρ a partir de r?

Queremos sacar conclusiones acerca del parámetro poblacional ρ a partir de la muestra de observaciones $(X_1, Y_1), \ldots, (X_n, Y_n)$. En el ejemplo, la pregunta que podríamos hacer es ¿qué podemos decir del vínculo entre inmunización contra la DPT y la tasa de mortalidad infantil para menores a cinco años? Sólo contamos

Figura 6: Interpretación de distintos grados de correlación. Inspirado en: Rosner [2006], pág. 137.



con observaciones de 20 países en 1992. El test que más nos interesará es el que tiene las siguientes hipótesis

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0, \end{aligned}$$

ya que la afirmación de la hipótesis nula, $\rho = 0$, puede escribirse como "no hay asociación lineal entre X e Y a nivel poblacional", mientras que la hipótesis alternativa postula que sí hay tal relación entre las variables. O sea, en el caso del ejemplo, sabemos que la correlación muestral observada entre ambas variables fue $r = -0{,}791$, y la pregunta ¿será que entre las dos variables consideradas no hay asociación lineal, y sólo por casualidad en la muestra obtenida vemos un valor de

$r = -0{,}791$? ¿O será que $\rho \neq 0$? Como el coeficiente de correlación muestral r es un estimador del valor poblacional ρ, a través de él podemos proponer un test para estas hipótesis.

Test para $\rho = 0$ Los supuestos para llevar a cabo el test son que los pares de observaciones $(X_1, Y_1), \ldots, (X_n, Y_n)$ sean independientes entre sí, idénticamente distribuidos, y tengan distribución (conjunta) normal bivariada (ver la definición de esto en la Observación 1.1). En particular, esto implica que cada una de las muestras $X_1 \ldots, X_n$ e $Y_1 \ldots, Y_n$ tengan distribución normal. Si la hipótesis nula es verdadera, entonces el estadístico

$$T = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

que no es más que $\widehat{\rho}$ dividido por un estimador de su desvío estándar, tiene distribución t de Student con $n-2$ grados de libertad, lo cual notaremos

$$T \sim t_{n-2} \text{ bajo } H_0.$$

Si H_0 fuera cierto, ρ sería cero y su estimador $r = \widehat{\rho}$ debería tomar valores muy cercanos a cero. Lo mismo debería pasar con T que es $\widehat{\rho}$ estandarizado. Por lo tanto rechazaríamos la hipótesis nula cuando T tome valores muy alejados de 0, tanto positivos como negativos. El test rechaza H_0 cuando T toma valores muy grandes o muy pequeños, es decir, rechazamos la hipótesis nula con nivel $1-\alpha$ cuando

$$T \geq t_{n-2,1-\frac{\alpha}{2}} \text{ ó } T \leq -t_{n-2,1-\frac{\alpha}{2}}$$

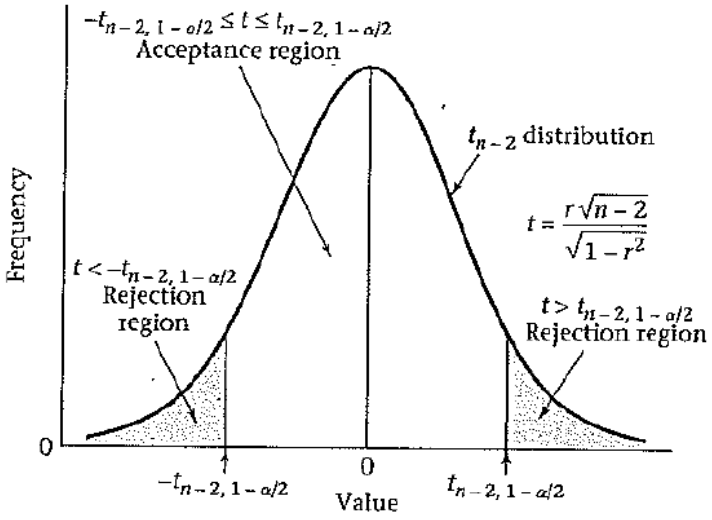
donde $t_{n-2,1-\frac{\alpha}{2}}$ es el percentil $1-\frac{\alpha}{2}$ de una distribución t_{n-2}, o sea el valor que deja a su izquierda un área de $1-\frac{\alpha}{2}$. Es un test bilateral. La región de rechazo aparece dibujada en la Figura 7. El p-valor puede calcularse como

$$p - valor = P\left(|T| \geq |T_{obs}|\right),$$

donde $T \sim t_{n-2}$ y en general lo devuelve el paquete estadístico. Si el tamaño de la muestra fuera suficientemente grande, aplicando una versión del teorema central del límite no necesitaríamos que la muestra fuera normal bivariada.

Ejemplo 1.3 *En la Tabla 3 aparece la salida del software libre* `R` *R Core Team [2015] para los datos del Ejemplo 1.1. Hemos llamado* `immunized` *al porcentaje de chicos vacunados contra la DPT y* `under5` *a la tasa de mortalidad para chicos menores a 5 años. Vemos que en este caso el p−valor del test resulta ser menor a* 0,05*, por lo que rechazamos la hipótesis nula y concluimos que el coeficiente de*

Figura 7: Región de rechazo y aceptaci ón para el test de t para una correlación. Fuente: Rosner [2006], pág. 457.



correlación poblacional ρ es no nulo, mostrando que la tasa de vacunación y la tasa de mortalidad infantil menor a 5 años están correlacionadas. Confiaremos en esta conclusión si somos capaces de creer que los datos de la muestra conjunta provienen de una distribución normal bivariada. En particular, debe cumplirse que ambas variables tengan distribución normal. Para validar al menos este último supuesto deberíamos realizar gráficos de probabilidad normal (qq-plots), histogramas o boxplots, y tests de normalidad (por ejemplo, el test de Shapiro-Wilks) y ver que ambos conjuntos de datos pasan las comprobaciones. Sin embargo, para estos conjuntos de datos no puede asumirse la distribución normal ya que ambos tienen distribución asimétrica: el porcentaje de niños vacunados con cola pesada a la derecha, la tasa de mortalidad con cola pesada a izquierda, como puede observarse en la Figura 8. Por lo tanto, no puede asumirse que conjuntamente tengan distribución normal bivariada, y no puede aplicarse el test de correlación antes descripto.

Observación 1.1 *¿Qué quiere decir que las observaciones $(X_1, Y_1), \ldots, (X_n, Y_n)$ tengan distribución conjunta normal bivariada? Es un término técnico. Decir que un vector aleatorio (X, Y) tenga dicha distribución conjunta quiere decir que existen cinco números reales $\mu_1, \mu_2, \sigma_1 > 0, \sigma_2 > 0$ y $-1 < \rho < 1$ tales que la función*

Tabla 3: Cálculo de la correlación entre el porcentaje de chicos vacunados contra la DPT y la tasa de mortalidad para chicos menores a 5 años, con el cálculo del p-valor para el test de las hipótesis $H_0 : \rho = 0$, versus $H_1 : \rho \neq 0$, e intervalo de confianza para ρ. Salida del R.

```
> cor.test(immunized,under5, method = "pearson")

        Pearson's product-moment correlation

data:  immunized and under5
t = -5.4864, df = 18, p-value = 3.281e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9137250 -0.5362744
sample estimates:
       cor
-0.7910654
```

de densidad conjunta para el vector (X, Y) *está dada por*

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right)\right]\right\}, \quad (1)$$

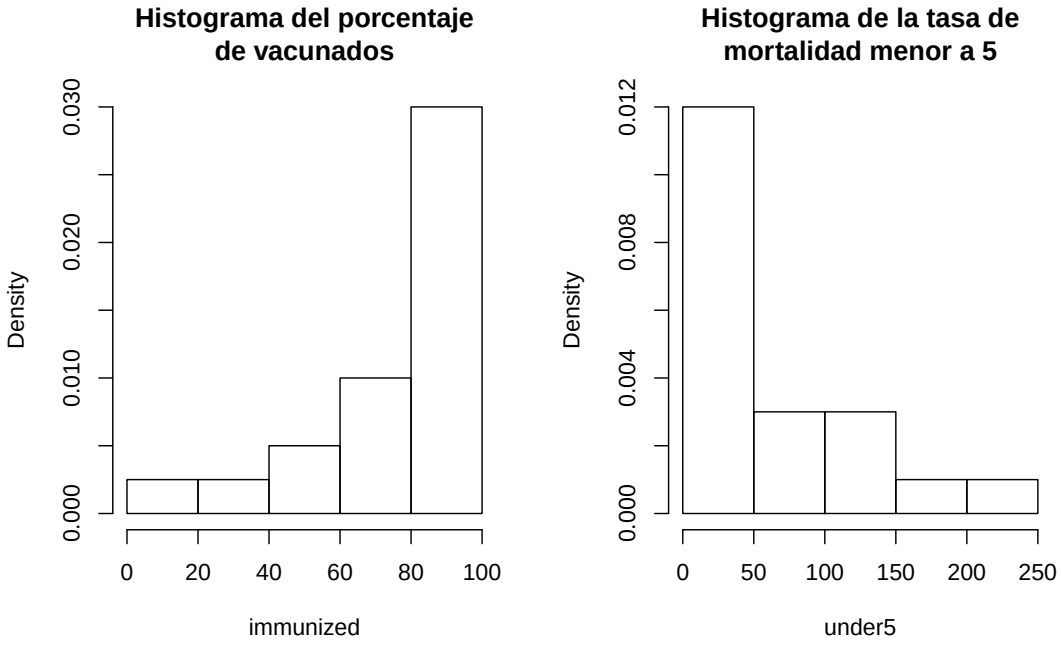
de modo que dada una determinada región A *contenida en* $\mathbb{R}^2$, *la probabilidad de que el vector* (X, Y) *pertenezca a dicha región está dada por*

$$P((X,Y) \in A) = \int\int_A f_{XY}(x,y)\,dxdy.$$

Es decir, se calcula hallando el área bajo la función f_{XY} *definida en (1) y sobre la región* A. *En la Figura 9 puede verse el gráfico de la densidad conjunta, que es una superficie en el espacio tridimensional. A los números* $\mu_1, \mu_2, \sigma_1, \sigma_2$ *y* ρ *se los denomina parámetros de la distribución normal bivariada (ya que una vez que se fija sus valores numéricos, la densidad queda determinada).*

Como ya mencionamos, puede probarse que cuando (X, Y) *tiene distribución normal bivariada, entonces cada una de las variables* X *e* Y *tienen distribución normal. Más aún,* μ_1 *es la media de* X *y* σ_1^2 *es su varianza, lo cual notamos* $X \sim$

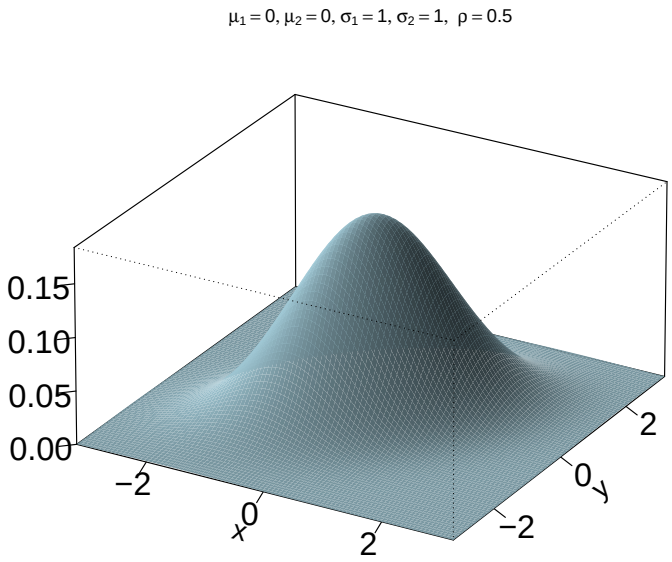
Figura 8: Histograma para los datos de porcentaje de niños vacunados y tasas de mortalidad infantil, para los datos del Ejemplo 1.1.



$N(\mu_1, \sigma_1^2)$ *y también* $Y \sim N(\mu_2, \sigma_2^2)$. *Además* ρ *es el coeficiente de correlación entre* X *e* Y. *La recíproca no es siempre cierta, sin embargo: si sabemos que tanto* X *como* Y *tienen distribución normal, entonces no siempre la distribución conjunta del vector* (X, Y) *es la normal bivariada dada por (1).*

Observación 1.2 *¿Cómo es un scatterplot de observaciones* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *independientes que tienen distribución normal bivariada? Por supuesto, el gráfico de dispersión dependerá de los valores de los parámetros. En general, los puntos yacerán en una zona que puede ser razonablemente bien descripta como una elipse con centro en* (μ_1, μ_2). *En la Figura 10 se ven los gráficos de dispersión correspondientes a distintas combinaciones de parámetros. En ellos vemos que a medida que* ρ *se acerca a uno, los puntos se acercan más a una recta, cuya pendiente y ordenada al origen depende de los valores de los cinco parámetros. La pendiente será positiva si* ρ *es positiva, y será negativa en caso contrario.*

Figura 9: Densidad conjunta normal bivariada, definida en (1.1) construida con los valores de parámetros especificados más abajo ($\mu_1 = 0, \mu_2 = 0, \sigma_1 = 1, \sigma_2 = 1, \rho =$ 0,5).



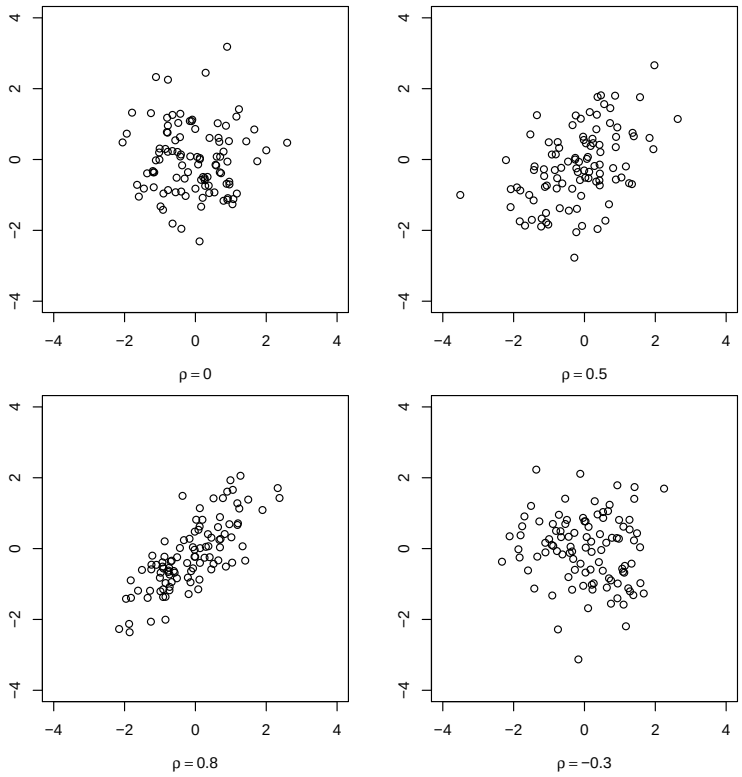
Test para $\rho = \rho_0$ A veces es de interés testear si la verdadera correlación poblacional es igual a un cierto valor ρ_0 predeterminado. Es decir, se quieren testear las hipótesis

$$\begin{aligned} H_0 &: \rho = \rho_0 \\ H_1 &: \rho \neq \rho_0. \end{aligned}$$

Por supuesto, esto no ocurre muy frecuentemente, pero puede surgir una pregunta de este tipo en algunas aplicaciones. La cuestión es que cuando $\rho = \rho_0$ el estadístico T descripto en la sección anterior no tiene distribución t de Student, sino que tiene una distribución sesgada.

Para testear las hipótesis recién propuestas, está el test basado en la transformación z de Fisher. Como en el anterior se requiere que las observaciones $(X_1, Y_1), \ldots, (X_n, Y_n)$ sean independientes entre sí, idénticamente distribuidos y

Figura 10: Gráficos de dispersión de datos bivariados con distribución normal bivariada con parámetros: $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1 = 1$, $\sigma_2 = 1$ y distintos valores de ρ, que se indican en cada gráfico.



tengan distribución conjunta normal bivariada. El test se realiza de la siguiente forma. Primero se calcula la transformación z de Fisher sobre el coeficiente de correlación, que es

$$z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right).$$

Bajo H_0, puede probarse que la distribución de z es aproximadamente

$$N \left(\frac{1}{2} \ln \left(\frac{1+\rho_0}{1-\rho_0} \right), \frac{1}{n-3} \right).$$

Luego, esta distribución se utiliza para calcular el p-valor del test, o dar la región de rechazo de nivel α. El p-valor se obtendrá estandarizando el valor de z observado y calculando la probabilidad de obtener un valor tan alejado del cero o más alejado aún como el observado, usando la función de distribución acumulada

normal estándar, es decir

$$\begin{aligned} z_{\text{est}} &= \frac{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\sqrt{\frac{1}{n-3}}} \\ p-valor &= P\left(|Z| \geq |z_{\text{est}}|\right). \end{aligned}$$

Esto lo realiza el paquete estadístico. En el ejemplo no puede aplicarse este test puesto que hemos visto ya que ninguna de las dos muestras es normal (por lo tanto los pares no pueden tener distribución conjunta normal bivariada), y este test es aún más sensible que el anterior al supuesto de normalidad.

Intervalo de confianza para ρ Puede resultar de interés disponer de un intervalo de confianza para el verdadero coeficiente de correlación poblacional, ρ, que nos dé indicios de qué parámetros poblacionales pueden describir apropiadamente a nuestros datos. Para construirlo se recurre a la transformación z presentada en la sección anterior. Luego se utiliza la distribución normal para encontrar los percentiles adecuados para describir el comportamiento del z estandarizado, y finalmente se aplica la inversa de la transformación z para obtener un intervalo de confianza para ρ. Los supuestos para llevar a cabo este procedimiento son los mismos que los presentados para ambos tests de las subsecciones anteriores. Finalmente el intervalo de confianza de nivel $1-\alpha$ para ρ está dado por $[\rho_I, \rho_D]$ donde

$$\begin{aligned} z_{\text{obs}} &= \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \\ \rho_I &= \frac{e^{2\left(z_{\text{obs}} - z_{1-\frac{\alpha}{2}}\cdot\frac{1}{\sqrt{n-3}}\right)} - 1}{e^{2\left(z_{\text{obs}} - z_{1-\frac{\alpha}{2}}\cdot\frac{1}{\sqrt{n-3}}\right)} + 1} \\ \rho_D &= \frac{e^{2\left(z_{\text{obs}} + z_{1-\frac{\alpha}{2}}\cdot\frac{1}{\sqrt{n-3}}\right)} - 1}{e^{2\left(z_{\text{obs}} + z_{1-\frac{\alpha}{2}}\cdot\frac{1}{\sqrt{n-3}}\right)} + 1} \end{aligned}$$

y $z_{1-\frac{\alpha}{2}}$ es el percentil $1-\frac{\alpha}{2}$ de la normal estándar. En el caso del ejemplo no tiene sentido mirarlo porque no se cumplen los supuestos, pero puede verse la salida del `R` en la Tabla 3 donde aparece calculado por la computadora, y da $[-0{,}91, -0{,}54]$.

1.3. Coeficiente de correlación de Spearman

Existen otras medidas de asociación entre dos variables que no son tan sensibles a observaciones atípicas como el coeficiente de correlación de Pearson, ni necesitan el supuesto de normalidad para testearse. La más difundida de ellas es

el coeficiente de correlación de Spearman, que presentamos en esta sección. El coeficiente de correlación de Spearman se encuadra entre las técnicas estadísticas no paramétricas, que resultan robustas bajo la presencia de outliers ya que reemplazan los valores observados por los rangos o rankings de las variables. Se calcula del siguiente modo.

1. Se ordena cada muestra por separado, de menor a mayor. A cada observación se le calcula el ranking que tiene (o rango, o número de observación en la muestra ordenada). De este modo, la observación más pequeña de las $X's$ recibe el número 1 como rango, la segunda recibe el número 2, etcétera, la más grande de todas las $X's$ recibirá el rango n. Si hubiera dos o más observaciones empatadas en algún puesto (por ejemplo, si las dos observaciones más pequeñas tomaran el mismo valor de X, entonces se promedian los rangos que les tocarían: cada una tendrá rango 1,5, en este ejemplo, ya que $\frac{1+2}{2} = 1{,}5$. En el caso en el que las tres primeras observaciones fueran empatadas, a las tres les tocaría el promedio entre 1, 2 y 3, que resultará ser $\frac{1+2+3}{3} = 2$). A este proceso se lo denomina *ranquear las observaciones* X. Llamemos $R(X_i)$ al rango obtenido por la i-ésima observación X.

2. Se reemplaza a cada observación X_i por su rango $R(X_i)$.

3. Se ranquean las observaciones Y, obteniéndose $R(Y_i)$ de la misma forma en que se hizo en el ítem 1 para las $X's$.

4. Se reemplaza a cada observación Y_i por su rango $R(Y_i)$. Observemos que conocemos la suma de todos los rangos de ambas muestras (es la suma de $1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}$).

5. Se calcula el coeficiente de correlación de Pearson entre los pares $(R(X_i), R(Y_i))$. El valor obtenido es el coeficiente de correlación de Spearman, que denotaremos r_S.

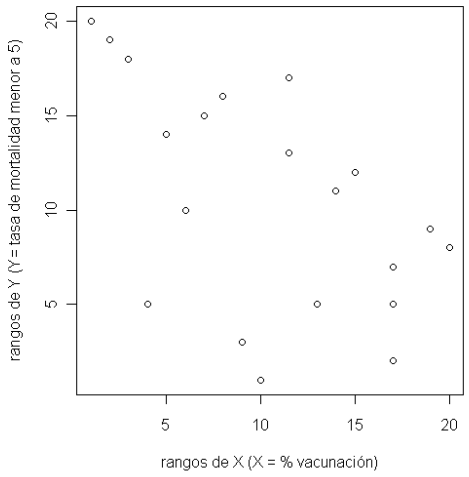
Ilustramos el procedimiento con los datos del Ejemplo 1.1, de la vacunación de DPT, en la Tabla 4. Allí figuran las originales X e Y en las columnas 1 y 3, y los rangos de cada muestra: los rangos de las $X's$ en la columna 2 y los rangos de las $Y's$ en la columna 4. Ahí vemos que Etiopía es el país de la muestra con menor tasa de vacunación, por eso su valor X recibe el rango 1. Lo sigue Camboya. Observamos que hay dos países cuyo porcentaje de vacunación es 89 %: Egipto e India. Ambos empatan en los puestos 11 y 12 de la muestra ordenada, por eso reciben el rango 11,5. Y también hay 3 países con un 95 % de bebés vacunados (Finlandia, Francia e Italia) que, a su vez, empatan en los puestos 16, 17 y 18 y reciben el rango promedio de esos tres valores, o sea, 17. Es interesante observar

que Etiopía recibe el rango 1 (el menor) para el porcentaje de vacunación, y el rango 20 (el mayor) para la tasa de mortalidad menor a 5 años, Camboya, a su vez, recibe el rango 2 (el segundo más chico) para el porcentaje de vacunación, y el rango 19 (el penúltimo) para la tasa de mortalidad. En ambos órdenes, lo sigue Senegal, esto muestra la asociación negativa entre ambas variables. Para evaluar si esto ocurre con el resto de los países, hacemos un scatter plot de los rangos de Y versus los rangos de X en la Figura 11. En ella se ve una asociación negativa entre los rangos de ambas variables, aunque no se trata de una asociación muy fuerte, sino más bien moderada. Los tres puntos con menores rangos de X mantienen una relación lineal perfecta, como habíamos observado. Sin embargo, ese ordenamiento se desdibuja en las demás observaciones.

Tabla 4: Datos para los 20 países, con las variables, X : porcentaje de niños vacunados a la edad de un año en cada país, rangos de la X : ranking que ocupa la observación en la muestra ordenada de las X's, Y : tasa de mortalidad infantil de niños menores de 5 años en cada país, rangos de la Y : posición que ocupa la observación en la muestra ordenada de las Y's.

País	Porcentaje vacunado (X)	Rangos de X	Tasa de mortalidad menor a 5 años (Y)	Rangos de Y
Bolivia	77,0	8	118,0	16
Brasil	69,0	5	65,0	14
Camboya	32,0	2	184,0	19
Canadá	85,0	9	8,0	3
China	94,0	15	43,0	12
República Checa	99,0	20	12,0	8
Egipto	89,0	11,5	55,0	13
Etiopía	13,0	1	208,0	20
Finlandia	95,0	17	7,0	2
Francia	95,0	17	9,0	5
Grecia	54,0	4	9,0	5
India	89,0	11,5	124,0	17
Italia	95,0	17	10,0	7
Japón	87,0	10	6,0	1
México	91,0	14	33,0	11
Polonia	98,0	19	16,0	9
Federación Rusa	73,0	6	32,0	10
Senegal	47,0	3	145,0	18
Turquía	76,0	7	87,0	15
Reino Unido	90,0	13	9,0	5

Figura 11: Gráfico de dispersión entre los rangos de Y (es decir, los rangos de la tasa de mortalidad menor a 5 años) y los rangos de X (es decir, del porcentaje de niños menores a un año vacunados contra la DPT). Se ve una asociación negativa, aunque no muy estrecha.



¿Cómo resumimos el grado de asociación observado entre los rangos? Con el cálculo del coeficiente de correlación entre ellos. En este caso da $r_S = -0{,}543$, como puede verse en la Tabla 5. Este número resulta menor en magnitud que el coeficiente de correlación de Pearson, pero sugiere una moderada relación entre las variables. Esta asociación es negativa.

Otro test de asociación entre variables. El coeficiente de correlación de Spearman puede usarse para testear las hipótesis

H_0 : No hay asociación entre X e Y

H_1 : Hay asociación entre X e Y : hay una tendencia de que los valores más grandes de X se apareen con los valores más grandes de Y, o bien la tendencia es que los valores más pequeños de X se apareen con los valores más grandes de Y

Como H_1 incluye la posibilidad de que la asociación sea positiva o negativa, este es un test de dos colas. El test rechazará para valores grandes de $|r_S|$ (valor absoluto de r_S). La distribución de r_S bajo H_0 no es difícil de obtener. Se basa en el hecho de

que, bajo H_0, para cada rango de X_i, $R(X_i)$, todos los rangos de Y_i son igualmente probables, siempre que no haya asociación entre ambas variables. El $p-$valor puede calcularse de manera exacta si $n < 10$ y no hay empates en la muestra, y de manera aproximada para n mayores.

Si n es muy grande, se utiliza la misma distribución t de la Sección anterior, t_{n-2}. La ventaja de este test por sobre el test de Pearson es que requiere menos supuestos para llevarlo a cabo. Basta con que los pares de observaciones $(X_1, Y_1), \ldots, (X_n, Y_n)$ sean independientes entre sí e idénticamente distribuidos. No es necesario asumir nada respecto de la distribución de cada muestra, de hecho basta que la escala de las observaciones sea ordinal para poder aplicarlo. Puede utilizarse si hay observaciones atípicas. La desventaja radica en la potencia del test. El test de Spearman tiene una potencia menor en el caso en el que ambas muestras son normales (en cualquier otro caso, el de Pearson no puede aplicarse). Pero, por supuesto que si con el test de Spearman se logra rechazar la hipótesis nula, ya no es necesario preocuparse por la potencia, ni utilizar el coeficiente de Pearson, que resulta más eficiente.

Tabla 5: Cálculo de la correlación de Spearman entre el porcentaje de chicos vacunados contra la DPT (`immunized`) y la tasa de mortalidad para chicos menores a 5 años (`under5`), con el cálculo del p-valor con el coeficiente de Spearman, para el test de las hipótesis H_0 : no hay asociación entre las variables, versus H_1 : las variables están positiva o negativamente asociadas. Salida del `R`.

```
> cor.test(immunized,under5, method = "spearman")

        Spearman's rank correlation rho

data:  immunized and under5
S = 2052.444, p-value = 0.01332
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.5431913
```

En el ejemplo vemos que el $p-$valor del test de Spearman es 0,013 que al ser menor a 0,05 nos permite rechazar la hipótesis nula y concluir que la verdadera correlación poblacional entre el porcentaje de niños vacunados y la tasa de mortalidad menor a 5 años, es distinta de cero.

Otra medida no paramétrica de asociación entre dos variables está dada por el τ de Kendall. Resume la asociación a través de los rangos de las observaciones

de ambas muestras, pero de una manera diferente. Puede consultarse Kendall y Gibbons [1990] para una discusión completa del tema.

Un último comentario respecto de la correlación en el contexto del estudio de regresión lineal. En este contexto, no estaremos tan interesados en los tests presentados para probar si existe correlación entre las variables, sino más bien en el uso de la correlación a nivel descriptivo. Nos servirá en una primera etapa exploratoria de los datos para ver si las variables bajo nuestra consideración están asociadas con una variable Y que nos interesa explicar, y qué grado de fuerza tiene esa asociación lineal. Y también nos servirá para entender ciertos comportamientos extraños que presenta la regresión lineal múltiple cuando se ajusta para modelos con muchas covariables muy correlacionadas entre sí.

1.4. Ejercicios

Con `R` hacer scatterplots es muy sencillo. Además es tan útil lo que puede aprenderse de los datos que vale la pena entrenarse exponiéndose a muchos ejemplos. Con el tiempo se gana familiaridad con los tipos de patrones que se ven. De a poco uno aprende a reconocer cómo los diagramas de dispersión pueden revelar la naturaleza de la relación entre dos variables.

En esta ejercitación trabajaremos con algunos conjuntos de datos que están disponibles a través del paquete `openintro` de `R`. Brevemente:

`mammals`: El conjunto de datos de mamíferos contiene información sobre 62 especies diferentes de mamíferos, incluyendo su peso corporal, el peso del cerebro, el tiempo de gestación y algunas otras variables.

`bdims`: El conjunto de datos `bdims` contiene medidas de circunferencia del cuerpo y diámetro esquelético para 507 individuos físicamente activos.

`smoking`: El conjunto de datos `smoking` contiene información sobre los hábitos de fumar de 1.691 ciudadanos del Reino Unido.

`cars`: El conjunto de datos `cars` está compuesto por la información de 54 autos modelo 1993. Se relevan 6 variables de cada uno (tamaño, precio en dólares, rendimiento en ciudad (millas por galón), tipo de tracción, cantidad de pasajeros, peso).

Para ver una documentación más completa, utilice las funciones `?` ó `help()`, una vez cargado el paquete. Por ejemplo, `help(mammals)`. Esta práctica se resuelve con el `script_correlacion.R`

Ejercicio 1.1 *Mamíferos, Parte I. Usando el conjunto de datos de* `mammals`*, crear un diagrama de dispersión que muestre cómo el peso del cerebro de un mamífero (*`BrainWt`*) varía en función de su peso corporal (*`BodyWt`*).*

Ejercicio 1.2 *Medidas del cuerpo, Parte I. Utilizando el conjunto de datos* `bdims`*, realizar un diagrama de dispersión que muestre cómo el peso de una persona (*`wgt`*) varía en función de su altura (*`hgt`*). Identifique el género de las observaciones en el scatterplot, para ello pinte de rojo a las mujeres y de azul a los hombres, use la instrucción* `col` *de* `R`*. Observar que en esta base de datos,* `sex = 1` *para los hombres y* `sex = 0` *para las mujeres.*

Ejercicio 1.3 *Utilizando el conjunto de datos* `smoking`*, realizar un diagrama de dispersión que ilustre cómo varía la cantidad de cigarrillos que fuma por día una persona durante el fin de semana (*`amtWeekends`*), en función de su edad (*`age`*).*

Ejercicio 1.4 *Utilizando el conjunto de datos* `cars`*, realizar un scatter plot del rendimiento del auto en la ciudad (*`mpgCity`*) en función del peso del auto (*`weight`*).*

Ejercicio 1.5 *Para cada uno de los cuatro scatterplots anteriores describa la forma, la dirección y la fuerza de la relación entre las dos variables involucradas. Respuestas posibles:*

- *forma: lineal, no lineal (cuadrática, exponencial, etc.)*
- *dirección: positiva, negativa*
- *fuerza de la relación: fuerte, moderada, débil, no asociación. Tiene que ver con cuán dispersos están las observaciones respecto del patrón descripto en la forma.*

Ejercicio 1.6 *¿Para cuáles de los 4 conjuntos de datos tiene sentido resumir la relación entre ambas variables con el coeficiente de correlación muestral de Pearson? Para los casos en los cuales contestó que era apropiado,*

(a) calcúlelo usando R.

(b) Testee las siguientes hipótesis

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

para cada uno de esos conjuntos. Antes de hacerlo defina a ρ en palabras. Observe que en el ítem 1.6 (a) calculó un estimador de esta cantidad, para cada conjunto. ¿En qué casos rechaza la hipótesis nula, a nivel 0.05?

Ejercicio 1.7 *Mamíferos, Parte II. El conjunto de datos de* `mammals` *presenta un scatterplot que no es razonable resumir con el coeficiente de correlación muestral. El gráfico no es lindo por varios motivos, básicamente las observaciones parecen estar en escalas distintas, hay muchas observaciones superpuestas, necesitaríamos hacer un zoom del gráfico en la zona cercana al origen, a expensas de perder las dos observaciones con valores mucho más grandes que el resto. Podemos comparar lo que pasaría si no hubiéramos observado el diagrama de dispersión y quisiéramos resumir los datos con el coeficiente de correlación.*

(a) Calcule el coeficiente de correlación muestral de Pearson para los 62 mamíferos.

(b) Identifique las dos observaciones que tienen valores de peso corporal y cerebral más grandes que el resto. Realice un scatter plot de las restantes 60 variables. ¿Cómo podría describir este gráfico? Calcule el coeficiente de correlación muestral de Pearson para estas 60 observaciones.

(c) El gráfico hecho en el ítem anterior no corrige el problema original del todo. La forma general podría describirse como un abanico: claramente las variables están asociadas, la asociación es positiva (ambas crecen simultáneamente) pero la dispersión de los datos parece aumentar a medida que ambas variables aumentan. Esta forma es frecuente en los conjuntos de datos, suelen corresponder a observaciones que están medidas en escalas que no son comparables entre sí y suele corregirse al tomar logaritmo en ambas variables. Para ver el efecto de transformar las variables, realice un scatterplot con todas las observaciones, del logaritmo (en base 10, o en base e) del peso del cerebro en función del logaritmo del peso corporal. Observe el gráfico. ¿Cómo lo describiría? Calcule la correlación de Pearson para los datos transformados.

(d) Para ambos conjuntos de datos (transformados por el logaritmo y sin transformar) calcule la correlación de Spearman.

Ejercicio 1.8 *¿Con qué coeficiente de correlación, Pearson o Spearman, resumiría los datos de* `cars`*? (*`weight,mpgCity`*)*

2. Regresión lineal simple

2.1. Introducción

Antes de presentar el modelo lineal, comencemos con un ejemplo.

Ejemplo 2.1 *Datos publicados en Leviton, Fenton, Kuban, y Pagano [1991], tratados en el libro de Pagano et al. [2000].*

Los datos corresponden a mediciones de 100 niños nacidos con bajo peso (es decir, con menos de 1500g.) en Boston, Massachusetts. Para dichos bebés se miden varias variables. La variable que nos interesa es el perímetro cefálico al nacer (medido en cm.). Los datos están en el archivo `low birth weight infants.txt`*, la variable* `headcirc` *es la que contiene los datos del perímetro cefálico. No tiene sentido tipear los 100 datos, pero al menos podemos listar algunos, digamos los primeros 14 datos: 27, 29, 30, 28, 29, 23, 22, 26, 27, 25, 23, 26, 27, 27. La lista completa está en el archivo. Asumamos que entra ahora una madre con su bebé recién nacido en mi consultorio de niños de bajo peso, y quiero predecir su perímetro cefálico, con la información que me proporciona la muestra de los 100 bebés. ¿Cuál debiera ser el valor de perímetro cefálico que le predigo? O sea, me estoy preguntando por el mejor estimador del perímetro cefálico medio de un bebé de bajo peso, sin otra información a mano más que la muestra de 100 bebés antes descripta. Si llamamos Y a la variable aleatoria:*

$$Y = \text{perímetro cefálico (medido en cm.) de un bebé recién nacido con bajo peso,}$$

estamos interesados en estimar a la media poblacional $E(Y)$. Sabemos que la media muestral $\overline{Y}_{100}$ será el mejor estimador que podemos dar para la media poblacional $E(Y)$. Los estadísticos de resumen para la muestra dada figuran en la Tabla 6.

Tabla 6: Medidas de resumen de los datos de perímetro cefálico.

Variable	n	Media muestral	Desvío estándar muestral
Perímetro cefálico	100	26,45	2,53

Luego, nuestro valor predicho será 26,45 *cm. de perímetro cefálico. El desvío estándar muestral es* 2,53*. Más aún, podríamos dar un intervalo de confianza para la media poblacional, basado en la muestra (ver la Tabla 7).*

Tabla 7: Intervalo de confianza para el perímetro cefálico medio, basado en los 100 datos disponibles (calculado con R).

```
> t.test(headcirc)

        One Sample t-test

data:  headcirc

t = 104.46, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 25.94757 26.95243
sample estimates:
mean of x
    26.45
```

Ejemplo 2.2 *Por lo tanto, el intervalo de confianza para* $E(Y)$ *resulta ser* [25,95, 26,95], *ver la Tabla 7.*

Pero ¿qué pasaría si contáramos con información adicional a la ya descripta en la muestra de 100 bebés de bajo peso? Además del perímetro cefálico al nacer, se miden otras variables en los 100 bebés en cuestión. La siguiente tabla las exhibe, para los primeros 14 bebés. Las iremos describiendo en la medida en la que las analicemos.

Comenzaremos por estudiar dos de estas variables conjuntamente. Es decir, miraremos `headcirc`: "perímetro cefálico al nacer (medido en cm.)" y `gestage`: "edad gestacional, es decir, duración del embarazo (medida en semanas)". La idea es ver si podemos predecir de una mejor manera el perímetro cefálico de un bebé al nacer si conocemos su edad gestacional. Podemos pensar en estas observaciones como en $n = 100$ observaciones apareadas (X_i, Y_i) con $1 \leq i \leq n$, donde Y_i es la variable respuesta medida en el i-ésimo individuo (o i-ésima repetición o i-ésima unidad experimental, según el caso), y X_i es el valor de la variable predictora en el i-ésimo individuo. En el ejemplo,

$Y_i =$ perímetro cefálico del *i-ésimo bebé* de bajo peso (`headcirc`)
$X_i =$ edad gestacional o duración de la gestación del *i-ésimo bebé* de bajo peso (`gestage`)

En la Figura 12 vemos un scatter plot (gráfico de dispersión) del perímetro cefálico versus la edad gestacional, para los 100 niños.

Tabla 8: Primeros 14 datos de los bebés de bajo peso

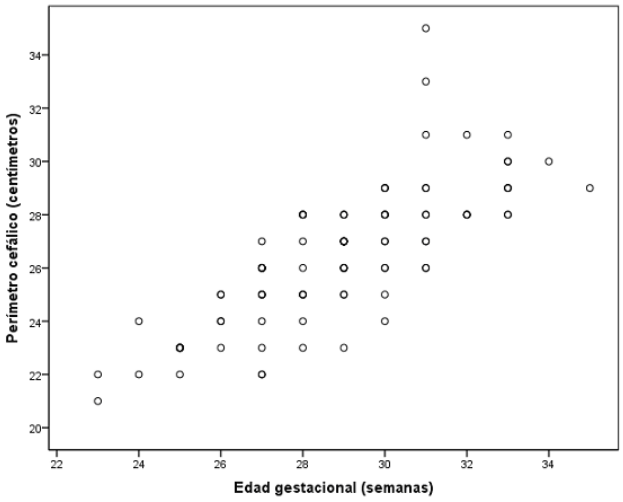
Caso	headcirc	length	gestage	birthwt	momage	toxemia
1	27	41	29	1360	37	0
2	29	40	31	1490	34	0
3	30	38	33	1490	32	0
4	28	38	31	1180	37	0
5	29	38	30	1200	29	1
6	23	32	25	680	19	0
7	22	33	27	620	20	1
8	26	38	29	1060	25	0
9	27	30	28	1320	27	0
10	25	34	29	830	32	1
11	23	32	26	880	26	0
12	26	39	30	1130	29	0
13	27	38	29	1140	24	0
14	27	39	29	1350	26	0

El scatter plot del perímetro cefálico versus la edad gestacional sugiere que el perímetro cefálico aumenta al aumentar la edad gestacional. Y que dicho aumento pareciera seguir un patrón lineal.

Observemos que, como ya dijimos, a veces el gráfico de dispersión no permite ver la totalidad de las observaciones: el scatter plot recién presentado contiene información correspondiente a 100 bebés, pero parece que hubiera menos de 100 puntos graficados. Esto se debe a que los resultados de las dos variables graficadas están redondeados al entero más cercano, muchos bebés aparecen con valores idénticos de perímetro cefálico y edad gestacional; en consecuencia algunos pares de datos son graficados sobre otros.

Si calculamos el coeficiente de correlación lineal para estos datos nos da 0,781, indicando fuerte asociación lineal entre X e Y, ya que el valor obtenido está bastante cerca de 1. Antes de realizar inferencias que involucren al coeficiente de correlación hay que verificar que se cumplen los supuestos de normalidad conjunta. Estos son difíciles de testear. Sin embargo, el gráfico de dispersión puede describirse globalmente mediante una elipse. Además podemos chequear la normalidad de ambas muestras (haciendo, por ejemplo un test de Shapiro-Wilks y un qqplot de los datos). Una vez verificado el supuesto de normalidad, podemos analizar el test. (Si los datos no sustentaran la suposición de normalidad, deberíamos usar el coeficiente de correlación de Spearman para evaluar la correlación existente entre ellos). Los resultados aparecen en la Figura 9. Recordemos que el $p-$valor obtenido en el test (menor a 0,0001 da casi cero trabajando con 4 decimales de precisión)

Figura 12: Gráfico de dispersión de perímetro cefálico versus edad gestacional, para 100 bebés de bajo peso.



significa que en el test de $H_0 : \rho = 0$ versus la alternativa $H_1 : \rho \neq 0$ donde ρ es el coeficiente de correlación poblacional, rechazamos la hipótesis nula a favor de la alternativa y resulta que ρ es significativamente distinto de cero, indicando que efectivamente hay una relación lineal entre ambas variables.

Observemos que si bien ahora sabemos que ambas variables están linealmente asociadas, todavía no podemos usar esta información para mejorar nuestra predicción del perímetro cefálico de un bebé recién nacido, de bajo peso. Para hacerlo, proponemos el modelo lineal.

2.2. Modelo lineal simple

El modelo de regresión lineal es un modelo para el vínculo de dos variables aleatorias que denominaremos $X =$ *variable predictora o covariable* e $Y =$ *variable dependiente o de respuesta*. El modelo lineal (simple pues sólo vincula una variable predictora con Y) propone que

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{2}$$

donde ε es el término del error. Esto es que para cada valor de X, la correspondiente observación Y consiste en el valor $\beta_0 + \beta_1 X$ más una cantidad ε, que puede ser positiva o negativa, y que da cuenta de que la relación entre X e Y no es exactamente lineal, sino que está expuesta a variaciones individuales que hacen que el

Tabla 9: Correlación entre perímetro cefálico y edad gestacional, en R.

```
> cor.test(gestage,headcirc)

        Pearson's product-moment correlation

data:  gestage and headcirc

t = 12.367, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6900943 0.8471989

 sample estimates:
      cor
0.7806919
```

par observado (X, Y) no caiga exactamente sobre la recta, sino cerca de ella, como puede anticiparse viendo el scatter plot de los datos que usualmente se modelan con este modelo (ver, por ejemplo, la Figura 12). En el modelo (2) los números β_0 y β_1 son constantes desconocidas que se denominan *parámetros* del modelo, o *coeficientes* de la ecuación. El modelo se denomina "lineal" pues propone que la Y depende linealmente de X. Además, el modelo es lineal en los parámetros: los $\beta' s$ no aparecen como exponentes ni multiplicados o divididos por otros parámetros. Los parámetros se denominan

$$\begin{aligned} \beta_0 &= \text{ordenada al origen} \\ \beta_1 &= \text{pendiente.} \end{aligned}$$

Otra forma de escribir el mismo modelo es pensando en las observaciones (X_i, Y_i). En tal caso, el modelo (2) adopta la forma

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \tag{3}$$

donde ε_i es el término del error para el individuo i-ésimo, que **no es observable**.

Antes de escribir los supuestos del modelo, hagamos un breve repaso de ecuación de la recta, en un ejemplo sencillo.

2.3. Ecuación de la recta

Estudiemos el gráfico y el vínculo entre x e y que impone la ecuación de la recta. Miremos en particular la recta

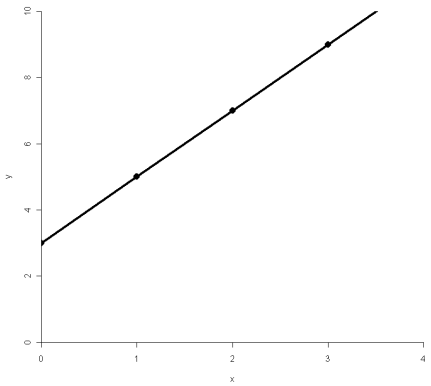
$$y = 2x + 3$$

En este caso la pendiente es $\beta_1 = 2$, y la ordenada al origen es $\beta_0 = 3$. Antes de graficarla armamos una tabla de valores de la misma.

x	y
0	3
1	5
2	7
3	9

Grafiquemos. Nos basta ubicar dos puntos sobre la misma, por ejemplo el $(0, 3)$ y el $(1, 5)$.

Figura 13: Gráfico de la recta $y = 2x + 3$.



Observemos que al pasar de $x = 0$ a $x = 1$, el valor de y pasa de 3 a 5, es decir, se incrementa en 2 unidades. Por otra parte, al pasar de $x = 1$ a $x = 2$, el valor de y pasa de 5 a 7, o sea, nuevamente se incrementa en 2 unidades. En general, al pasar de cualquier valor x a $(x + 1)$, el valor de y pasa de $2x + 3$ a $2(x + 1) + 3$, es decir, se incrementa en

$$\begin{aligned}[2(x+1)+3] - [2x+3] &= 2x + 2 + 3 - 2x - 3 \\ &= 2\end{aligned}$$

que es la pendiente. Por lo tanto, la pendiente representa el cambio en y cuando x aumenta una unidad.

Luego de este breve repaso, retomemos el modelo lineal, escribiendo los supuestos bajo los cuales es válido.

2.4. Supuestos del modelo lineal

Tomando en cuenta el repaso realizado de la ecuación de la recta, podemos decir que en el scatter plot de la Figura 12, hemos visto que una relación lineal indica la tendencia general por la cual el perímetro cefálico varía con la edad gestacional. Se puede observar que la mayoría de los puntos no caen exactamente sobre una línea. La dispersion de los puntos alrededor de cualquier línea que se dibuje representa la variación del perímetro cefálico que no está asociada con la edad gestacional, y que usualmente se considera que es de naturaleza aleatoria. Muchas veces esta aleatoriedad se debe a la falta de información adicional (datos genéticos del niño y sus padres, abultada información acerca del embarazo que incluyan tratamientos seguidos por la madre, datos de alimentación, raza, edad de la madre, etc.) y de un modelo complejo que pueda dar un adecuado vínculo funcional entre estos datos y la variable respuesta (en este caso el perímetro cefálico del recién nacido de bajo peso). Por otro lado, como se espera que todos estos componentes diversos se sumen entre sí y tengan un aporte muy menor a la explicación de la variable respuesta comparada con el de la explicativa considerada, se los puede modelar adecuadamente asumiendo que todas estas características independientes de la edad gestacional y asociadas al individuo las incluyamos en el término del error, que al ser suma de muchas pequeñas variables independientes (y no relevadas) podemos asumir que tiene distribución normal. Lo cual no se alejará mucho de la realidad en muchos de los ejemplos prácticos de aplicación del modelo de regresión.

Los supuestos bajo los cuales serán válidas las inferencias que haremos más adelante sobre el modelo

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \tag{4}$$

son los siguientes:

1. los ε_i tiene media cero, $E(\varepsilon_i) = 0$.

2. los ε_i tienen todos la misma varianza desconocida que llamaremos σ^2 y que es el otro parámetro del modelo, $Var(\varepsilon_i) = \sigma^2$. A este requisito se lo suele llamar *homoscedasticidad.*

3. los ε_i tienen distribución normal

4. los ε_i son independientes entre sí, y son no correlacionados con las X_i.

El hecho de que los errores no estén correlacionados con las variables explicativas apunta a que el modelo esté identificado. Observemos que estos cuatro supuestos pueden resumirse en la siguiente expresión

$$\varepsilon_i \sim N\left(0, \sigma^2\right), \quad 1 \leq i \leq n, \quad \text{independientes entre sí.} \tag{5}$$

Remarquemos que en la ecuación (4) lo único que se observa es (X_i, Y_i) : desconocemos tanto a β_0 como a β_1 (que son números fijos), a ε_i no lo observamos. Notemos que si bien en la ecuación (4) sólo aparecen dos parámetros desconocidos, β_0 y β_1, en realidad hay tres parámetros desconocidos, el tercero es σ^2.

Otra manera de escribir los supuestos es observar que a partir de la ecuación (4) o (2) uno puede observar que **para cada valor fijo de la variable** X, el valor esperado de la respuesta Y depende de X de manera lineal, es decir escribir el modelo en términos de la esperanza de Y condicional a las X's que notaremos $E\left(Y \mid X\right)$. Esto constituye un modo muy utilizado de escribir el modelo de regresión lineal simple. En este caso los supuestos son:

1. La esperanza condicional de Y depende de X de manera lineal, es decir

$$E\left(Y \mid X\right) = \beta_0 + \beta_1 X \tag{6}$$

   o, escrito de otro modo

$$E\left(Y \mid X = x_i\right) = \beta_0 + \beta_1 x_i \tag{7}$$

   donde β_0, β_1 son los parámetros del modelo, o coeficientes de la ecuación. A la ecuación (6) se la suele llamar **función de respuesta**, es una recta.

2. La varianza de la variable respuesta Y dado que la predictora está fijada en $X = x$ la denotaremos por $Var\left(Y \mid X = x\right)$. Asumimos que satisface

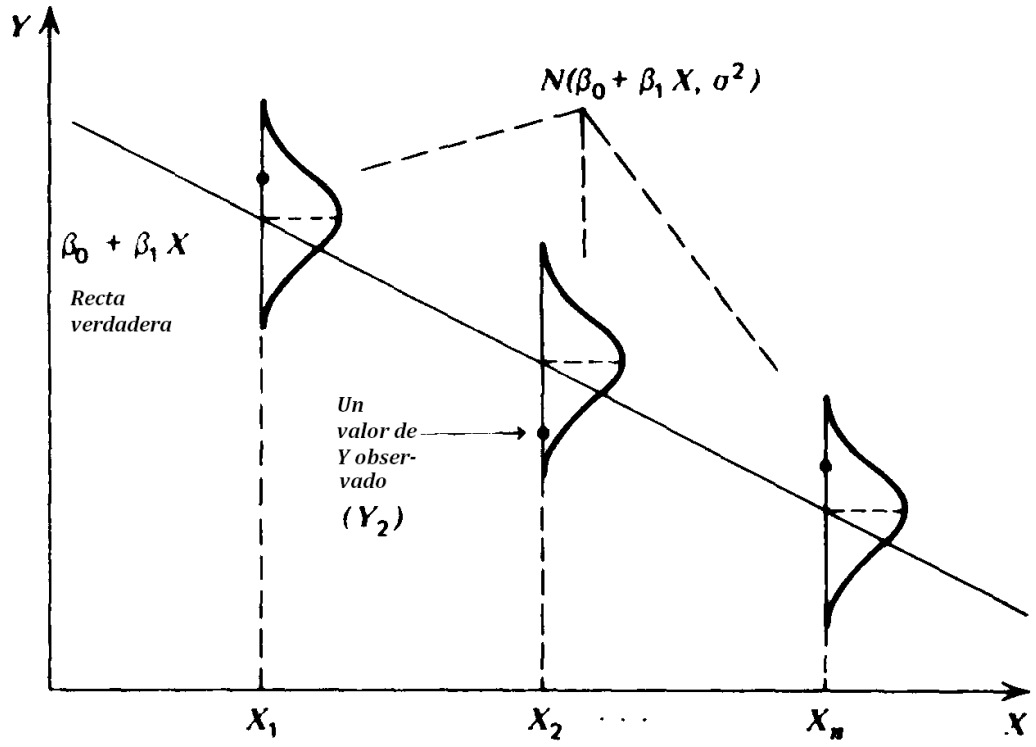
$$Var\left(Y \mid X = x_i\right) = \sigma^2,$$

   o sea, es constante (una constante desconocida y positiva) y no depende del valor de X.

3. Las Y_i, es decir, el valor de la variable Y cuando X toma el valor i-ésimo observado, (o sea, el valor de $Y \mid X = x_i$) tienen distribución normal, es decir, $Y \mid X = x_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$.

4. Las Y_i son independientes entre sí.[1]

Figura 14: Suponemos que cada observación de la variable respuesta proviene de una distribución normal centrada verticalmente en el nivel implicado por el modelo lineal asumido. Asumimos que la varianza de cada distribución normal es la misma, σ^2. Fuente: Draper y Smith [1998], p. 34.



Ejemplificamos gráficamente los supuestos en la Figura 14.

Si para algún conjunto de datos estos supuestos no se verifican (por ejemplo, las observaciones no son independientes porque hay varias mediciones de los mismos pacientes, o la varianza de Y crece a medida que crece X) no se puede aplicar el modelo de regresión lineal a dichos datos. Es necesario trabajar con modelos más refinados, que permitan incluir estas estructuras en los datos, por ejemplo, modelos de ANOVA con alguna predictora categórica que agrupe observaciones realizadas a los mismos individuos, o modelo lineal estimado con mínimos cuadrados pesados, que permiten incluir ciertos tipos de heteroscedasticidades.

[1]En realidad, se pueden hacer supuestos más débiles aún: asumir que $E(\varepsilon_i \mid X_i) = 0$, y $Var(\varepsilon_i \mid X_i) = \sigma^2$. Para los test se asume que $\varepsilon_i \mid X_i \sim N(0, \sigma^2)$, $1 \leq i \leq n$, ver Wasserman [2010].

El modelo de regresión lineal tiene tres parámetros a ser estimados, β_0, β_1 y σ^2. ¿Qué nos interesa resolver?

1. Estimar los parámetros a partir de las observaciones.
2. Hacer inferencias sobre los pámetros (tests e intervalos de confianza para β_0, β_1 y σ^2).
3. Dar alguna medida de la adecuación del modelo a los datos.
4. Evaluar si se cumplen los supuestos (resúmenes, gráficos, tests).
5. Estimar la esperanza condicional de Y para algún valor de X observado o para algún valor de X que no haya sido observado en la muestra, y construir un intervalo de confianza para dicha esperanza, como para tener idea del error a que se está expuesto.
6. Dar un intervalo de predicción para el valor de Y de una nueva observación para la cual tenemos el valor de X.
7. Describir los alcances y los problemas del modelo de regresión lineal.

2.5. Estimación de los parámetros β_0 y β_1

Los coeficientes del modelo se estiman a partir de la muestra aleatoria de n observaciones (X_i, Y_i) con $1 \leq i \leq n$. Llamaremos $\widehat{\beta}_0$ y $\widehat{\beta}_1$ a los estimadores de β_0 y β_1. Los valores $\widehat{\beta}_0$ y $\widehat{\beta}_1$ corresponderán a la recta de ordenada al origen $\widehat{\beta}_0$ y pendiente $\widehat{\beta}_1$ que "mejor ajuste" a los datos $(X_1, Y_1), \ldots, (X_n, Y_n)$ observados. Para encontrarlos, debemos dar una noción de bondad de ajuste de una recta cualquiera con ordenada al origen a y pendiente b a nuestros datos. Tomemos las distancias verticales entre los puntos observados (X_i, Y_i) y los puntos que están sobre la recta $y = a + bx$, que están dados por los pares $(X_i, a + bX_i)$. La distancia entre ambos es $Y_i - (a + bX_i)$. Tomamos como función que mide el desajuste de la recta a los datos a

$$g(a, b) = \sum_{i=1}^{n} (Y_i - (a + bX_i))^2, \tag{8}$$

es decir, la suma de los cuadrados de las distancias entre cada observación y el valor que la recta candidata $y = a + bx$ propone para ajustar dicha observación. Esta expresión puede pensarse como una función g que depende de a y b, y que toma a los valores $(X_1, Y_1), \ldots, (X_n, Y_n)$ como números fijos. Cuánto más cerca esté la recta de ordenada al origen a y pendiente b, menor será el valor de g evaluado en

el par (a, b). Los estimadores de mínimos cuadrados de β_0 y β_1 serán los valores de a y b que minimicen la función g. Para encontrarlos, derivamos esta función con respecto a a y b y luego buscamos los valores $\widehat{\beta}_0$ y $\widehat{\beta}_1$ que anulan sus derivadas. Sus derivadas son

$$\frac{\partial g(a,b)}{\partial a} = \sum_{i=1}^{n} 2\left(Y_i - (a + bX_i)\right)(-1)$$
$$\frac{\partial g(a,b)}{\partial b} = \sum_{i=1}^{n} 2\left(Y_i - (a + bX_i)\right)(-X_i)$$

Las igualamos a cero para encontrar $\widehat{\beta}_0$ y $\widehat{\beta}_1$, sus puntos críticos. Obtenemos

$$\sum_{i=1}^{n} \left(Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right)\right) = 0 \tag{9}$$

$$\sum_{i=1}^{n} \left(Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right)\right) X_i = 0. \tag{10}$$

Las dos ecuaciones anteriores se denominan las *ecuaciones normales* para regresión lineal. Despejamos de ellas las estimaciones de los parámetros que resultan ser

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}, \tag{11}$$

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}. \tag{12}$$

La pendiente estimada también se puede escribir de la siguiente forma

$$\widehat{\beta}_1 = \frac{\frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} = \frac{\widehat{cov}(X,Y)}{\widehat{Var}(X)},$$

es decir, el cociente entre la covarianza muestral y la varianza muestral de las $X's$. Por supuesto, un estudio de las segundas derivadas mostrará (no lo haremos acá) que este procedimiento hace que el par $\widehat{\beta}_0$ y $\widehat{\beta}_1$ no sea sólo un punto crítico, sino también un mínimo. Afortunadamente, en la práctica, los cálculos para hallar a $\widehat{\beta}_0$ y $\widehat{\beta}_1$ son realizados por un paquete estadístico.

Observación 2.1 *La función g propuesta no es la única función de desajuste posible, aunque sí la más difundida. La elección de otra función g para medir el desajuste que proporciona la recta $y = a + bx$ a nuestros datos, como*

$$g(a,b) = \textit{mediana}\left\{[Y_1 - (a + bX_1)]^2, \ldots, [Y_n - (a + bX_n)]^2\right\}$$

da lugar al ajuste, conocido por su nombre en inglés, de ***least median of squares****. Obtendremos distintos estimadores de* β_0 *y* β_1 *que los que se obtienen por mínimos cuadrados. También se utiliza como funcion* g *a la siguiente*

$$g(a,b) = \sum_{i=1}^{n} \rho\left(Y_i - (a + bX_i)\right),$$

donde ρ *es una función muy parecida al cuadrado para valores muy cercanos al cero, pero que crece más lentamente que la cuadrática para valores muy grandes. Estos últimos se denominan* ***M-estimadores de regresión****, y, en general, están programados en los paquetes estadísticos usuales.*

2.6. Recta ajustada, valores predichos y residuos

Una vez que tenemos estimadores para β_0 y β_1 podemos armar la recta ajustada (o modelo ajustado), que es

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

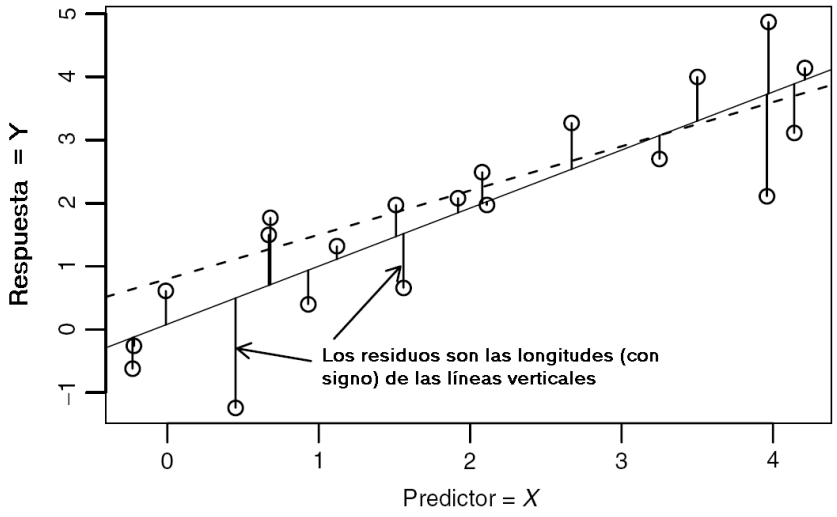
Definición 2.1 *El valor* $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ *calculado para el valor* X_i *observado se denomina (valor)* ***predicho o ajustado i-ésimo****.*

Definición 2.2 *Llamamos* ***residuo de la observación i-ésima*** *a la variable aleatoria*

$$\begin{aligned} e_i &= Y_i - \widehat{Y}_i \\ &= Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i \end{aligned}$$

El residuo i-ésimo representa la distancia vertical entre el punto observado (X_i, Y_i) y el punto predicho por el modelo ajustado, $\left(X_i, \widehat{Y}_i\right)$, como puede observarse en la Figura 15. Los residuos reflejan la inherente asimetría en los roles de las variables predictora y respuesta en los problemas de regresión. Hay herramientas estadísticas distintas para tratar problemas donde no se da esta asimetría, hemos visto el coeficiente de correlación como una de ellas. Las herramientas del análisis multivariado (no se verán en este curso), en general, se abocan a modelar problemas en los que no está presente esta asimetría.

Figura 15: Un gráfico esquemático de ajuste por mínimos cuadrados a un conjunto de datos. Cada par observado está indicado por un círculo pequeño, la línea sólida es la recta ajustada por el método de mínimos cuadrados, la línea punteada es la recta verdadera (desconocida) que dio lugar a los datos. Las líneas verticales entre los puntos y la recta ajustada son los residuos. Los puntos que quedan ubicados bajo la línea ajustada dan residuos negativos, los que quedan por encima dan residuos positivos. Fuente: Weisberg [2005], p.23.



2.6.1. Aplicación al ejemplo

Ajuste con el R Volvamos al ejemplo correspondiente a las mediciones de 100 niños nacidos con bajo peso. El modelo propone que para cada edad gestacional, el perímetro cefálico se distribuye normalmente, con una esperanza que cambia linealmente con la edad gestacional y una varianza fija. Asumimos que las 100 observaciones son independientes. El modelo propuesto es

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

Ajustemos el modelo de regresión lineal simple a los datos. Presentamos la tabla de coeficientes estimados en la Tabla 10.

La recta ajustada a esos datos es

$$\widehat{Y} = 3{,}9143 + 0{,}7801 \cdot X,$$

a veces se anota de la siguiente forma, para enfatizar el nombre de las variables

$$\widehat{headcirc} = 3{,}9143 + 0{,}7801 \cdot gestage.$$

Tabla 10: Coeficientes estimados para el modelo de regresión lineal aplicado a los datos de bebés recién nacidos.

```
> ajuste<-lm(headcirc ~ gestage)
> ajuste
Call:
lm(formula = headcirc ~ gestage)

Coefficients:
(Intercept)      gestage
     3.9143       0.7801
```

Es decir, la ordenada al origen estimada resulta ser 3,9143 y la pendiente de la recta estimada es 0,7801.

Significado de los coeficientes estimados Teóricamente, el valor de la ordenada al origen, es decir, 3,91 es el valor de perímetro cefálico esperado para una edad gestacional de 0 semanas. En este ejemplo, sin embargo, la edad 0 semanas no tiene sentido. La pendiente de la recta es 0,7801, lo cual implica que para cada incremento de una semana en la edad gestacional, el perímetro cefálico del bebé aumenta 0,7801 centímetros en promedio. A veces (no en este caso), tiene más sentido emplear un aumento de la variable explicativa mayor a una unidad, para expresar el significado de la pendiente, esto sucede cuando las unidades de medida de la covariable son muy pequeñas, por ejemplo.

Ahora podemos calcular los valores predichos basados en el modelo de regresión. También podríamos calcular los residuos. Por ejemplo, calculemos el valor predicho de perímetro cefálico medio para un bebé con 25 semanas de gestación (caso $i = 6$, ver los valores observados en la Tabla 11), nuestro valor predicho sería de

$$\widehat{Y}_6 = 3{,}9143 + 0{,}7801 \cdot 25 = 23{,}417$$

y el residuo sería

$$e_6 = Y_6 - \widehat{Y}_6 = 23 - 23{,}417 = -0{,}417$$

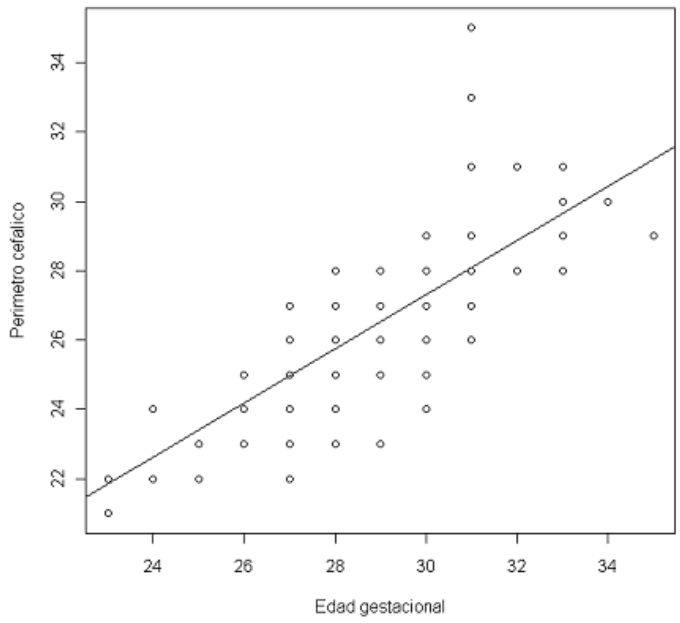
Si quisiéramos predecir el valor del perímetro cefálico medio para un bebé con 29 semanas de gestación ($i = 1$), nuestro valor predicho sería

$$\widehat{Y}_1 = 3{,}9143 + 0{,}7801 \cdot 29 = 26{,}537$$

y el residuo sería

$$e_1 = Y_1 - \widehat{Y}_1 = 27 - 26{,}537 = 0{,}463$$

Figura 16: Gráfico de dispersión del perímetro cefálico versus la edad gestacional, con la recta ajustada por mínimos cuadrados.



Si quisiéramos predecir el valor del perímetro cefálico medio para un bebé con 33 semanas de gestación ($i = 3$), nuestro valor predicho sería

$$\widehat{Y}_3 = 3{,}9143 + 0{,}7801 \cdot 33 = 29{,}658$$

y el residuo sería

$$e_3 = Y_3 - \widehat{Y}_3 = 30 - 29{,}658 = 0{,}342$$

Resumimos esta información en la Tabla 11. Además, en la Figura 16 superponemos al scatter plot la recta estimada por mínimos cuadrados.

Volviendo a la pregunta que motivó la introducción del modelo lineal, si entra una madre con su bebé recién nacido, de bajo peso, al consultorio y quiero predecir su perímetro cefálico, ahora contamos con una herramienta que (confiamos) mejorará nuestra predicción. Le podemos preguntar a la madre la duración de la gestación del niño. Si contesta 25 semanas, predeciré, 23,417 cm. de perímetro cefálico; si contesta 29 semanas, predeciré 26,537, si contesta 33 semanas, predeciré 29,658. Si dice x_0 semanas, diremos $3{,}9143 + 0{,}7801 \cdot x_0$ cm. ¿Qué error tiene

Tabla 11: Tres datos de los bebés de bajo peso analizados en el texto, con el valor predicho y el residuo respectivo

Caso (i)	$Y_i =$ (`headcirc`)	$X_i =$ (`gestage`)	$\widehat{Y}_i$ (predicho)	e_i (residuo)
1	27	29	26,537	0,463
3	30	33	29,658	0,342
6	23	25	23,417	−0,417

esta predicción? Para contestar a esta pregunta, tenemos que estimar la varianza condicional de Y, es decir, σ^2.

2.7. Ejercicios (primera parte)

Estos ejercicios se resuelven con el `script_reglinealsimple1.R`

Ejercicio 2.1 *Medidas del cuerpo, Parte II. Datos publicados en Heinz, Peterson, Johnson, y Kerk [2003], base de datos* `bdims` *del paquete* `openintro`.

(a) Realizar un diagrama de dispersión que muestre la relación entre el peso medido en kilogramos (`wgt`) y la circunferencia de la cadera medida en centímetros (`hip.gi`), ponga el peso en el eje vertical. Describa la relación entre la circunferencia de la cadera y el peso.

(b) ¿Cómo cambiaría la relación si el peso se midiera en libras mientras que las unidades para la circunferencia de la cadera permanecieran en centímetros?

(c) Ajuste un modelo lineal para explicar el peso por la circunferencia de cadera, con las variables en las unidades originales. Escriba el modelo (con papel y lápiz, con betas y epsilones). Luego, escriba el modelo ajustado (sin epsilones). Interprete la pendiente estimada en términos del problema. Su respuesta debería contener una frase que comience así: "Si una persona aumenta un cm. de contorno de cadera, en promedio su peso aumentará ... kilogramos".

(d) Superponga la recta ajustada al scatterplot. Observe el gráfico. ¿Diría que la recta describe bien la relación entre ambas variables?

(e) Elegimos una persona adulta físicamente activa entre los estudiantes de primer año de la facultad. Su contorno de cadera mide 100 cm. Prediga su peso en kilogramos.

(f) Esa persona elegida al azar pesa 81kg. Calcule el residuo.

(g) Estime el peso esperado para la población de adultos cuyo contorno de cadera mide 100 cm.

Ejercicio 2.2 *Medidas del cuerpo, Parte III. Base de datos* `bdims` *del paquete* `openintro`.

(a) Realizar un diagrama de dispersión que muestre la relación entre el peso medido en kilogramos (`wgt`*) y la altura (*`hgt`*).*

(b) Ajuste un modelo lineal para explicar el peso por la altura. Escriba el modelo (con papel y lápiz, con betas y epsilones). Luego, escriba el modelo ajustado (sin epsilones). Interprete la pendiente estimada en términos del problema. Interprete la pendiente. ¿Es razonable el signo obtenido para la pendiente estimada? Superponer al scatterplot anterior la recta estimada.

(c) La persona elegida en el ejercicio anterior, medía 187 cm. de alto, y pesaba 81 kg. Prediga su peso con el modelo que tiene a la altura como covariable. Calcule el residuo de dicha observación.

Ejercicio 2.3 *Mamíferos, Parte III. Base de datos* `mammals` *del paquete* `openintro`.

(a) Queremos ajustar un modelo lineal para predecir el peso del cerebro de un mamífero (`BrainWt`*) a partir del peso corporal (*`BodyWt`*) del animal. Habíamos visto en el Ejercicio 1.7 que si graficamos el peso del cerebro en función del peso corporal, el gráfico era bastante feo. Y que todo mejoraba tomando logaritmo (en cualquier base, digamos base 10) de ambas variables. Ajuste un modelo lineal para explicar a* log_{10} (`BrainWt`) *en función del* log_{10} (`BodyWt`). *Como antes, escriba el modelo teórico y el ajustado. Una observación: en el help del* `openitro` *se indica que la variable* `BrainWt` *está medida en kg., sin embargo, esta variable está medida en gramos.*

(b) Repita el scatterplot de las variables transformadas y superpóngale la recta ajustada.

(c) La observación 45 corresponde a un chancho. Prediga el peso del cerebro del chancho con el modelo ajustado, sabiendo que pesa 192 kilos. Recuerde transformar al peso corporal del chancho antes de hacer cálculos. Marque esa observación en el gráfico, con color violeta.

(d) La observación 34 corresponde a un ser humano. Prediga el peso del cerebro de un ser humano con el modelo ajustado, sabiendo que pesa 62 kilos. Recuerde transformar al peso corporal del chancho antes de hacer cálculos. Marque esa observación en el gráfico, con color rojo.

Ejercicio 2.4 *Resuelva (en clase) el Taller 1 que figura en el Apéndice A.*

2.8. Estimación de σ^2

Escribamos nuevamente el modelo poblacional y el modelo ajustado

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i, && \text{Modelo poblacional} \\ \widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i, && \text{Modelo ajustado} \end{aligned} \tag{13}$$

Si queremos hacer aparecer el valor observado Y_i a la izquierda en ambos, podemos escribir el modelo ajustado de la siguiente forma

$$Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + e_i, \qquad \text{Modelo ajustado.} \tag{14}$$

ya que los residuos se definen por $e_i = Y_i - \widehat{Y}_i$. El error iésimo (ε_i) es la variable aleatoria que representa la desviación (vertical) que tiene el i-ésimo par observado (X_i, Y_i) respecto de la recta poblacional o teórica que asumimos es el modelo correcto para nuestros datos (ecuación (13)). El residuo i-ésimo (e_i) , en cambio, es la variable aleatoria que representa la desviación (vertical) que tiene el i-ésimo par observado (X_i, Y_i) respecto de la recta ajustada que calculamos en base a nuestros datos (ecuación (14)). Recordemos que uno de los supuestos del modelo es que la varianza de los errores es σ^2, $Var(\varepsilon_i) = \sigma^2$. Si pudiéramos observar los errores, entonces podríamos construir un estimador de la varianza a partir de ellos, que sería

$$\frac{1}{n-1} \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2 .$$

Pero los errores (ε_i) no son observables, lo que podemos observar son su correlato empírico, los residuos (e_i). Desafortunadamente, el residuo i-ésimo no es una estimación del error i-ésimo: en estadística sabemos estimar números fijos que llamamos parámetros. El error, sin embargo, es una variable aleatoria, así que no lo podemos estimar. Tanto los ε_i como los e_i son variables aleatorias, pero muchas de las cualidades de los errores no las heredan los residuos. Los errores ε_i son independientes, pero los residuos e_i no lo son. De hecho, suman 0. Esto puede verse si uno escribe la primera ecuación normal que vimos en la Sección 2.5, la ecuación (9) en términos de los e_i

$$0 = \sum_{i=1}^{n} \left(Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i \right) \right) = \sum_{i=1}^{n} e_i. \tag{15}$$

Luego, $\overline{e} = \sum_{i=1}^{n} e_i = 0$. Si escribimos la segunda ecuación normal en términos de los residuos vemos también que

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \left(Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right)\right) X_i \\ &= \sum_{i=1}^{n} e_i X_i = \sum_{i=1}^{n} (e_i - \overline{e}) X_i = \sum_{i=1}^{n} (e_i - \overline{e}) \left(X_i - \overline{X}\right) \end{aligned} \tag{16}$$

La segunda igualdad de (16) se debe a que por (15) el promedio de los residuos $\overline{e}$, es igual a cero, y la tercera puede verificarse haciendo la distributiva correspondiente. Observemos que si calculamos el coeficiente de correlación muestral entre las X_i y los e_i, el numerador de dicho coeficiente es el que acabamos de probar que vale 0, es decir,

$$r = r\left((X_1, e_i), \ldots, (X_n, e_n)\right) = \frac{\sum_{i=1}^{n} (e_i - \overline{e}) \left(X_i - \overline{X}\right)}{\sqrt{\sum_{i=1}^{n} (e_i - \overline{e})^2} \sqrt{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}} = 0.$$

Luego, los residuos satisfacen dos ecuaciones lineales (las dadas por (15) y (16)) y por lo tanto, tienen más estructura que los errores. Además, los errores tienen todos la misma varianza, pero los residuos no. Más adelante las calcularemos.

El estimador de σ^2 que usaremos será

$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (e_i - \overline{e})^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2. \tag{17}$$

Al numerador de la expresión anterior ya lo encontramos cuando hallamos la solución al problema de mínimos cuadrados: es la suma de los cuadrados de los residuos que notamos también por SSRes donde las siglas vienen de la expresión en inglés (*residual sum of squares*). De él deviene otra manera de nombrar al estimador de σ^2 : MSRes, es decir, *mean squared residuals*, o cuadrado medio de los residuos:

$$\widehat{\sigma}^2 = \frac{1}{n-2} \text{SSRes} = \text{MSRes}.$$

Hallémoslo en el caso del ejemplo. Para ello, vemos una tabla más completa de la salida del `R` en la Tabla 12. Más adelante analizaremos en detalle esta salida, por ahora sólo nos interesa la estimación de σ que resulta ser 1,59, indicada por `Residual standard error`. Luego la estimación de σ^2 que proporciona el modelo lineal es su cuadrado. Si comparamos la estimación de σ con la obtenida sin el modelo de regresión, cuando sólo disponíamos de la variable Y, vemos que el desvío estándar se redujo considerablemente (el desvío estándar muestral de las Y's es 2,53, ver la Tabla 6). Esta información además nos permite proponer tests e intervalos de confianza para β_0 y β_1.

Tabla 12: Salida del ajuste de regresión lineal, con p-valores, para los 100 bebés de bajo peso.

```
> ajuste<-lm(headcirc~gestage)
> summary(ajuste)

Call:
lm(formula = headcirc ~ gestage)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5358 -0.8760 -0.1458  0.9041  6.9041

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.91426    1.82915    2.14   0.0348 *
gestage      0.78005    0.06307   12.37   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 1.59 on 98 degrees of freedom
Multiple R-squared:  0.6095,        Adjusted R-squared:  0.6055
F-statistic: 152.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

2.9. Inferencia sobre β_1

Intentaremos construir un intervalo de confianza y tests para β_1, la pendiente de la recta del modelo lineal poblacional o teórico que describe a la población de la que fueron muestreados nuestros datos. Recordemos que el modelo lineal es un modelo para la esperanza condicional de Y conocidos los valores de la variable X. La estimación y la inferencia se realizan bajo este contexto condicional. Para hacer inferencias, tomaremos las X_i como constantes, para no escribir oraciones del estilo $E\left(\widehat{\beta}_1 \mid X_1, \ldots, X_n\right)$. Para el estimador $\widehat{\beta}_1$ puede probarse que, si los datos siguen el modelo lineal (2), es decir, si

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon.$$

con los supuestos que hemos descrito (homoscedasticidad, independencia y normalidad de los errores), entonces

$$\begin{aligned} E\left(\widehat{\beta}_1\right) &= \beta_1 \\ Var\left(\widehat{\beta}_1\right) &= \frac{\sigma^2}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}. \end{aligned}$$

y[2] también

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\right).$$

Un estimador de la varianza es

$$\widehat{Var}\left(\widehat{\beta}_1\right) = \frac{\widehat{\sigma}^2}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2} = \frac{\text{SSRes}/\left(n-2\right)}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}.$$

Finalmente, bajo los supuestos del modelo, puede probarse que

$$\frac{\widehat{\beta}_1-\beta_1}{\sqrt{\widehat{Var}\left(\widehat{\beta}_1\right)}} = \frac{\widehat{\beta}_1-\beta_1}{se_{\widehat{\beta}_1}}$$

tiene distribución *t de Student con $n-2$ grados de libertad* si los datos siguen el modelo lineal, donde

$$se_{\widehat{\beta}_1} = \sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}} = \sqrt{\frac{\text{SSRes}/\left(n-2\right)}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}}.$$

Esto es lo que se conoce como la distribución de muestreo de $\widehat{\beta}_1$. Los grados de libertad son $n-2$ puesto que los residuos satisfacen dos ecuaciones lineales, es decir, conociendo $n-2$ de ellos se pueden reconstruir los dos restantes. A la raíz cuadrada de una varianza estimada se la llama error estándar (*standard error*), por lo que usamos el símbolo $se_{\widehat{\beta}_1}$ para el error estándar de $\widehat{\beta}_1$, o sea $se_{\widehat{\beta}_1}$ es un estimador de la desviación estándar de la distribución de muestreo de $\widehat{\beta}_1$.

Con esta distribución podemos construir un intervalo de confianza de nivel $1-\alpha$ para β_1 que resultará

$$\begin{aligned} &\widehat{\beta}_1 \pm t_{n-2;1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}}, \text{ o bien} \\ &\widehat{\beta}_1 \pm t_{n-2;1-\frac{\alpha}{2}} \cdot se_{\widehat{\beta}_1} \end{aligned} \tag{18}$$

[2] En realidad, $E\left(\widehat{\beta}_1\right) = \beta_1$ y $Var\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}$, pero omitiremos este nivel de detalle en adelante.

donde $t_{n-2,1-\frac{\alpha}{2}}$ es el percentil $1-\frac{\alpha}{2}$ de la distribución t_{n-2} (el valor que deja a su izquierda un área $1-\frac{\alpha}{2}$)[3]. Esto también permite realizar tests para la pendiente. La forma general de las hipótesis para estos tests es

$$\begin{aligned} H_0 &: \beta_1 = b \\ H_1 &: \beta_1 \neq b. \end{aligned}$$

donde b es un valor fijado de antemano. Sin embargo, el test de mayor interés para el modelo lineal es el que permite decidir entre estas dos hipótesis

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0, \end{aligned} \tag{19}$$

(es decir, tomar $b = 0$ como caso particular). Si $\beta_1 = 0$, las Y_i no dependen de las X_i, es decir, no hay asociación lineal entre X e Y, en cambio, la hipótesis alternativa indica que sí hay un vínculo lineal entre ambas variables. Para proponer un test, debemos dar la distribución de un estadístico basado en el estimador bajo la hipótesis nula. En este caso resulta que, bajo H_0, $Y_i \mid X_i \sim N(\beta_0, \sigma^2)$, es decir, son variables aleatorias independientes e idénticamente distribuidas. Como además el estimador de β_1 (y también el de β_0) puede escribirse como una combinación lineal de los Y_i :

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{j=1}^n \left(X_j - \overline{X}\right)^2} = \frac{1}{\sum_{j=1}^n \left(X_j - \overline{X}\right)^2} \sum_{i=1}^n \left(X_i - \overline{X}\right) Y_i \\ &= \sum_{i=1}^n \frac{\left(X_i - \overline{X}\right)}{\sum_{j=1}^n \left(X_j - \overline{X}\right)^2} Y_i = \sum_{i=1}^n c_i Y_i \end{aligned} \tag{20}$$

donde

$$\begin{aligned} c_i &= \frac{\left(X_i - \overline{X}\right)}{\sum_{j=1}^n \left(X_j - \overline{X}\right)^2} = \frac{\left(X_i - \overline{X}\right)}{S_{XX}}, \\ S_{XX} &= \sum_{j=1}^n \left(X_j - \overline{X}\right)^2. \end{aligned} \tag{21}$$

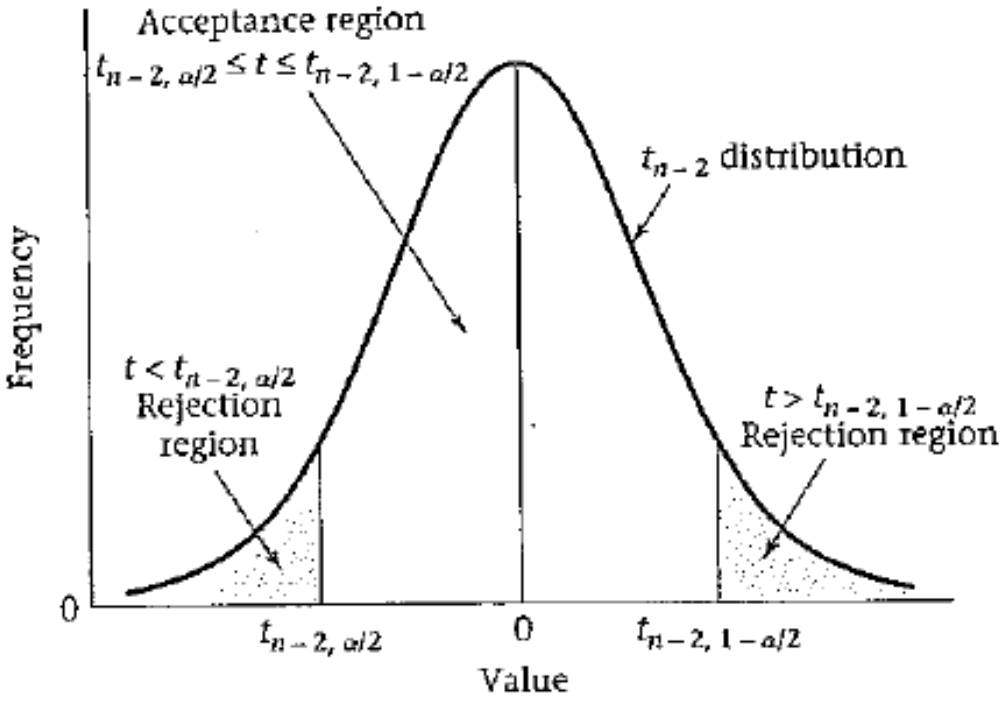
Entonces, la distribución de $\widehat{\beta}_1$ será normal. Si el supuesto de normalidad de los errores no valiera, pero la muestra fuera suficientemente grande, y se cumpliera una condición sobre los c_i la distribución de $\widehat{\beta}_1$ seguiría siendo aproximadamente normal. Luego, si $\beta_1 = 0$ el estadístico T descripto a continuación

$$T = \frac{\widehat{\beta}_1 - 0}{\sqrt{\widehat{Var}\left(\widehat{\beta}_1\right)}} = \frac{\widehat{\beta}_1}{se_{\widehat{\beta}_1}} = \frac{\widehat{\beta}_1}{\sqrt{\frac{\text{SSRes}}{(n-2)\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}}} \tag{22}$$

[3]En `R` al percentil $t_{n-2,1-\frac{\alpha}{2}}$ lo encontramos con el comando $\texttt{qt}\left(1-\frac{\alpha}{2}, \texttt{df} = n-2\right)$.

tiene distribución t_{n-2}. Finalmente, un test de nivel α para las hipótesis (19) rechazará H_0 cuando el valor de T observado en la muestra sea mayor que el percentil $1-\frac{\alpha}{2}$ de la distribución t_{n-2}, es decir, $t_{n-2,1-\frac{\alpha}{2}}$, o menor que $t_{n-2,\frac{\alpha}{2}} = -t_{n-2,1-\frac{\alpha}{2}}$, según la Figura 17.

Figura 17: Región de rechazo y aceptación para el test t para la pendiente del modelo lineal simple, se grafica la densidad de una t de Student con $n-2$ grados de libertad. Fuente Rosner [2006], pág. 442.



Es decir, el test rechaza H_0 con nivel α si

$$T_{obs} \le t_{n-2,\frac{\alpha}{2}} \text{ ó } t_{n-2,1-\frac{\alpha}{2}} \le T_{obs},$$

donde T_{obs} es el valor del estadístico T definido en (22) calculado en base a las observaciones $(X_1, Y_1), \ldots, (X_n, Y_n)$. O bien, se puede calcular el $p-valor$ del test de la siguiente forma

$$p - valor = 2P\left(T \ge |T_{obs}|\right),$$

ya que se trata de un test a dos colas. Reportar el p-valor cuando uno realiza un test sobre un conjunto de datos siempre permite al lector elegir su punto de corte respecto de aceptar o rechazar una hipótesis.

Un comentario final. Hay una importante distinción entre significatividad estadística, es decir, la observación de un p-valor suficientemente pequeño, y la significatividad científica (médica, biológica, económica, dependiendo del contexto) en el hecho de considerar significativo un efecto de una cierta magnitud. La significatividad científica requerirá examinar, en la mayoría de las aplicaciones, el contexto, la evidencia científica existente, las magnitudes de las variables relacionadas, el estado del arte en el tema en cuestión, más que sólo un p-valor.

2.9.1. Aplicación al ejemplo

Para el ejemplo de los 100 bebés de bajo peso, si volvemos a mirar la tabla de coeficientes estimados (Tabla 12) obtenemos el estimador del error estándar de $\widehat{\beta}_1$, o sea

$$se_{\widehat{\beta}_1} = 0{,}063.$$

Otra forma de obtener este valor es a partir de la Tabla 12 y la Figura 18. De la primera obtenemos que

$$\text{SSRes}/\,(n-2) = 247{,}883/98 = 2{,}529$$

Figura 18: Estadísticos descriptivos para la edad gestacional

	N	Mínimo	Máximo	Media	Desv. típ.
Edad gestacional (semanas)	100	23	35	28,89	2,534
N válido (según lista)	100				

En la segunda vemos que $\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{n-1}} = 2{,}534$ que es el desvío estándar muestral de las $X's$. De aquí obtenemos

$$S_{XX} = \sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 = 2{,}534^2\,(n-1) = 2{,}534^2\,(99) = 635{,}69$$

Finalmente,

$$\begin{aligned} se_{\widehat{\beta}_1} &= \sqrt{\frac{\text{SSRes}/(n-2)}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}} = \sqrt{\frac{247{,}883/98}{635{,}69}} \\ &= \sqrt{\frac{2{,}529418}{635{,}69}} = 0{,}06307941 \end{aligned}$$

El percentil resulta ser $t_{n-2;1-\frac{\alpha}{2}} = t_{98;0,975} = 1{,}984467$. Luego, un intervalo de confianza de nivel $0{,}95 = 1-\alpha$ para β_1 será

$$\begin{aligned} &\widehat{\beta}_1 \pm t_{n-2;1-\frac{\alpha}{2}} \cdot se_{\widehat{\beta}_1} \\ &0{,}7801 \pm 1{,}984467 \cdot 0{,}06307941 \\ &[0{,}654921, 0{,}905279] \end{aligned}$$

Es decir, como el intervalo está íntegramente contenido en los reales positivos, el verdadero valor de la pendiente, β_1, será positivo, confirmando que la asociación positiva que encontramos en la muestra se verifica a nivel poblacional. Observemos también que el intervalo es bastante preciso, esto se debe a que la muestra sobre la que sacamos las conclusiones es bastante grande. Notemos que la variabilidad de $\widehat{\beta}_1$ disminuye (la estimación es más precisa o el intervalo de confianza más pequeño), ver la expresión (18) cuando:

- La varianza de los errores σ^2 disminuye.
- La varianza muestral de la variable regresora aumenta, o sea, mientras más amplio el rango de valores de la covariable, mayor la precisión en la estimación de la pendiente.
- El tamaño de muestra aumenta. Para ver el efecto de aumentar el n, podemos escribir

$$se^2_{\widehat{\beta}_1} = \frac{\widehat{\sigma}^2}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2} = \frac{\widehat{\sigma}^2}{\left[\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2/(n-1)\right]} \cdot \frac{1}{(n-1)}.$$

  El primer factor convergerá a $\frac{\sigma^2}{Var(X)}$ si las X‘s son una muestra al azar. Como el segundo factor tiende a cero, el producto tiende a cero al aumentar el n.

Si en vez del intervalo de confianza queremos hacer un test de nivel 0,05 para las hipótesis siguientes

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0, \end{aligned}$$

entonces en la Tabla 12 vemos calculado el estadístico del test $T = 12{,}367$ que se obtuvo al dividir el estimador de β_1 por el estimador del desvío estándar del estimador de β_1 :

$$T_{obs} = \frac{\widehat{\beta}_1}{se_{\widehat{\beta}_1}} = \frac{0{,}7801}{0{,}06307941} = 12{,}36695.$$

Para decidir la conclusión del test debemos comparar el valor T_{obs} con el percentil $t_{n-2;1-\frac{\alpha}{2}} = t_{98,0{,}975} = 1{,}984467$. Como claramente $T_{obs} = 12{,}367 > t_{98,0{,}975} = 1{,}984$, entonces rechazamos H_0, concluyendo que el parámetro poblacional que mide la pendiente del modelo lineal es distinto de cero. Como sabemos, una forma alternativa de llevar a cabo este test es calcular el $p-valor$, que en este caso será

$$p - valor = 2P\left(T > T_{obs}\right) = 2P\left(T > 12{,}367\right) \simeq 0$$

como figura en la última columna de la Tabla 12. Como $p-valor < 0{,}05$, se rechaza la hipótesis nula.

Observemos que el intervalo de confianza para β_1 construido en base a los datos es más informativo que el test, ya que nos permite decir que para los tests de hipótesis

$$\begin{aligned} H_0 &: \beta_1 = b \\ H_1 &: \beta_1 \neq b. \end{aligned}$$

la hipótesis nula será rechazada para todo b fijo que no quede contenido en el intervalo $[0{,}655, 0{,}905]$ en base a la muestra observada (esto es lo que se conoce como dualidad entre intervalos de confianza y tests).

En la Tabla 13 pueden verse los intervalos de confianza para ambos parámetros calculados en `R`.

Tabla 13: Intervalos de confianza de nivel 0,95 para los coeficientes lineales del ajuste en `R`, para los 100 bebés de bajo peso.

```
> ajuste<-lm(headcirc~gestage)
> confint(ajuste, level = 0.95)
                2.5 %    97.5 %
(Intercept) 0.2843817 7.5441466
gestage     0.6548841 0.9052223
```

2.10. Inferencia sobre β_0

Esta inferencia despierta menos interés que la de β_1. Aunque los paquetes estadísticos la calculan es infrecuente encontrarla en aplicaciones. Bajo los supuestos del modelo lineal, puede calcularse la esperanza y varianza del estimador de β_0, que resultan ser

$$\begin{aligned} E\left(\widehat{\beta}_0\right) &= \beta_0 \\ Var\left(\widehat{\beta}_0\right) &= \sigma^2\left(\frac{1}{n}+\frac{\overline{X}^2}{\sum_{j=1}^{n}\left(X_j-\overline{X}\right)^2}\right). \end{aligned}$$

Nuevamente, las conclusiones son condicionales a los valores de los X's observados. La varianza puede estimarse por

$$\widehat{Var}\left(\widehat{\beta}_0\right)=\widehat{\sigma}^2\left(\frac{1}{n}+\frac{\overline{X}^2}{\sum_{j=1}^{n}\left(X_j-\overline{X}\right)^2}\right)$$

Nuevamente, el estadístico $\widehat{\beta}_0$ tiene distribución normal, su distribución es $N\left(\beta_0,\sigma^2\left(\frac{1}{n}+\frac{\overline{X}^2}{\sum_{j=1}^{n}\left(X_j-\overline{X}\right)^2}\right)\right)$, luego

$$\frac{\widehat{\beta}_0-\beta_0}{\sqrt{Var\left(\widehat{\beta}_0\right)}}\sim t_{n-2}$$

y el intervalo de confianza para la ordenada al origen resulta ser

$$\widehat{\beta}_0 \pm t_{n-2;\frac{\alpha}{2}}\cdot\widehat{\sigma}\cdot\sqrt{\frac{1}{n}+\frac{\overline{X}^2}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}} \tag{23}$$

Esto quiere decir que el $(1-\alpha)\cdot 100$ por ciento de los intervalos construidos de esta forma contendrán al verdadero valor β_0 con el que fueron generados los datos.

Ejemplo 2.3 *Para el ejemplo de los 100 bebés vemos en la Tabla 12 que el estadístico T observado en este caso vale* 2,14 *y el p-valor para testear*

$$\begin{aligned} H_0 &: \beta_0 = 0 \\ H_1 &: \beta_0 \neq 0, \end{aligned}$$

es 0,035, *indicando que se rechaza la H_0 y la ordenada al origen poblacional es no nula. También en la Tabla 13 puede observarse el intervalo de confianza de nivel* 0,95 *para β_0 que resulta ser* $[0{,}284, 7{,}544]$.

2.11. Intervalo de confianza para la respuesta media de Y cuando $X = x_h$

Nos interesa construir un intervalo de confianza para $E(Y_h \mid X = x_h)$ que escribiremos $E(Y_h)$, es decir, un intervalo de confianza para la respuesta media para algun valor prefijado de la covariable en x_h. Observemos que x_h, el nivel de X para el que queremos estimar la respuesta media puede o no ser un valor observado en la muestra (pero siempre tiene que estar dentro del rango de valores observados para X, es decir, entre el mínimo y máximo valor observado para X). El parámetro poblacional a estimar es, entonces

$$E(Y_h \mid X = x_h) = \beta_0 + \beta_1 x_h.$$

El estimador puntual está dado por

$$\widehat{Y}_h = \widehat{\beta}_0 + \widehat{\beta}_1 x_h.$$

La esperanza y la varianza de dicho estimador son

$$\begin{aligned} E\left(\widehat{Y}_h\right) &= E(Y_h) \\ Var\left(\widehat{Y}_h\right) &= \sigma^2 \cdot \left[\frac{1}{n} + \frac{\left(x_h - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right]. \end{aligned}$$

Observemos que la variabilidad de nuestra estimación de Y_h se ve afectada esencialmente por dos componentes:

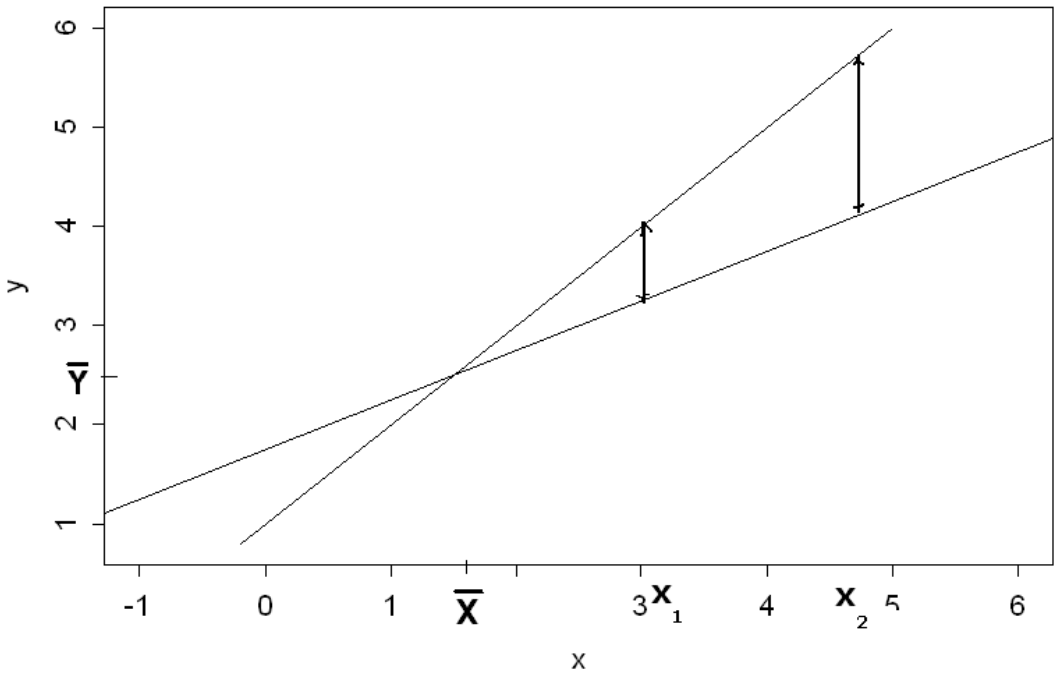
- por σ^2, la variabilidad de las $Y's$ cuando conocemos el valor de X,
- y por cuan lejos está ese valor particular de x_h (de X) del promedio observado en la muestra $\overline{X}$.

Notar que la variabilidad de la estimación de $E(Y_h \mid X = x_h)$ será menor cuanto más cercano a la media muestral $\overline{X}$ esté el valor de x_h que nos interesa. Esto puede tomarse en cuenta en los (raros) casos en los cuales los valores de $X_1, \ldots, X_n$ son fijados por el experimentador. Esto último se debe a que, por construcción, la recta de mínimos cuadrados **siempre** pasa por el punto $\left(\overline{X}, \overline{Y}\right)$, ya que

$$\widehat{\beta}_0 + \widehat{\beta}_1 \overline{X} = \underbrace{\overline{Y} - \widehat{\beta}_1 \overline{X}}_{\widehat{\beta}_0} + \widehat{\beta}_1 \overline{X} = \overline{Y}$$

Luego, si tomamos muchas muestras de observaciones $(X_1, Y_1), \ldots, (X_n, Y_n)$ con los mismos valores $X_1, \ldots, X_n$, resultará que el valor $\overline{X}$ no variará, y el valor

Figura 19: Dos rectas ajustadas por mínimos cuadrados para dos muestras con los mismos X_i, ambas pasan por el mismo $\left(\overline{X}, \overline{Y}\right)$, se observa la variabilidad mayor en el valor predicho (o ajustado) para $E\left(Y \mid X = x_2\right)$ que para $E\left(Y \mid X = x_1\right)$ si la distancia al $\overline{X}$ es mayor para x_2 que para x_1.



$\overline{Y}$ será parecido en las diversas muestras. Todas las rectas ajustadas por mínimos cuadrados pasarán por sus respectivos centros $\left(\overline{X}, \overline{Y}\right)$, que al no diferir demasiado en su valor en $\overline{Y}$, darán una estimación más precisa de $E\left(Y_h \mid X = x_h\right)$ cuando x_h esté cerca de $\overline{X}$ que cuando esté lejos, ver la Figura 19.

A partir de la definición de $\widehat{\beta}_0$, y las ecuaciones (20) y (21), podemos escribir

$$\begin{aligned}\widehat{Y}_h &= \widehat{\beta}_0 + \widehat{\beta}_1 x_h \\ &= \overline{Y} - \widehat{\beta}_1 \overline{X} + \widehat{\beta}_1 x_h \\ &= \overline{Y} + \widehat{\beta}_1 \left(x_h - \overline{X}\right) \\ &= \sum_{i=1}^{n} \frac{1}{n} Y_i + \sum_{i=1}^{n} c_i Y_i \left(x_h - \overline{X}\right) \\ &= \sum_{i=1}^{n} \left[\frac{1}{n} + c_i \left(x_h - \overline{X}\right)\right] Y_i\end{aligned}$$

con $c_i = \frac{(X_i - \overline{X})}{S_{XX}}$. De la normalidad de los errores se deduce la normalidad de $\widehat{Y}_h$. Luego, un intervalo de confianza (que abreviaremos IC) de nivel $1-\alpha$ para $E(Y_h)$ resulta ser

$$\widehat{Y}_h \pm t_{n-2;1-\frac{\alpha}{2}} \cdot \widehat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{\left(x_h - \overline{X}\right)^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}}. \tag{24}$$

2.12. Intervalo de Predicción de una nueva observación Y medida cuando $X = x_h$

Consideramos ahora el problema de predecir una nueva observación Y correspondiente a un nivel de X dado.

En el ejemplo de los bebés nacidos con bajo peso, queremos predecir el perímetro cefálico de un bebé que tiene 29 semanas de gestación (y sabemos que nació con bajo peso).

Esta nueva observación debe ser obtenida en forma independiente de las observaciones $(X_i, Y_i)_{1 \leq i \leq n}$ en las cuales se basó la estimación de la recta de regresión. En el caso del ejemplo se trata de predecir el perímetro cefálico de un bebé que no está entre los 100 bebés sobre los cuales se basó el ajuste de la regresión.

Denotamos por x_h el valor de X y por $Y_{h(nuevo)}$ al valor de Y. A diferencia del intervalo de confianza (IC) para $E(Y_h)$ que hallamos antes, ahora predecimos un **resultado individual** proveniente de la distribución de Y, o sea, tenemos ahora dos fuentes de variabilidad:

- la incerteza en la estimación de $E(Y_h)$ alrededor de la cual yacerá la nueva observación

- la variabilidad de Y alrededor de su media (que deviene de su distribución).

Lo que queremos es un intervalo de extremos aleatorios $[a_n, b_n]$ tal que

$$P\left(a_n \leq Y_{h(nuevo)} \leq b_n\right) = 1 - \alpha.$$

Enfaticemos la diferencia entre ambos procedimientos.

Estimación (es decir, el cálculo del intervalo de confianza para la esperanza de Y condicional al valor de X, $E\left(Y_h \mid X = x_h\right)$): Es una regla para calcular a partir de los datos un valor que nos permita "adivinar" el valor que puede tomar un **parámetro poblacional**, en este caso, la esperanza de Y cuando la variable X toma el valor x_h. En el ejemplo, el parámetro es el perímetro cefálico medio de todos los bebés de bajo peso con x_h (por ejemplo, 29) semanas de gestación.

Predicción (es decir, el cálculo del intervalo de predicción de una nueva observación $Y_{h(nueva)}$ medida cuando $X = x_h$): Es una regla para calcular a partir de los datos un valor que nos permita "adivinar" el valor que puede tomar una **variable aleatoria**.

Nuestra mejor predicción es nuevamente

$$\widehat{Y}_h = \widehat{\beta}_0 + \widehat{\beta}_1 x_h,$$

pero ahora el error asociado será mayor. Estimamos el **error estándar** de la **predicción** con

$$\widehat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(x_h - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}$$

A partir de este error estándar podemos construir un intervalo de predicción (que abreviaremos IP) de nivel $(1 - \alpha)$ para el valor predicho de Y cuando $X = x_h$ por

$$\widehat{Y}_h \pm t_{n-2;1-\frac{\alpha}{2}} \cdot \widehat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(x_h - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}$$

2.12.1. Aplicación al ejemplo

Calculemos los intervalos de confianza de nivel 0,95 para $\mathrm{E}(Y_h \mid X = x_h)$ y de predicción para una nueva observación Y_h realizada cuando $X = x_h$, para algunos valores considerados previamente (y otros más) en el ejemplo de los 100 niños de bajo peso al nacer: (recordemos que $X =$ edad gestacional, $Y =$ perímetro cefálico)

En R: Creamos un vector con todos los valores de x_h para los cuales queremos hallar los intervalos de confianza, lo llamamos `xx` en la lista de comandos que sigue. En `R`, el nivel de los intervalos, por default, es 0.95.

$X = x_h$	$\widehat{Y}_h$	Intervalo de confianza		Longitud del IC
23	21.86	[21.05	22.66]	1.60
25	23.42	[22.84	24.00]	1.16
28	25.76	[25.42	26.09]	0.67
29	26.54	[26.22	26.85]	0.63
33	29.66	[29.05	30.26]	1.21
35	31.22	[30.39	32.04]	1.65

$X = x_h$	$\widehat{Y}_h$	Intervalo de predicción		Longitud del IP
23	21.86	[18.60	25.11]	6.51
25	23.42	[20.21	26.62]	6.42
28	25.76	[22.58	28.93]	6.35
29	26.54	[23.36	29.71]	6.34
33	29.66	[26.44	32.87]	6.43
35	31.22	[27.95	34.48]	6.53

```
> ajuste<-lm(headcirc~gestage)
> xx<-c(23,25,28,29,33,35)
> IC<-predict(ajuste,newdata=data.frame(gestage=xx),
interval="confidence",level=0.95)
> IP<-predict(ajuste,newdata=data.frame(gestage=xx),
interval="prediction",level=0.95)
> IC
       fit      lwr      upr
1 21.85549 21.05352 22.65745
2 23.41559 22.83534 23.99584
3 25.75575 25.42106 26.09045
4 26.53581 26.21989 26.85172
5 29.65602 29.05247 30.25956
6 31.21612 30.38878 32.04347
> IP
       fit      lwr      upr
1 21.85549 18.59907 25.11190
2 23.41559 20.20657 26.62461
3 25.75575 22.58193 28.92957
4 26.53581 23.36391 29.70770
5 29.65602 26.44271 32.86933
```

Hagamos las cuentas en detalle para $x_h = 29$. Sabemos que $\widehat{Y}_h = 26{,}537$. La

teoría nos dice que el IC de nivel 0,95 para $E(Y_h \mid X = x_h)$ se obtiene por

$$\widehat{Y}_h \pm t_{n-2;1-\frac{\alpha}{2}} \cdot \widehat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{\left(x_h - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}$$

Sabemos que (ver los estadísticos descriptivos de la edad gestacional) calculados en la Sección 2.9

$$\begin{aligned} \overline{X} &= 28{,}89 \\ S_{XX} &= 635{,}69 \\ n &= 100 \end{aligned}$$

La varianza estimada por la regresión es

$$\widehat{\sigma}^2 = \frac{SS\text{Res}}{n-2} = 2{,}529$$

de dónde surge

$$\widehat{\sigma} = s = \sqrt{2{,}529} = 1{,}5903$$

y

$$t_{n-2;1-\frac{\alpha}{2}} = t_{98;0,975} = 1{,}984467.$$

Luego, el intervalo de confianza de nivel 0,95 para $E(Y_h \mid X = 29)$ se obtiene por

$$\widehat{Y}_h \pm t_{n-2;\frac{\alpha}{2}} \cdot \widehat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x_h - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

$$26{,}537 \pm 1{,}984467 \cdot 1{,}5903 \cdot \sqrt{\frac{1}{100} + \frac{(29 - 28{,}89)^2}{635{,}69}}$$

$$26{,}537 \pm 0{,}3159$$

$$[26{,}22; \quad 26{,}85]$$

que coincide con lo hallado por el `R`: $[26{,}21989; \quad 26{,}85172]$.

En cuanto al intervalo de predicción para una nueva observación de perímetro cefálico a realizarse en un bebé de 29 semanas de gestación, el intervalo de predicción de nivel $1 - \alpha = 0{,}95$ resulta ser

$$\widehat{Y}_h \pm t_{n-2;\frac{\alpha}{2}} \cdot \widehat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_h - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

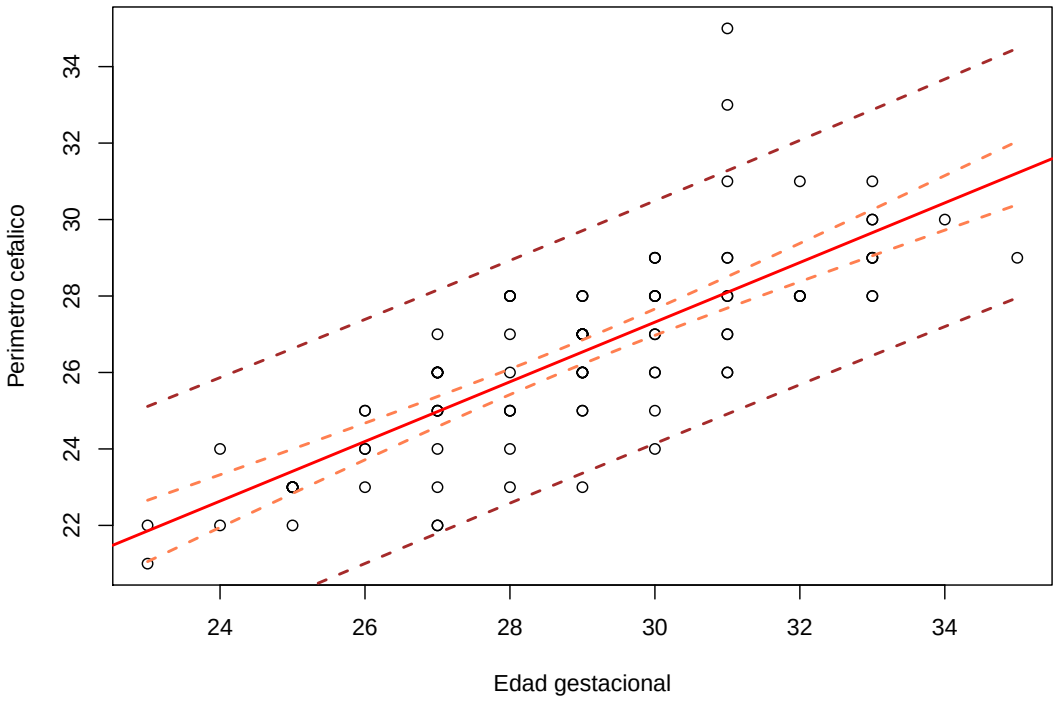
$$26{,}537 \pm 1{,}984467 \cdot 1{,}5903 \cdot \sqrt{1 + \frac{1}{100} + \frac{(29 - 28{,}89)^2}{635{,}69}}$$

$$26{,}537 \pm 3{,}1717$$

$$[23{,}365; \quad 29{,}709]$$

que coincide con lo hallado por el R: [23,36391; 29,70770] . Veámoslo gráficamente. Si construimos un IC y un IP para cada x_h tenemos el gráfico de la Figura 20.

Figura 20: Recta ajustada e intervalos de confianza y de predicción para el ejemplo de los 100 bebés.



Observemos que el IC para $E(Y_h \mid X = \overline{x})$ es el más corto. Y que los IP son mucho más anchos que los IC. De hecho, si aumentáramos el tamaño de muestra muchísimo (lo que matemáticamente se dice "hiciéramos tender n a infinito") y eligiéramos los X_i de manera tal que $\frac{(x_h-\overline{X})^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$ tendiera a cero, entonces la longitud de los IC tendería a cero, pero la longitud de los IP no. Una observación sobre el gráfico anterior es que las conclusiones tienen nivel de confianza $1-\alpha$ **para cada valor** (o nivel de predicción para cada IP) calculado, pero no hay nivel de confianza simultáneo. (O sea, la probabilidad de que un IC contenga al verdadero parámetro es $1-\alpha$, sin embargo la probabilidad de que simultáneamente el IC

calculado para $x_h = 29$ y el IC calculado para $x_{h+1} = 30$ ambos contengan a los dos verdaderos parámetros, no puede asegurarse que sea $1 - \alpha$).

2.13. Banda de confianza para la recta estimada

A veces uno quiere obtener una banda de confianza para **toda** la recta de regresión, $E(Y \mid X = x) = \beta_0 + \beta_1 x$. Es decir, una región que, con una confianza prefijada, que denominamos $1 - \alpha$, contenga a la recta completa. A su vez, esta región de confianza también tendrá nivel al menos $1 - \alpha$ para cada valor de x en particular.

La banda de confianza de Working–Hotelling de nivel $1 - \alpha$ para el modelo de regresión lineal tiene los siguientes dos límites, para cada valor x_h que pueda tomar la covariable X (x_h puede ser un valor de X observado en la muestra o no observado, mientras esté entre el mínimo y el máximo valor observado)

$$\widehat{Y}_h \pm W \cdot \widehat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{\left(x_h - \overline{X}\right)^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}}. \tag{25}$$

donde

$$W = \sqrt{2F_{1-\alpha,2,n-2}},$$

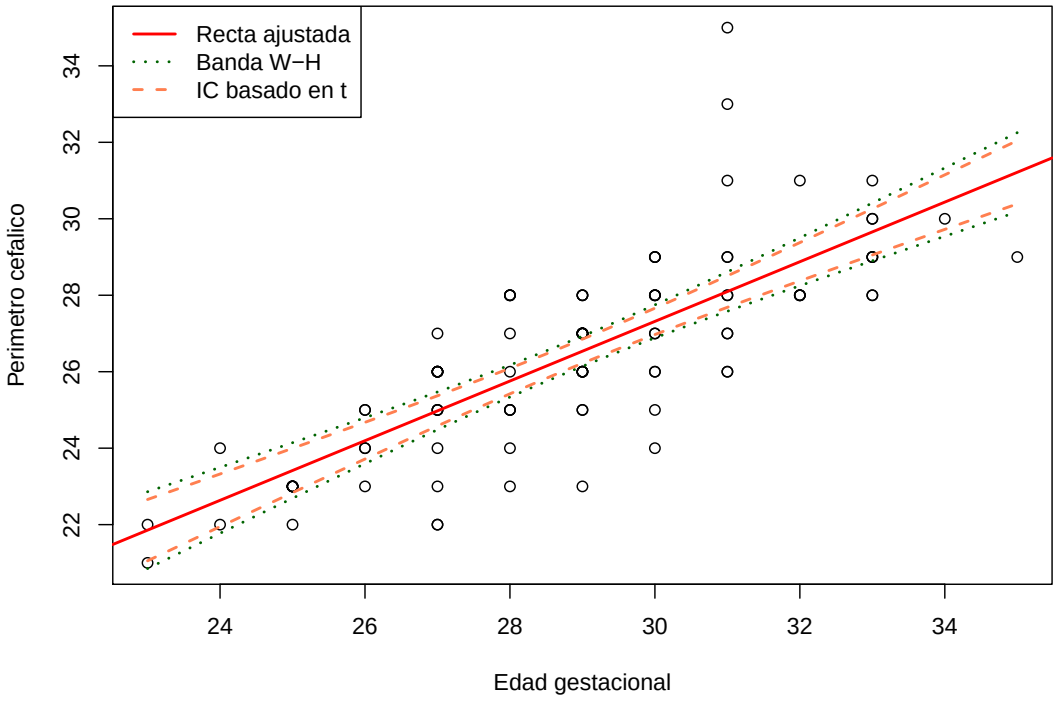
donde $F_{1-\alpha,2,n-2}$ es el cuantil $1-\alpha$ de una distribución F de Fisher con 2 grados de libertad en el numerador y $(n-2)$ en el denominador, que en `R` se calcula usando el comando `qf(`$1-\alpha$`,df1=` 2`,df2=` $n-2$). Observemos que la fórmula (25) tiene la misma forma que (24) para el intervalo de confianza para la esperanza condicional de Y cuando $X = x_h$, excepto que el cuantil t se modifica por el de la distribución W, que es más grande y cubre el nivel simultáneo. Si los comparamos para el ejemplo de 100 niños de bajo peso ($n = 100$), tomando nivel $1 - \alpha = 0{,}95$ tenemos

$$\begin{aligned} t_{n-2;\frac{\alpha}{2}} &= t_{98,0,975} = 1{,}984 \\ W &= \sqrt{2F_{1-\alpha,2,n-2}} = \sqrt{2F_{0,9,2,98}} = 2{,}1714. \end{aligned}$$

Para comparar ambas regiones, las presentamos en la Figura 21 para el ejemplo de los bebés de bajo peso.

Observación 2.2 *Los límites de la banda de confianza para la recta de regresión representan una curva que, matemáticamente, se denomina hipérbola. La región obtenida es más angosta para los valores de X cercanos a $\overline{X}$ y más ancha cuando nos alejamos de él.*

Figura 21: Banda de confianza de Working-Hotelling de nivel 0.95 para la recta esperada, comparada con los intervalos de confianza basados en la distribución t presentados en (24), para el ejemplo de bebés de bajo peso.



Observación 2.3 *La banda de confianza es válida para todos los valores de X para los cuáles vale el modelo lineal (entre el mínimo y máximo observados en la muestra), simultáneamente. Volveremos sobre los niveles de significatividad conjunta en la Sección 4.9.3. Es decir, si queremos hallar intervalos de confianza de nivel conjunto* 0,95 *para los niveles de edad gestacional 29, 31 y 33 semanas, debemos usar la banda de confianza de Working–Hotelling, para obtenerlos.*

Observación 2.4 *Para calcularlos en* `R`*, se puede hacer la cuenta "a mano", es decir, usando los percentiles de la F y los desvíos estándares que calcula el* `lm`*, o bien de manera automática con el paquete* `investr`*, y el comando* `predFit()`*, como vemos a continuación:*

```
> library(investr)
```

```
> ajuste <- lm(headcirc ~ gestage)
> equis<- c(29,31,33)
> predFit(ajuste, newdata = data.frame(gestage = equis),
+         interval = 'confidence', adjust = 'Scheffe')
       fit      lwr      upr
1 26.53581 26.14011 26.93150
2 28.09591 27.58044 28.61138
3 29.65602 28.90005 30.41199
```

2.14. Descomposición de la suma de cuadrados (ANOVA para regresión)

El análisis de la varianza provee un método apropiado para comparar el ajuste que dos o más modelos proporcionan a los mismos datos. La metodología presentada aquí será muy útil en regresión múltiple, y con modificaciones no demasiado importantes, en la mayoría de los problemas de regresión más generales. Queremos comparar el ajuste proporcionado por el modelo de regresión con el modelo más simple disponible.

¿Cuál es el modelo más simple a nuestra disposición? Es el modelo en el que no contamos con la variable explicativa X y sólo tenemos las observaciones $Y_1, \ldots, Y_n$. A falta de algo mejor proponemos el modelo

$$\begin{aligned} &\text{Modelo A:} \quad E(Y \mid X) = \mu, \text{ o escrito de otro modo} \\ &\text{Modelo A:} \quad Y_i = \mu + u_i \qquad \text{con } u_i \sim N\left(0, \sigma_Y^2\right), \quad 1 \leq i \leq n, \\ &\qquad\qquad \text{independientes entre sí.} \end{aligned}$$

Es lo que se conoce como el modelo de posición para las $Y's$. Un estimador puntual de μ es $\overline{Y}$ y un estimador de la varianza o variabilidad de las Y's bajo el *modelo A* es la varianza muestral

$$\frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2.$$

En este contexto, la varianza muestral es una medida de la variabilidad de Y que no queda explicada por el Modelo A. A la cantidad $\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$ se la denomina *suma de los cuadrados total* (SSTo). Estos sumandos tienen $n-1$ grados de libertad ya que si uno conoce los valores de $\left(Y_1 - \overline{Y}\right), \ldots, \left(Y_{n-1} - \overline{Y}\right)$ puede deducir el valor de $\left(Y_n - \overline{Y}\right)$ pues todos ellos suman 0.

Si ahora usamos los pares $(X_1, Y_1), \ldots, (X_n, Y_n)$ para estimar la recta de re-

gresión tenemos el modelo

$$\begin{array}{ll} \text{Modelo B:} & E(Y \mid X) = \beta_0 + \beta_1 X, \text{ o escrito de otro modo} \\ \text{Modelo B:} & Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad \text{con } \varepsilon_i \sim N\left(0, \sigma^2\right), \quad 1 \leq i \leq n, \\ & \text{independientes entre sí.} \end{array} \tag{26}$$

Ahora la variabilidad de las Y_i que no queda explicada por el modelo de regresión (*modelo B*) puede estimarse por

$$\frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 = \frac{1}{n-2}\text{SSRes}$$

es decir, la variación de las observaciones alrededor de la recta ajustada. Como ya comentamos, a la cantidad $\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$ se la denomina *suma de los cuadrados de los residuos* (SSRes). Estos sumandos (los residuos) tienen $n-2$ grados de libertad pues si uno conoce los valores de $\left(Y_1 - \widehat{Y}_1\right), \ldots, \left(Y_{n-2} - \widehat{Y}_{n-2}\right)$ puede deducir el valor de $e_{n-1} = \left(Y_{n-1} - \widehat{Y}_{n-1}\right)$ y $e_n = \left(Y_n - \widehat{Y}_n\right)$ ya que los residuos satisfacen las dos ecuaciones normales (suman 0 y su correlación muestral con las $X's$ es cero, las ecuaciones (15) y (16)).

Si comparamos los dos modelos disponibles para las $Y's$ vemos que el Modelo A está incluído en el Modelo B, ya que tomando $\beta_0 = \mu$ y $\beta_1 = 0$ en el Modelo B obtenemos el Modelo A como un caso particular del modelo B. Estadísticamente se dice que ambos modelos están anidados. Es decir, que ajustar bajo el Modelo A corresponde a encontrar la mejor recta *horizontal* que ajuste a los datos, mientras que ajustar bajo el Modelo B es encontrar la mejor recta (no vertical) que ajuste a los datos. La Figura 22 muestra los ajustes de ambos modelos para un mismo conjunto de datos.

Si todas las Y_i cayeran sobre la recta, SSResiduos sería igual a cero. Cuánto mayor sea la variación de las Y_i alrededor de la recta ajustada, mayor será la SSResiduos.

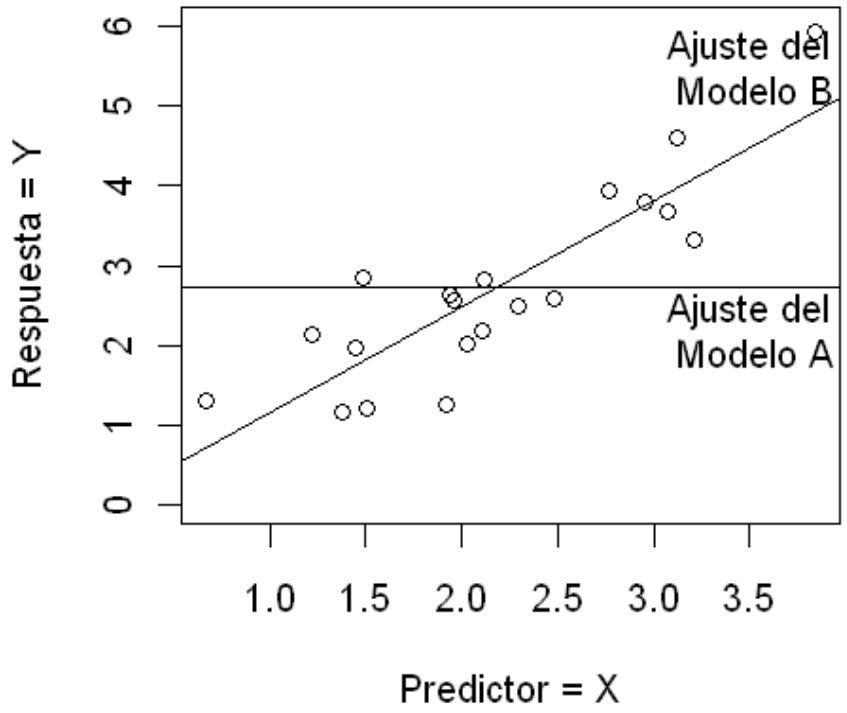
¿Cuál de las dos será mayor: SSTotal o SSRes? Vale que

$$\text{SSRes} \leq \text{SSTotal}$$

pues $\widehat{\beta}_0$ y $\widehat{\beta}_1$ son los estimadores de mínimos cuadrados, es decir, son aquellos valores de ordenada al origen a y pendiente b que minimizan la suma de los cuadrados siguiente

$$g(a,b) = \sum_{i=1}^{n}\left(Y_i - (a + bX_i)\right)^2.$$

Figura 22: Las dos esperanzas o medias condicionales ajustadas bajo ambos modelos, para un conjunto de veinte datos



Por lo tanto,

$$\begin{aligned}\text{SSRes} = g\left(\widehat{\beta}_0, \widehat{\beta}_1\right) &= \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right)\right)^2 \\ &\leq g\left(a, b\right) = \sum_{i=1}^{n}\left(Y_i - \left(a + bX_i\right)\right)^2 \qquad \text{para todo } a \text{ y } b. \end{aligned} \tag{27}$$

En particular, tomando $a = \overline{Y}$ y $b = 0$ tenemos $g\left(\overline{Y}, 0\right) = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$ y de (27) tenemos

$$\text{SSRes} \leq \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \text{SSTo}. \tag{28}$$

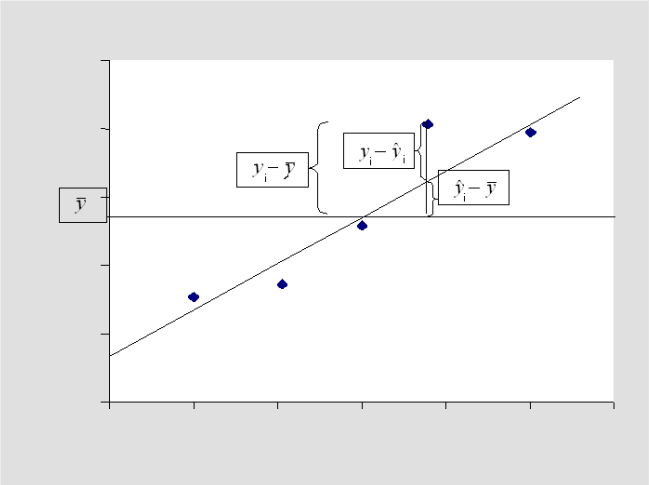
Podemos interpretar a SSTo como una medida de la variabilidad de las Y que no queda explicada por el modelo A. Es una medida del desajuste del modelo A a los datos. Lo mismo puede decirse de SSRes: es una medida de la variabilidad

de la Y que no queda explicada por el modelo de regresión lineal (modelo B). La desigualdad (28) nos dice que la mejor recta ajusta mejor a los datos que la mejor recta horizontal, como ya discutimos, y graficamos en la Figura 22. Podemos hacer la siguiente descomposición de cada uno de los sumandos de SSTo

$$\underbrace{Y_i - \overline{Y}}_{\text{desviación total}} = \underbrace{Y_i - \widehat{Y}_i}_{\substack{\text{desvío alrededor} \\ \text{de la recta de regresión} \\ \text{ajustada}}} + \underbrace{\widehat{Y}_i - \overline{Y}}_{\substack{\text{desvío de los predichos} \\ \text{respecto de la media}}} \tag{29}$$

En la Figura 23 vemos estas diferencias graficadas para una observación. La desviación total $Y_i - \overline{Y}$ mide la distancia vertical (con signo) de la observación a la recta horizontal que corta al eje vertical en $\overline{Y}$, $Y_i - \widehat{Y}_i$ mide la distancia vertical (con signo, es decir puede ser positivo o negativo, según dónde esté ubicada la observación) de la observación a la recta ajustada por mínimos cuadrados y $\widehat{Y}_i - \overline{Y}$ mide la distancia vertical (con signo) entre los puntos que están ubicados sobre ambas rectas y tienen la misma coordenada X_i. Cada una de estas cantidades puede ser positiva, negativa o nula para distintas observaciones.

Figura 23: Los tres términos que aparecen en la igualdad (29) para una observación.



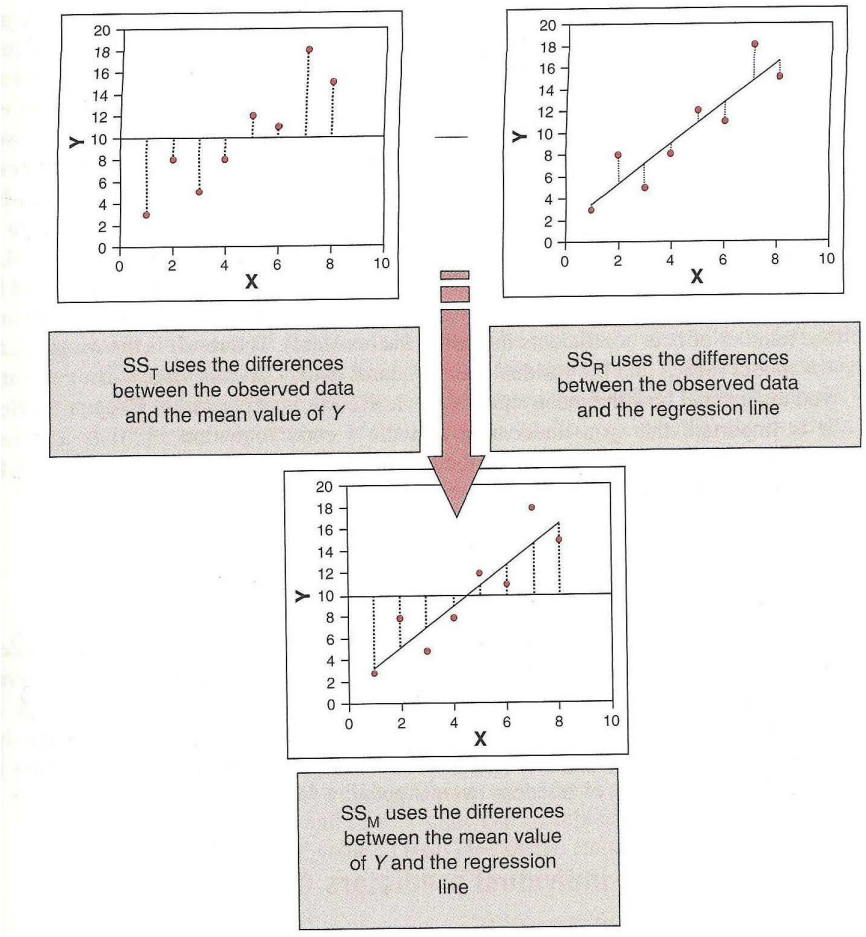
Obviamente es falso que el cuadrado del término de la izquierda en la igualdad (29) anterior sea igual a la suma de los cuadrados de los términos de la derecha es decir,

$$\left(Y_i - \overline{Y}\right)^2 \neq \left(Y_i - \widehat{Y}_i\right)^2 + \left(\widehat{Y}_i - \overline{Y}\right)^2 \qquad \text{para cada } i.$$

Sin embargo, puede probarse que vale la siguiente igualdad, cuando sumamos sobre todas las observaciones

$$\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 + \sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2. \tag{30}$$

Figura 24: El primer gráfico contiene las distancias (con signo) que intervienen en la SSTo, es decir, las diferencias entre los valores observados de Y y la media muestral $\overline{Y}$, el segundo tiene las diferencias entre las observaciones y los valores predichos por la recta ajustada, que conforman la SSRes y el tercer gráfico muestra la diferencia entre los valores predichos por el modelo lineal y el promedio $\overline{Y}$, que forman la SSReg o SSM. Fuente: Field [2005], pág. 149.



El tercer término involucrado en esta suma recibe el nombre de *suma de cuadrados de la regresión* (SSReg, algunos autores lo llaman suma de cuadrados del

modelo, SSM), y por la igualdad anterior, puede escribirse la siguiente igualdad

$$\begin{aligned}\text{SSReg} &= \sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 - \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 \\ &= \text{SSTo} - \text{SSRes}.\end{aligned}$$

En la Figura 24 pueden verse los tres sumandos de esta descomposición en forma gráfica para un conjunto de datos.

Como la SSReg queda completamente determinada al quedar determinada la inclinación de la recta (recordemos que los valores de X_i están fijos), es decir, la pendiente de la recta, decimos que la SSReg tiene un sólo grado de libertad.

Con estas cantidades se construye la tabla de análisis de la varianza que aparece en la salida de cualquier paquete estadístico en lo que se conoce como tabla de ANOVA (*Analysis of Variance table*). Resumimos esta información en la Tabla 14

Tabla 14: Tabla de ANOVA para el modelo de Regresión Lineal Simple

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	p-valor
Regresión	SSReg	1	MSReg	$\frac{\text{MSReg}}{\text{MSRes}}$	$P\left(F_{1,n-2} \geq F_{obs}\right)$
Residuos	SSRes	$n-2$	MSRes		
Total	SSTo	$n-1$			

donde

$$\begin{aligned}&\text{SSReg} = \textstyle\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2 \qquad \text{MSReg} = \frac{\text{SSReg}}{1} \\ &\text{SSRes} = \textstyle\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 \qquad \text{MSRes} = \frac{\text{SSRes}}{n-2} \\ &\text{SSTo} = \textstyle\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \qquad F = \frac{\text{MSReg}}{\text{MSRes}} = \frac{\text{SSReg}(n-2)}{\text{SSRes}}\end{aligned}$$

La primer columna tiene las sumas de cuadrados, la segunda los respectivos grados de libertad, la tercera tiene el cociente entre la primera y la segunda, es decir, esta columna tiene lo que denominamos los *cuadrados medios* (o media cuadrática, *mean square*, en inglés). Explicaremos las dos últimas columnas de la tabla de ANOVA en la Sección 2.16.

Observemos también, que la última fila de la tabla es la suma de las primeras dos, lo cual es consecuencia de la ecuación (30) es decir

$$\text{SSTo} = \text{SSRes} + \text{SSRegresión}.$$

El valor de las sumas de cuadrados depende de la escala en la que está medida la variable Y. Cambia si cambiamos las unidades de medida de las Y: por ejemplo de cm. a metros, de pesos a pesos (en miles) o de kg. a g.

Ejemplo 2.4 *En la Tabla 15 exhibimos la tabla de ANOVA que proporciona la salida del R para los datos de bebés con bajo peso. Observemos que las dos primeras columnas están intercambiadas respecto de la descripción hecha antes (primero los grados de libertad y luego las sumas de cuadrados). En la tercer columna puede verse el mean square residual, que en este caso vale 2.53. Este es el estimador de* σ^2 *dado por el modelo, es decir,* $MSRes = SSRes/(n-2)$. *En la salida del modelo lineal en la Tabla 12, veíamos la raíz cuadrada del mismo. Por otro lado, vemos que los valores numéricos exhibidos en la tabla no nos dan información que nos permita evaluar la bondad de la regresión.*

Tabla 15: Tabla de ANOVA, salida de R con el comando `anova`, para los 100 bebés con bajo peso.

```
> ajuste<-lm(headcirc~gestage)
> anova(ajuste)
Analysis of Variance Table

Response: headcirc
          Df Sum Sq Mean Sq F value    Pr(>F)
gestage    1 386.87  386.87  152.95 < 2.2e-16 ***
Residuals 98 247.88    2.53
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

En la siguiente sección nos ocuparemos de construir una medida para evaluar la bondad del modelo de regresión, en cuanto al ajuste a nuestros datos, que no sea dependiente de la escala en la que esté medida la variable Y, a partir de la tabla de ANOVA.

2.15. El coeficiente de determinación R^2

Trataremos de construir una medida de la fuerza de la relación entre la variable dependiente e independiente, que nos indique cuán buen predictor de Y es X. Se trata de decidir si el hecho de conocer el valor de X mejora nuestro conocimiento de Y. O sea, si uno puede predecir Y mucho mejor usando la recta de regresión

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

que sin conocer el valor de X, entonces las variables están asociadas. Para ello usaremos la descomposición de la suma de cuadrados vista en la sección anterior.

Por lo descripto allí, la mejora en el ajuste a los datos conseguida por la inclusión del modelo B resulta ser SSTo − SSRes. ¿Cuánto de la variabilidad total de las Y queda explicada por la regresión? Podemos plantear la siguiente regla de tres simple:

$$100\,\%\text{ de variabilidad} \quad \text{——} \quad \text{SSTo}$$

$$\%\text{ de variabilidad explicada} \quad \text{——} \quad \text{SSTo} - \text{SSRes}$$

Luego el porcentaje de variabilidad explicada es

$$\frac{\text{SSTo} - \text{SSRes}}{\text{SSTo}} \times 100\,\%.$$

A la cantidad

$$\frac{\text{SSTo} - \text{SSRes}}{\text{SSTo}} = \frac{\text{SSReg}}{\text{SSTo}}$$

se la denomina R^2, o **coeficiente de determinación**.

R^2 nos dice qué proporción de la variabilidad total en la variable Y puede ser explicada por la variable regresora, en consecuencia es una medida de la capacidad de *predicción* del modelo.

R^2 también puede verse como una medida de la fuerza de la *asociación lineal* entre X e Y. (Hacemos énfasis en la palabra lineal porque fue obtenido bajo un modelo lineal).

2.15.1. Propiedades de R^2

- $0 \leq R^2 \leq 1$
- No depende de las unidades de medición.
- Es el cuadrado del coeficiente de correlación de Pearson para la muestra $\{(X_i, Y_i)\}_{1\leq i\leq n}$. También es el cuadrado del coeficiente de correlación de Pearson para los pares $\left\{\left(\widehat{Y}_i, Y_i\right)\right\}_{1\leq i\leq n}$, es decir, entre los valores de la covariable observados y los predichos por el modelo lineal.
- Mientras mayor es R^2 mayor es la fuerza de la variable regresora (X) para predecir a la variable respuesta (Y).
- Mientras mayor sea R^2 menor es la SSRes y por lo tanto, más cercanos están los puntos a la recta.
- Toma el mismo valor cuando usamos a X para predecir a Y o cuando usamos a Y para predecir a X.

Ejemplo 2.5 *Para los datos de la regresión de perímetro cefálico versus edad gestacional, en la salida del modelo lineal en la Tabla 12, vemos que*

$$R^2 = 0{,}6095$$

Este valor implica una relación lineal moderadamente fuerte entre la edad gestacional y el perímetro cefálico. En particular, el $60{,}95\,\%$ *de la variabilidad observada en los valores de perímetro cefálico queda explicada por la relación lineal entre el perímetro cefálico y la edad gestacional. El restante*

$$100\,\% - 60{,}95\,\% = 39{,}05\,\%$$

de la variabilidad no queda explicada por esta relación.

El R^2 no se usa para testear hipótesis del modelo sino como una medida de la capacidad predictiva de la relación lineal ajustada.

2.16. Test F (otro test para $H_0 : \beta_1 = 0$)

A partir de la Tabla de ANOVA es posible derivar un test para $H_0 : \beta_1 = 0$.

En el contexto de regresión lineal simple ya hemos obtenido el test t que resuelve este punto. El test F será más importante en Regresión Múltiple.

El razonamiento es el siguiente. Bajo los supuestos del modelo de regresión, puede probarse que

1. La distribución de muestreo de MSRes = SSRes$/(n-2)$ tiene esperanza σ^2.
2. La distribución de muestreo de MSReg = SSReg$/1$ tiene esperanza $\sigma^2 + \beta_1^2 \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$.

Entonces, cuando H_0 es verdadera ambos cuadrados medios (el residual y el de regresión) deberían parecerse mucho, o su cociente debería parecerse a uno, y cuando H_0 no es cierta, el numerador tenderá a tomar valores mucho más grandes que el denominador. Por lo tanto, es razonable considerar el estadístico

$$F = \frac{\text{MSReg}}{\text{MSRes}} = \frac{\frac{\text{SSReg}}{1}}{\frac{\text{SSRes}}{n-2}} = \frac{\text{SSReg}}{\text{SSRes}/\left(n-2\right)}$$

como candidato para testear la hipótesis $H_0 : \beta_1 = 0$. Esperamos que F esté cerca de 1 (o sea menor a 1) si H_0 es verdadera y que F sea mucho más grande cuando H_0 es falsa.

Puede probarse que, bajo los supuestos del modelo lineal y cuando H_0 es verdadera, F tiene distribución de Fisher con 1 grado de libertad en el numerador y $n-2$ grados de libertad en el denominador. Por lo tanto, un test de nivel α para

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

rechazará la hipótesis nula si el valor del estadístico observado F_{obs} cumple que $F_{obs} > F_{1,n-2,1-\alpha}$. O cuando su p-valor es menor a α, siendo

$$\text{p-valor} = P\left(F\left(1, n-2\right) > F_{obs}\right).$$

Las dos últimas columnas de la tabla de ANOVA descripta en la Tabla 14 presentan estos valores.

Observación 2.5 *El test F que obtendremos aquí es el mismo que el test t presentado en la Sección 2.9 para testear la hipótesis $H_0 : \beta_1 = 0$, ya que F se define como el cuadrado del estadístico empleado en el test t. Para comprobarlo, observemos que a partir de la ecuación que define a $\widehat{\beta}_0$ (12) tenemos*

$$\begin{aligned}
\widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{X}. \\
MSReg &= SSReg = \sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} \left(\left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right) - \overline{Y}\right)^2 \\
&= \sum_{i=1}^{n} \left(\left(\overline{Y} - \widehat{\beta}_1 \overline{X} + \widehat{\beta}_1 X_i\right) - \overline{Y}\right)^2 = \sum_{i=1}^{n} \left(-\widehat{\beta}_1 \overline{X} + \widehat{\beta}_1 X_i\right)^2 \\
&= \widehat{\beta}_1^2 \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \\
MSRes &= \frac{SSRes}{n-2}
\end{aligned}$$

Luego, si recordamos el estadístico T definido en las ecuaciones (22) para testear la hipótesis de pendiente igual a cero, tenemos

$$T = \frac{\widehat{\beta}_1 - 0}{\sqrt{\widehat{Var}\left(\widehat{\beta}_1\right)}} = \frac{\widehat{\beta}_1}{se_{\widehat{\beta}_1}} = \frac{\widehat{\beta}_1}{\sqrt{\frac{\text{SS}Res/(n-2)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}}$$

y el estadístico F que resulta ser

$$F = \frac{MSReg}{MSRes} = \frac{\widehat{\beta}_1^2 \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{\frac{SSRes}{n-2}} = \frac{\widehat{\beta}_1^2}{\frac{\text{SS}Res/(n-2)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}},$$

vemos que

$$F = T^2$$

y el p-valor del test t se calculaba

$$\begin{aligned} p-valor &= 2P\left(T \geq |T_{obs}|\right) = P\left(T \geq |T_{obs}| \text{ ó } T \leq |T_{obs}|\right) \\ &= P\left(|T| \geq |T_{obs}|\right) = P\left(|T|^2 \geq |T_{obs}|^2\right) = P\left(T^2 \geq T_{obs}^2\right) \\ &= P\left(F \geq F_{obs}\right) \end{aligned}$$

dando el mismo p-valor que el test de Fisher.

Ejemplo 2.6 *Si miramos la tabla de ANOVA para el ejemplo de los 100 bebés (Figura 15), vemos que el estadístico del test F toma el valor*

$$F = 152{,}947.$$

Su raíz cuadrada es $\sqrt{152{,}947} = 12{,}367$, *que es el valor del estadístico T para testear si la pendiente es o no nula, como puede verse en la Tabla 12.*

2.17. Ejercicios (segunda parte)

Estos ejercicios se resuelven con el `script_reglinealsimple2.R`

Ejercicio 2.5 *Medidas del cuerpo, Parte IV. Base de datos* **`bdims`** *del paquete* ***`openintro`***.

*(a) Compare los ajustes realizados en los ejercicios 2.1 y 2.2. En ambos se ajusta un modelo lineal para explicar el peso medido en kilogramos (***`wgt`***): en el ejercicio 2.1 por la circunferencia de la cadera medida en centímetros (***`hip.gi`***), en el ejercicio 2.2 por la altura media en centímetros (***`hgt`***). ¿Cuál de los dos covariables explica mejor al peso? ¿Qué herramienta utiliza para compararlos?*

(b) Para el ajuste del peso usando la circunferencia de cadera como única covariable, halle un intervalo de confianza de nivel 0.95 cuando el contorno de cadera mide 100 cm. Compárelo con el intervalo de predicción para ese mismo contorno de cadera.

(c) Para el ajuste del peso usando la altura como única covariable, halle un intervalo de confianza de nivel 0.95 cuando la altura es de 176 cm. Compárelo con el intervalo de predicción para esa misma altura. ¿Cuál de los dos modelos da un intervalo de predicción más útil?

(d) *Construya un intervalo de confianza para el peso esperado cuando el contorno de cintura es de 80cm., 95cm., 125cm. de nivel 0.95. Estos tres intervalos, ¿tienen nivel simultáneo 0.95? Es decir, la siguiente afirmación ¿es verdadera o falsa? Justifique. En aproximadamente 95 de cada 100 veces que yo construya los IC basados en una (misma) muestra, cada uno de los 3 IC contendrán al verdadero valor esperado del peso.*

(e) *Construya los intervalos de predicción para el peso esperado cuando de nivel (individual) 0.95 cuando el contorno de cintura es de 80cm., 95cm. y 125cm. Compare las longitudes de estos tres intevalos entre sí. Compárelos con los IC de nivel individual.*

(f) *Construya los intervalos de confianza para el peso esperado cuando de nivel simultáneo 0.95 cuando el contorno de cintura es de 80cm., 95cm. y 125cm.*

(g) *Estime la varianza del error (σ^2) en ambos modelos.*

(h) *Realice un scatterplot del peso en función del contorno de cintura. Superponga los IC y los IP al gráfico, de nivel 0.95 (no simultáneo).*

Ejercicio 2.6 *(Del Libro de Weisberg [2005]) Uno de los primeros usos de la regresión fue estudiar el traspaso de ciertos rasgos de generación en generación. Durante el período 1893–1898, E. S. Pearson organizó la recolección de las alturas de $n = 1375$ madres en el Reino Unido menores de 65 años y una de sus hijas adultas mayores de 18 años. Pearson y Lee (1903) publicaron los datos, y usaremos estos datos para examinar la herencia. Los datos (medidos en pulgadas) pueden verse en el archivo de datos* `heights.txt` *del paquete* `alr3` *de* `R`*. Nos interesa estudiar el traspaso de madre a hija, así que miramos la altura de la madre, llamada* `Mheight`*, como la variable predictora y la altura de la hija,* `Dheight`*, como variable de respuesta. ¿Será que las madres más altas tienden a tener hijas más altas? ¿Las madres más bajas tienden a tener hijas más bajas?*

(a) *Realice un scatterplot de los datos, con la altura de las madres en el eje horizontal.*

   i. *Como lo que queremos es comparar las alturas de las madres con la de las hijas, necesitamos que en el scatterplot las escalas de ambos ejes sean las mismas (y que por lo tanto el gráfico sea cuadrado).*

   ii. *Si cada madre e hija tuvieran exactamente la misma altura que su hija, ¿cómo luciría este scatterplot? Resuma lo que observa en este gráfico. Superpóngale la figura que describió como respuesta a la pregunta anterior. ¿Describe esta figura un buen resumen de la relación entre ambas variables?*

iii. Los datos originales fueron redondeados a la pulgada más cercana. Si trabajamos directamente con ellos, veremos menos puntos en el scatterplot, ya que varios quedarán superpuestos. Una forma de lidiar con este problema es usar el jittering, es decir, sumar un pequeño número uniforme aleatorio se a cada valor. Los datos de la librería alr3 tienen un número aleatorio uniforme en el rango de -0.5 a +0.5 añadidos. Observemos que si se redondearan los valores del archivo `heights` *se recuperarían los datos originalmente publicados. En base al scatterplot, ¿parecería ser cierto que las madres más altas suelen tener hijas más altas y viceversa con las más bajas?*

(b) Ajuste el modelo lineal a los datos. Indique el valor de la recta ajustada. Superpóngala al scatter plot. ¿Presenta visualmente un mejor ajuste que la recta identidad postulada en el ítem anterior? Dé los estimadores de los coeficientes de la recta, sus errores estándares, el coeficiente de determinación, estime la varianza de los errores. Halle un intervalo de confianza de nivel 0.95 para la pediente. Testee la hipótesis $E(\texttt{Dheight} \mid \texttt{Mheight}) = \beta_0$ *versus la alternativa que* $E(\texttt{Dheight} \mid \texttt{Mheight}) = \beta_0 + \beta_1 \texttt{Mheight}$. *Escriba su conclusión al respecto en un par de renglones.*

(c) Prediga y obtenga un intervalo de predicción para la altura de una hija cuya madre mide 64 pulgadas. Observe que para que esta predicción sea razonable, hay que pensar que la madre vivía en Inglaterra a fines del siglo XIX.

(d) Una pulgada equivale a 2.54cm. Convierta ambas variables a centímetros (`Dheightcm` *y* `Mheightcm`*) y ajuste un modelo lineal a estas nuevas variables. ¿Deberían cambiar los estimadores de* β_0 *y* β_1*? ¿De qué manera? ¿Y los errores estándares? ¿Y los p-valores? ¿Y el coeficiente de determinación? ¿Y la estimación del desvío estándar de los errores? Compare ambos resultados, y verifique si sus conjeturas resultaron ciertas. En estadística, que un estimador se adapte al cambio de escala en las variables (covariable y respuesta) se dice: "el estimador es equivariante (afín y por escala)".*

Ejercicio 2.7 *Simulación 1. El objetivo de este ejercicio es generar datos para los cuales conocemos (y controlamos) el modelo del que provienen y la distribución que siguen.*

(a) Generar $n = 22$ *datos que sigan el modelo lineal simple*

$$Y = 10 + 5X + \varepsilon, \tag{31}$$

donde $\varepsilon \sim N(0, \sigma^2)$, *con* $\sigma^2 = 49$. *Las* n *observaciones las generamos independientes entre sí.*

i. *Para hacer esto en R, conviene primero definir un vector de longitud 22 de errores, que tenga distribución normal. La instrucción que lo hace es* `rnorm`*. Visualice los errores con un histograma de los mismos.*

ii. *Inventamos los valores de X. Para eso, generamos 22 valores con distribución uniforme entre 0 y 10, con la instrucción* `runif`*. Para no trabajar con tantos decimales, redondeamos estos valores a dos decimales, con la instrucción* `round()`*.*

iii. *Ahora sí, definimos las Y usando todo lo anterior:*

$$Y_i = 10 + 5X_i + \varepsilon_i,$$

para cada $1 \leq i \leq n = 22$. Observar que nos hemos conseguido observaciones $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ independientes que siguen el modelo

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

¿Cuánto valen los verdaderos β_0 y β_1?

(b) *Haga un scatterplot de los datos generados.*

(c) *Ajuste el modelo lineal, guarde el resultado obtenido en el objeto* `ajuste`*. Observe si los parámetros estimados son significativos. Calcule intervalos de confianza para la ordenada al origen y la pendiente, de nivel 0.95. Para esto recuerde los comandos:* `lm` *y* `confint`*. ¿Los verdaderos β_0 y β_1 pertenecen a dichos intervalos? ¿Cuánto dio la pendiente estimada, $\widehat{\beta}_1$? ¿En qué parte de la salida del ajuste lineal podemos encontrar el estimador de σ? ¿Cuánto debería valer?*

(d) *Pídamosle al R que chequee si el 5 pertenece al IC de nivel 0.95 calculado en base a la muestra. El* `R` *nos devolverá "*`TRUE`*" o "*`FALSE`*" como respuesta a esta pregunta. La computadora codifica los "*`TRUE`*" como 1 y los "*`FALSE`*" como 0 para poder operar numéricamente con respuestas de este tipo. También guardemos la pendiente estimada en un objeto que se llame* `beta1est`*.*

(e) *Superpóngale al scatterplot de los datos la recta verdadera (en azul) y la estimada en base a ellos (en rojo).*

Ejercicio 2.8 *Simulación 2. Ahora hacemos un upgrade del desafío. Vamos a repetir lo hecho en el ejercicio 2.7 muchas veces, digamos lo replicaremos $B = 1000$ veces. Llamaremos replicación a cada repetición del ejercicio anterior. ¿Qué replicamos? Repetimos generar $n = 22$ observaciones del modelo (31) con errores normales (lo que llamamos elegir una muestra), ajustamos el modelo lineal, guardamos la pendiente estimada y nos fijamos si el 5 pertenece al intervalo de confianza para la pendiente.*

(a) *¿Puede usted anticipar, desde la teoría las respuestas de las preguntas que siguen?*

   i. *Las pendientes estimadas en las* $B = 1000$ *replicaciones, ¿serán siempre iguales o cambiarán de replicación en replicación?*

   ii. *¿Alrededor de qué número variarán las pendientes estimadas en las 1000 replicaciones?*

   iii. *Si hacemos un histograma de estas* $B = 1000$ *replicaciones, ¿a qué distribución debería parecerse?*

   iv. *Aproximadamente, ¿qué porcentaje de los 1000 intervalos de confianza para la pendiente estimados a partir de las 1000 muestras cubrirá al verdadero valor de la pendiente?*

   v. *Observe que si usted tuviera 22.000 observaciones de un modelo, nunca las dividiría en 1000 tandas de 22 observaciones para analizarlas: las consideraría todas juntas. Es por eso que este ejercicio es irreal, es simplemente una herramienta de aprendizaje.*

(b) *Antes de empezar, definamos vectores donde guardaremos la información. De longitud* $B = 1000$ *cada uno, necesitamos un vector para los* $\widehat{\beta}_1$ *y otro para guardar las respuestas respecto de si el 5 pertenece o no al intervalo de confianza. Llamémoslos:* `beta1est` *e* `icbeta`*. Inicialmente ponemos un* `NA` *en cada coordenada de estos vectores (*`NA` *es, usualmente, la notación reservada para una observación faltante, son las siglas de not available). La instrucción* `rep` *del R (que repite un número o una acción un número fijo de veces resultará muy útil).*

(c) *Los valores de* $X_1, \ldots, X_{22}$ *los dejaremos siempre fijos, en los valores que tomamos en el ejercicio 2.7. En cada replicación elegimos nuevos valores para los errores, y consecuentemente, nuevos valores para la variable respuesta* $Y_1, \ldots, Y_{22}$*. No nos interesará guardar ni a los errores ni a las* Y*. Para cada muestra, corra el ajuste lineal y guarde la pendiente estimada y la respuesta en forma de true o false respecto de si el intervalo de confianza para la pendiente contiene al verdadero valor de la pendiente. Todo esto puede realizarse con la instrucción* `for` *del* `R`*, que no es la manera óptima de programar, pero sí es la más comprensible.*

(d) *Haga un histograma de las pendientes estimadas. ¿Qué distribución parecen tener los datos?*

(e) *¿Qué proporción de los intervalos de confianza construidos contiene al verdadero valor de la pendiente?*

Ejercicio 2.9 *Mamíferos, Parte IV. conjunto de datos* `mammals` *del paquete* `openintro`. *Vimos, en los ejercicios 1.7 y 2.3, que el scatter plot de los datos originales no tiene la forma elipsoidal (o de pelota de rugby, más o menos achatada) que podemos describir con un modelo de regresión lineal. Por ello, ajustamos un modelo lineal para explicar a* $\log_{10}(\texttt{BrainWt})$ *en función del* $\log_{10}(\texttt{BodyWt})$,

$$\log_{10}(\texttt{BrainWt}) = \beta_0 + \beta_1 \log_{10}(\texttt{BodyWt}) + \varepsilon. \tag{32}$$

Una observación: en el help del `openitro` *se indica que la variable* `BrainWt` *está medida en kg., sin embargo, esta variable está medida en gramos.*

(a) A partir de $\log_{10}(10) = 1$ *y de recordar que*

$$\log_{10}(ab) = \log_{10}(a) + \log_{10}(b),$$

podemos observar que en el modelo lineal (32) aumentar una unidad de $log_{10}(\texttt{BodyWt})$ *es lo mismo que multiplicar a* `BodyWt` *por 10. Si dos animales difieren en el* `BodyWt` *por un factor de diez, dé un intervalo del 95 % de confianza para la diferencia en el* $\log_{10}(\texttt{BrainWt})$ *para estos dos animales.*

(b) Para un mamífero que no está en la base de datos, cuyo peso corporal es de 100 kg., obtenga la predicción y un intervalo de nivel 95 % de predicción del $\log_{10}(\texttt{BrainWt})$. *Prediga el peso del cerebro de dicho animal. Ahora queremos convertir el intervalo de predicción del* $\log_{10}(\texttt{BrainWt})$ *en un intervalo de predicción para el* `BrainWt`. *Para eso, observemos que si el intervalo* (a, b) *es un intevalo de predicción de nivel 95 % para* $\log_{10}(\texttt{BrainWt})$, *entonces, un intervalo para el* `BrainWt` *está dado por* $\left(10^a, 10^b\right)$. *¿Por qué? Use este resultado para obtener un intervalo de predicción del peso del cerebro del mamífero cuyo peso corporal es 100kg. Mirando los valores numéricos obtenidos, ¿parece muy útil el resultado obtenido?*

(c) Observe que si quisiéramos construir el intervalo de confianza de nivel 95 % para el peso del cerebro esperado de un mamífero cuyo peso corporal es es 100kg, no es posible hacer la conversión del ítem anterior de manera automática, ya que para cualquier funcion g en general

$$E[g(Y)] \neq g(E[Y]).$$

Si se quiere construir dicho intervalo, habrá que apelar a otras herramientas, por ejemplo el desarrollo de Taylor de la función g.

Ejercicio 2.10 *(Del Libro de Weisberg [2005]) La perca americana o lubina (smallmouth bass) es un pez que vive en lagos y cuya pesca constituye una actividad bastante difundida. En Estados Unidos, para garantizar un equilibrio saludable entre*

la conservación del medio ambiente y la explotación humana se implementan distintas políticas de regulación de su pesca. Entender los patrones de crecimiento de los peces es de gran ayuda para decidir políticas de conservación de stock de peces y de permisos de pesca. Para ello, la base de datos `wblake` *del paquete* `alr3` *registra la longitud en milímetros al momento de la captura (*`Length`*) y la edad (*`Age`*) para* $n = 439$ *percas medidas en el Lago West Bearskin en Minnesota, EEUU, en 1991. Ver help(*`wblake`*) para más información de los datos. Las escamas de los peces tienen anillos circulares como los árboles, y contándolos se puede determinar la edad (en años) de un pez. La base de datos también tiene la variable* `Scale` *que mide el radio de las escamas en mm., que no utilizaremos por ahora.*

(a) *Hacer un scatter plot de la longitud (*`Length`*) en función de la edad (*`Age`*). ¿Qué observa? La apariencia de este gráfico es diferente de los demás gráficos de dispersión que hemos hecho hasta ahora. La variable predictora* `Age` *sólo puede tomar valores enteros, ya que se calculan contando los anillos de las esacamas, de modo que realmente estamos graficando ocho poblaciones distintas de peces. Como es esperable, la longitud crece en general con la edad, pero la longitud del pez más largo de un año de edad excede la longitud del pez más corto de cuatro años de edad, por lo que conocer la edad de un pez no nos permitirá predecir su longitud de forma exacta.*

(b) *Calcule las medias y los desvíos estándares muestrales para cada uno de las ocho subpoblaciones de los datos de las percas. Dibuje un boxplot de la longitud para cada edad de las percas, todos en la misma escala. Describa lo que ve. La longitud, ¿parece aumentar con la edad? La dispersión de la longitud, ¿parece mantenerse más o menos constante con la edad? ?¿O crece? ¿O decrece?*

(c) *Ajuste un modelo lineal para explicar la longitud (*`Length`*) en función de la edad (*`Age`*). ¿Resulta significativa la pendiente? Resuma la bondad del ajuste con el* R^2*. Superponga la recta estimada al gráfico de dispersión, y también las medias muestrales por grupos. Halle el estimador de* σ *que proporciona el modelo lineal. ¿A qué valor debiera parecerse? ¿Se parece? Observar que no debiera parecerse a* `sd(Length)`*. ¿Le parece que el ajuste obtenido por el modelo lineal es satisfactorio?*

(d) *Obtenga intervalos de confianza de nivel 95 % para la longitud media a edades 2, 4 y 6 años (no simultáneos). ¿Sería correcto obtener IC para la longitud media a los 9 años con este conjunto de datos?*

3. Diagnóstico en Regresión

Las técnicas del diagnóstico en regresión se abocan a validar que los supuestos realizados por el modelo sean apropiados para los datos con los que se cuenta. Son realizadas a posteriori del ajuste (aunque filosóficamente se deberían realizar antes) y están basadas en general en los residuos (o versiones apropiadamente escaladas) de ellos. Constan principalmente de técnicas gráficas, aunque también en la exhibición de algunas medidas de bondad de ajuste. Si el modelo propuesto, una vez ajustado a los datos, no proporciona residuos que parezcan razonables, entonces comenzamos a dudar de que algun aspecto del modelo (o todos) sea apropiado para nuestros datos. Un tema relacionado es asegurarse que la estimación realizada no sea tremendamente dependiente de un sólo dato (o un pequeño subconjunto de datos) en el sentido en que si no se contara con dicho dato las conclusiones del estudio serían completamente diferentes. La identificación de estos puntos influyentes forma parte relevante del diagnóstico (y de esta sección).

3.1. Medidas de diagnóstico

3.1.1. Leverage de una observación

El valor predicho de un dato puede escribirse como combinación lineal de las observaciones

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i = \sum_{k=1}^{n} h_{ik} Y_k \tag{33}$$

donde

$$h_{ik} = \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)\left(X_k - \overline{X}\right)}{S_{XX}}$$

y como caso particular tenemos que

$$h_{ii} = \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{S_{XX}}. \tag{34}$$

Recordemos que hemos llamado S_{XX} a la cantidad

$$S_{XX} = \sum_{k=1}^{n} \left(X_k - \overline{X}\right)^2.$$

Vale que

$$\sum_{k=1}^{n} h_{ik} = 1, \quad \sum_{i=1}^{n} h_{ik} = 1 \tag{35}$$

$$\sum_{i=1}^{n} h_{ii} = 2$$

$$\frac{1}{n} \leq h_{ii} \leq \frac{1}{s} \leq 1. \tag{36}$$

donde s es la cantidad de observaciones con predictor igual a X_i en la muestra. La cantidad h_{ii} se denomina *leverage del dato i-ésimo*. Es una medida que resume cuán lejos cae el valor de X_i de la media muestral de las X. Mide, de alguna manera, cuánto es el aporte de la observación i-ésima a la varianza muestral de las X (que es $\frac{S_{XX}}{n-1}$). La traducción de leverage al castellano es usualmente palanca, o influencia. Observemos que es un concepto que no depende del valor Y_i observado.

3.1.2. Residuos

Dijimos en la Sección 2.8 que los residuos son cantidades observables, que representan de alguna manera el correlato empírico de los errores. Para verificar los supuestos del modelo lineal, suelen usarse métodos gráficos que involucran a los residuos. El modelo lineal

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

supone que los errores ε tienen media poblacional cero y varianza constante (que denominamos σ^2), y que son indendientes para distintas observaciones. Sin embargo, ya hemos visto que no ocurre lo mismo con los residuos. Vimos que los residuos no son independientes. Además, puede probarse que

$$\begin{aligned} E(e_i) &= 0 \\ Var(e_i) &= \sigma^2 (1 - h_{ii}) \end{aligned} \tag{37}$$

donde $h_{ii} = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{S_{XX}}$, es el leverage de la observación i-ésima. En consecuencia la varianza del residuo de un dato depende del valor de la covariable, y los residuos de distintos casos tienen diferentes varianzas. De la ecuación (37) vemos que cuánto mayor sea h_{ii}, menor será la varianza del e_i : mientras más cercano a uno sea h_{ii} más cercana a cero será la varianza del residuo de la observación i-ésima. Esto quiere decir que para observaciones con gran h_{ii}, $\widehat{Y}_i$ tenderá a estar cerca del valor observado Y_i, sin importar cuánto sea el valor Y_i observado. En el caso extremo e hipotético en que $h_{ii} = 1$, la recta ajustada sería forzada a pasar por el valor observado (X_i, Y_i).

3.1.3. Residuos estandarizados

Para hacer más comparables a los residuos entre sí, podemos dividir a cada uno de ellos por un estimador de su desvío estándar, obteniendo lo que se denominan *residuos estandarizados*:

$$rest_i = \frac{e_i}{\sqrt{\widehat{\sigma}^2 (1 - h_{ii})}}. \tag{38}$$

Recordemos que el estimador de σ^2 bajo el modelo de regresión está dado por

$$\widehat{\sigma}^2 = \frac{SS\text{Res}}{n-2}$$

Puede probarse que los residuos estandarizados tienen media poblacional cero (igual que los residuos), y varianza poblacional igual a uno, es decir

$$\begin{aligned} E(rest_i) &= 0 \\ Var(rest_i) &= 1, \quad \text{para todo } i. \end{aligned}$$

3.1.4. Los residuos cuando el modelo es correcto

Para chequear que los supuestos del modelo lineal son apropiados para un conjunto de datos, suelen hacerse una serie de gráficos. El más importante es el scatter plot de residuos versus la covariable. Esto se conoce como gráfico de residuos (o residual plot). En el caso de regresión lineal simple, los valores ajustados o predichos $\widehat{Y}_i$ representan un cambio de escala lineal respecto de los valores X_i ya que $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$. Luego, es equivalente al gráfico recién descripto el scatter plot de residuos versus los valores ajustados. ¿Cómo debe lucir este gráfico si el modelo es correcto?

1. Puede probarse que $E(e \mid X_1, \ldots, X_n) = 0$. Esto quiere decir que el scatter plot de los residuos versus las X debe estar centrado alrededor del cero (de la recta horizontal de altura cero).

2. Mencionamos que cuando el modelo es correcto, $Var(e_i \mid X_1, \ldots, X_n) = \sigma^2 (1 - h_{ii})$. Luego el gráfico de residuos versus la covariable debería mostrar menor variabilidad para los valores de X más alejados de la media muestral (serán los que tengan mayor leverage h_{ii}). Por este motivo, suele ser más frecuente graficar los residuos estandarizados versus la covariable. En ese caso, deberíamos ver la misma variabilidad para los distintos valores de la covariable.

3. Los residuos de distintas observaciones están correlacionados entre sí, pero esta correlación no es muy importante, no será visible en los gráficos de residuos.

En resumen, si el modelo es correcto, el gráfico de los residuos versus predichos o versus la covariable debería lucir como una nube de puntos sin estructura, ubicada alrededor del eje horizontal.

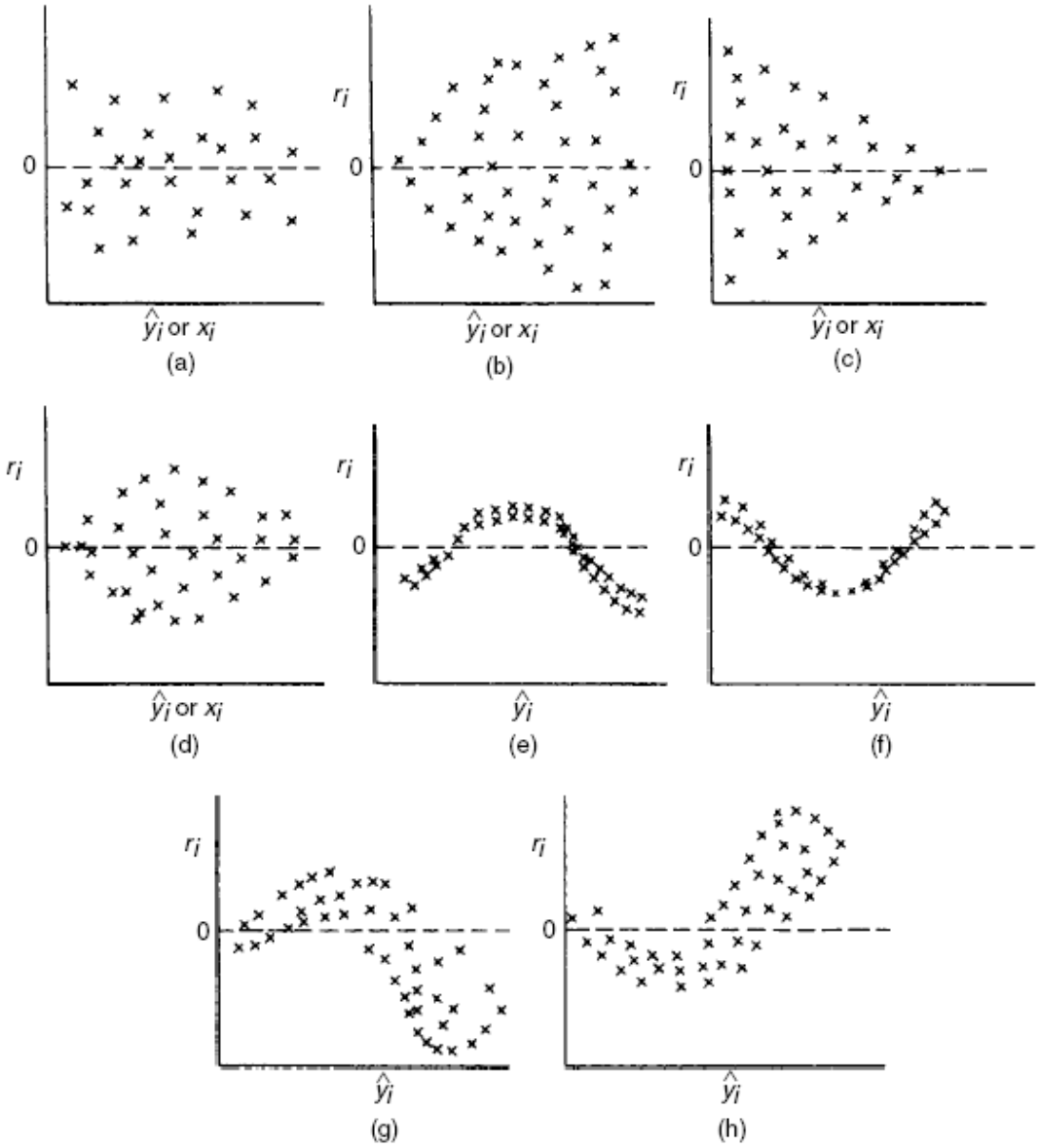
3.1.5. Los residuos cuando el modelo es incorrecto

En el caso en el que el modelo es incorrecto, el gráfico de residuos (o de residuos estandarizados) versus la variable predictora (o versus los valores predichos) suele tener algún tipo de estructura. En la Figura 25 se ven varios de estos posibles scatter plots (algo idealizados, claro).

El primero de ellos es una nube de puntos sin estructura que indica que no hay problemas con el modelo ajustado. De las Figuras 25(b) a 25(d) inferiríamos que el supuesto de homogeneidad de varianzas no se satisface: la varianza depende de la variable graficada en el eje horizontal. Las Figuras 25(e) a 25(h) son indicadoras de que se viola el supuesto de linealidad de la esperanza condicional, lo cual nos lleva a pensar que el vínculo entre el valor esperado de la variable respuesta Y y la covariable se ve mejor modelado por una función más complicada que la lineal (lo que genéricamente suele denominarse una curva). Las dos últimas figuras, las 25(g) y 25(h) sugieren la presencia simultánea de curvatura y varianza no constante.

En la práctica, los gráficos de residuos no son tan claros como estos... Es útil recordar que aún cuando todos los datos satisficieran todos los supuestos, la variabilidad muestral podría hacer que el gráfico tuviera pequeños apartamientos de la imagen ideal.

Figura 25: Gráficos de residuos: (a) nube de datos sin estructura, (b) varianza que crece con X (forma de megáfono abierto a la derecha), (c) varianza que decrece con X (forma de megáfono abierto a la izquierda), (d) varianza que depende de la covariable, (e)-(f) no linealidad, (g)-(h) combinación de no linealidad y función de varianza no constante. Fuente: Weisberg [2005], pág. 172.



3.1.6. Los residuos en el ejemplo

La Figura 26 muestra el gráfico de residuos en el ejemplo de los 100 bebés de bajo peso. Por ejemplo, el primer dato observado ($i = 1$) corresponde a un bebé de 29 semanas de gestación cuyo perímetro cefálico fue de 27 cm. El valor predicho para este caso es

$$\widehat{Y}_1 = 3{,}9143 + 0{,}7801 \cdot 29 = 26{,}537$$

y el residuo asociado a esa observación es

$$e_1 = Y_1 - \widehat{Y}_1 = 27 - 26{,}537 = 0{,}463,$$

como ya habíamos calculado. Luego, el punto $(26{,}537, 0{,}463)$ será incluido en el gráfico, que es un scatter plot de los puntos $\left(\widehat{Y}_i, e_i\right)$ para las 100 observaciones de la muestra.

En él vemos que hay un residuo en particular que es un poco más grande que el resto: este punto corresponde a la observación 31, que corresponde a un bebé cuya edad gestacional es de 31 semanas y cuyo perímetro cefálico es de 35 centímetros. De acuerdo al modelo, el valor predicho para su perímetro cefálico sería

$$\widehat{Y}_{31} = 3{,}9143 + 0{,}7801 \cdot 31 = 28{,}097$$

un valor mucho menor que el observado, por lo tanto el residuo resulta grande

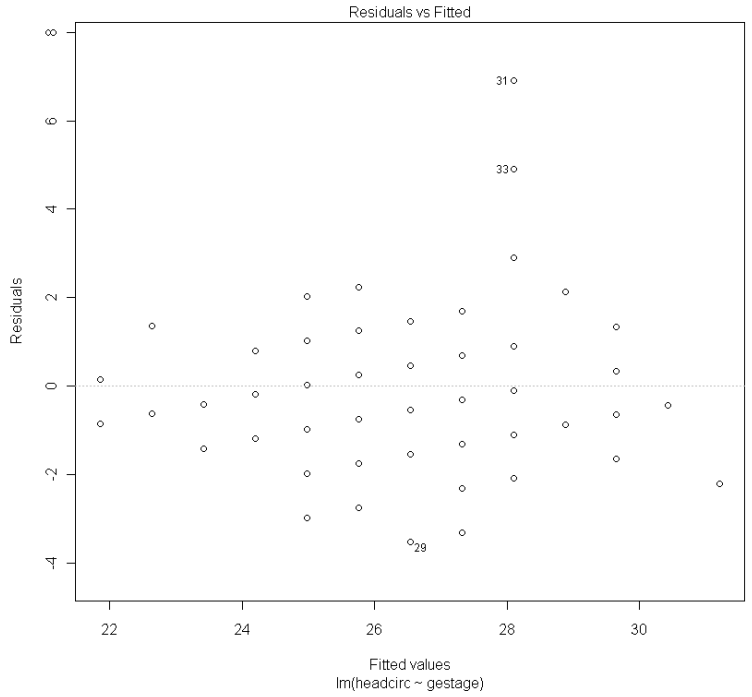
$$e_{31} = Y_{31} - \widehat{Y}_{31} = 35 - 28{,}097 = 6{,}903.$$

Podemos probar sacar este punto de la muestra, volver a realizar el ajuste y luego comparar los dos modelos para medir el efecto del punto en la estimación de los coeficientes de la recta. No lo haremos aquí puesto que en las secciones subsiguientes propondremos otros modelos que ajustarán mejor a nuestros datos. En cuanto al gráfico de residuos, este no muestra evidencia de que el supuesto de homoscedasticidad sea violado, o que haya algun tipo de curvatura en el vínculo entre los residuos y los predichos, indicando que el modelo ajusta bien a los datos.

3.1.7. ¿Cómo detectar (y resolver) la curvatura?

Para ayudarnos a decidir si un gráfico de residuos corresponde (o no) a una nube de puntos es posible hacer un test de curvatura. El más difundido es el test de no aditividad de Tuckey, que no describiremos aquí. Sin embargo, sí diremos que un remedio posible al problema de la curvatura consiste en transformar alguna de las variables X o Y (o ambas), y luego proponer un modelo lineal para las variables transformadas. Hay técnicas que ayudan a decidir qué transformaciones de los datos puede ser interesante investigar. Las transformaciones de Box-Cox son las más difundidas de estas técnicas, ver Kutner, Nachtsheim, Neter, y Li [2005].

Figura 26: Gráfico de residuos versus valores ajustados para el ajuste lineal de perímetro cefálico en función de la edad gestacional, en el caso de los 100 bebés de bajo peso.



Otra posibilidad consiste en proponer modelos más complejos que contemplen un vínculo más general entre X e Y, por ejemplo

$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2.$$

Es posible estudiar estos modelos como un caso particular de los modelos de regresión lineal, pero con dos covariables (X y X^2), lo cual nos lleva a tratarlos dentro de los modelos de regresión múltiple, que presentaremos más adelante.

3.1.8. ¿Qué hacer si la varianza no es constante?

En un gráfico de residuos, una función de la varianza no constante puede indicar que el supuesto de varianza constante es falso. Hay por lo menos cuatro remedios básicos en este caso, que describiremos siguiendo a Weisberg [2005], Sección 8.3. El primero es el uso de una *transformación estabilizadora de la varianza* para transformar a las Y, ya que el reemplazo de Y por `Ytransformada` puede inducir

varianza constante en la escala transformada. Una segunda opción es encontrar los pesos que podrían ser utilizados en los *mínimos cuadrados ponderados*. El método de mínimos cuadrados ponderados o pesados es una técnica estadística que ataca una versión más general del problema de regresión que hemos descrito hasta ahora. Lo presentamos a continuación, en su caso más simple. Seguimos trabajando bajo el supuesto de linealidad de la esperanza

$$E\left(Y \mid X = x_i\right) = \beta_0 + \beta_1 x_i,$$

pero ahora relajamos el supuesto de que la función de varianza $Var(Y \mid X)$ sea la misma para todos los valores de X. Supongamos que podemos asumir que

$$Var\left(Y \mid X = x_i\right) = Var\left(\varepsilon_i\right) = \frac{\sigma^2}{w_i}$$

donde $w_1, \ldots, w_n$ son números positivos conocidos. La función de varianza todavía queda caracterizada por un único parámetro desconocido σ^2, pero las varianzas pueden ser distintas para distintos valores de X. Esto nos lleva al método de mínimos cuadrados pesados o ponderados (en inglés *weighted least squares*, o wls) en vez del método usual de mínimos cuadrados (*ordinary least squares*, ols) para obtener estimadores. En este caso, se buscan los valores de los parámetros que minimizan la función

$$g_{wls}\left(a, b\right) = \sum_{i=1}^{n} w_i \left(Y_i - \left(a + bX_i\right)\right)^2.$$

Existen expresiones explícitas para los parámetros estimados con este método, y los softwares más difundidos realizan el ajuste. En las aplicaciones, por supuesto, se agrega la complejidad extra de elegir los pesos w_i que en general no vienen con los datos. Muchas veces se usan pesos empíricos, que se deducen de algunos supuestos teóricos que se tengan sobre las variables, por ejemplo. Si hubiera replicaciones, es decir varias mediciones de la variable respuesta realizadas para el mismo valor de la covariable, podría estimarse la varianza dentro de cada grupo y conseguirse de este modo pesos aproximados. También es posible usar *modelos de mínimos cuadrados generalizados*, en los que se estiman simultáneamente los parámetros del modelo y los pesos, que exceden por mucho estas notas (consultar por ejemplo Pinheiro y Bates [2000], Sección 5.1.2).

La tercera posibilidad es no hacer nada. Los estimadores de los parámetros, ajustados considerando una función de varianza incorrecta o mal especificada, son de todos modos insesgados, aunque ineficientes. Los tests e intervalos de confianza calculados con la función de varianza errada serán inexactos, pero se puede recurrir a métodos de bootstrapping para obtener resultados más precisos.

La última opción es usar modelos de regresión que contemplan la posibilidad de una función de varianza no constante que dependa de la media. Estos modelos se denominan *modelos lineales generalizados*, de los cuales por ejemplo, los modelos de regresión logística forman parte. Puede consultarse el texto clásico McCullagh y Nelder [1989] y también el libro de Weisberg [2005], Sección 8.3 y Sección 12.

3.1.9. ¿Cómo validamos la independencia?

Si las observaciones con las que contamos fueron producto de haber tomado una muestra aleatoria de sujetos de alguna población, entonces en principio, tendremos observaciones independientes. Algunas situaciones en las que este supuesto puede fallar se describen a continuación.

Los estudios en los cuales los datos se recolectan secuencialmente pueden dar lugar a observaciones que no resulten independientes. Lo mismo puede suceder en las determinaciones de laboratorio hechas secuencialmente en el tiempo, ya que pueden mostrar un cierto patrón, dependiendo de cómo funcionan los equipos, los observadores, etc. El modo de detección de estas situaciones suele ser graficar los residuos versus la secuencia temporal en la que fueron relevados.

Si los datos fueron obtenidos por dos observadores distintos A y B, podríamos esperar que las observaciones de un observador tiendan a parecerse más entre ellas. La manera de detectar que esto sucede es graficar las Y versus las X identificando los puntos de cada grupo. En ocasiones, la variabilidad debida a la regresión puede ser explicada por la pertenencia al grupo. Tampoco serán independientes las observaciones si varias de ellas fueron realizadas sobre los mismos sujetos (o animales). Si este fuera el caso, puede considerarse un modelo de regresión múltiple donde el operador (o el sujeto) entre como covariable. Nos ocuparemos de discutir esto más adelante, ya que los modelos correctos para este tipo de situaciones son los modelos de ANOVA con efectos aleatorios, o los modelos de efectos mixtos, que exceden el contenido de estas notas. Ver para ello, el libro de Pinheiro y Bates [2000].

3.1.10. ¿Cómo validamos la normalidad?

El supuesto de normalidad de los errores juega un rol menor en el análisis de regresión. Es necesario para realizar inferencias en el caso de muestras pequeñas, aunque los métodos de bootstrap (o resampleo) pueden usarse si este supuesto no está presente. El problema con las muestras pequeñas es que chequear el supuesto de normalidad a través de los residuos cuando no hay muchas observaciones es muy difícil. Los gráficos cuantil cuantil de los residuos (qq-plots) y los tests de normalidad realizados sobre ellos pueden ayudar en esta tarea. Hay varios tests posibles que ayudan a descartar la normalidad, entre ellos el test de Shapiro-

Wilks (que está esencialmente basado en el cuadrado de la correlación entre las observaciones ordenadas y sus valores esperados bajo normalidad), o el test de Kolmogorov-Smirnov, que están implementados en los paquetes.

En la práctica los supuestos de normalidad y homoscedasticidad nunca se cumplen exactamente. Sin embargo, mientras más cerca estén nuestros datos de los supuestos del modelo lineal, más apropiados serán los tests e intervalos de confianza que construyamos.

Para muestras grandes el supuesto de distribución normal no es crucial. Una versión extendida del Teorema Central del Límite garantiza que el estimador de mínimos cuadrados de la pendiente tiene distribución de muestreo aproximadamente normal cuando n es grande.

3.2. Outliers y observaciones influyentes

3.2.1. Outliers

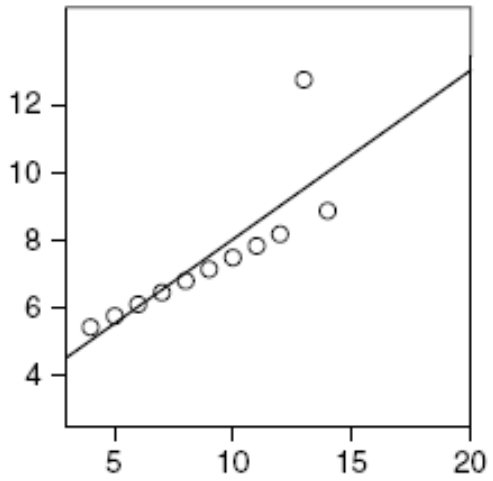
En algunos problemas, la respuesta observada para algunos pocos casos puede parecer no seguir el modelo que sí ajusta bien a la gran mayoría de los datos. Un ejemplo (de datos ficticios) puede verse en el scatter plot de la Figura 27. Los datos de este ejemplo sugieren que el modelo lineal puede ser correcto para la mayoría de los datos, pero uno de los casos está muy alejado de lo que el modelo ajustado le prescribe. Diremos que este dato alejado es un outlier. Observemos que el concepto de outlier (o sea, dato atípico) es un concepto relativo al modelo específico en consideración. Si se modifica la forma del modelo propuesto a los datos, la condición de ser outlier de un caso individual puede modificarse. O sea, un **outlier** es un caso que no sigue el mismo modelo que el resto de los datos. La identificación de estos casos puede ser útil. ¿Por qué? Porque el método de cuadrados mínimos es muy sensible a observaciones alejadas del resto de los datos. De hecho, las observaciones que caigan lejos de la tendencia del resto de los datos pueden modificar sustancialmente la estimación.

3.2.2. Un test para encontrar outliers

Si sospechamos que la observación i-ésima es un outlier podemos proceder del siguiente modo. Este procedimiento es clásico dentro de la regresión y corresponde a muchos otros procedimientos en estadística que son genéricamente conocidos como "*leave one out procedures*".

1. Eliminamos esa observación de la muestra, de modo que ahora tenemos una muestra con $n - 1$ casos.

Figura 27: Datos hipotéticos que muestran el desajuste de una observación al modelo ajustado.



2. Usando el conjunto de datos reducidos volvemos a estimar los parámetros, obteniendo $\widehat{\beta}_{0(i)}, \widehat{\beta}_{1(i)}$ y $\widehat{\sigma}^2_{(i)}$ donde el subíndice (i) está escrito para recordarnos que los parámetros fueron estimados sin usar la i-ésima observación.

3. Para el caso omitido, calculamos el valor ajustado $\widehat{Y}_{i(i)} = \widehat{\beta}_{0(i)} + \widehat{\beta}_{1(i)} X_i$. Como el caso i-ésimo no fue usado en la estimación de los parámetros, Y_i y $\widehat{Y}_{i(i)}$ son independientes. La varianza de $Y_i - \widehat{Y}_{i(i)}$ puede calcularse y se estima usando $\widehat{\sigma}^2_{(i)}$.

4. Escribamos

$$t_i = \frac{Y_i - \widehat{Y}_{i(i)}}{\sqrt{\widehat{Var}\left(Y_i - \widehat{Y}_{i(i)}\right)}},$$

la versión estandarizada del estadístico en consideración. Si la observación i-ésima sigue el modelo, entonces la esperanza de $Y_i - \widehat{Y}_{i(i)}$ debería ser cero. Si no lo sigue, será un valor no nulo. Luego, si llamamos δ a la esperanza poblacional de esa resta, $\delta = E\left(Y_i - \widehat{Y}_{i(i)}\right)$, y asumimos normalidad de los errores, puede probarse que la distribución de t_i bajo la hipótesis $H_0 : \delta = 0$ es una t de Student con $n-3$ grados de libertad, $t_i \sim t_{n-3}$ (recordar que hemos excluido una observación para el cálculo del error estándar que figura en el denominador, por eso tenemos un grado de libertad menos que con

los anteriores tests), y rechazar cuando este valor sea demasiado grande o demasiado pequeño.

Hay una fórmula computacionalmente sencilla para expresar a t_i sin necesidad de reajustar el modelo lineal con un dato menos, ya que es fácil escribir al desvío estándar estimado sin la observación i-ésima $\left(\widehat{\sigma}_{(i)}\right)$ en términos del leverage de la observación i-ésima (h_{ii}) y el desvío estándar estimado con toda la muestra $(\widehat{\sigma})$. Es la siguiente

$$t_i = \frac{e_i}{\widehat{\sigma}_{(i)}\sqrt{1-h_{ii}}} = rest_i\sqrt{\frac{n-3}{n-2-rest_i}} \tag{39}$$

donde el residuo estadarizado $rest_i$ lo definimos en la ecuación (38). Esta cantidad se denomina el **residuo estudentizado** i-ésimo. La ecuación (39) nos dice que los residuos estudentizados y los residuos estandarizados llevan la misma información, ya que pueden ser calculados uno en función de otro. Vemos entonces que para calcular los residuos estudentizados no es necesario descartar el caso i-ésimo y volver a ajustar la regresión (cosa que tampoco nos preocuparía mucho ya que es la computadora la que realiza este trabajo).

Para completar el test, nos queda únicamente decidir contra qué valor comparar el t_i para decidir si la i-ésima observación es o no un outlier. Si el investigador sospecha de antemano a realizar el ajuste que la observación i-ésima es un outlier lo justo sería comparar el valor absoluto de t_i con el percentil $1-\frac{\alpha}{2}$ de la t de student con $n-3$ grados de libertad. Pero es rara la ocasión en la que se sospecha de un dato antes de hacer el análisis. Si en cambio el analista hace el ajuste, luego computa los residuos estudentizados, y, a raíz de lo obtenido sospecha de aquella observación con mayor valor absoluto de t_i, entonces en el fondo está realizando n tests de significatividad, uno para cada observación. Para tener controlada la probabilidad de cometer un error de tipo I en alguno de los n tests (es decir, decidir falsamente que una observación que en realidad no es outlier sea declarada como tal), puede usarse un procedimiento conservativo conocido como método de Bonferroni para comparaciones múltiples. Este procedimiento propone rechazar H_0 : *ninguna de las n observaciones es un outlier*, versus H_1 : *hay al menos un outlier,* cuando alguno de los $|t_i|$ es mayor que el percentil $1-\frac{\alpha}{2n}$ de la t_{n-3}. Por ejemplo, si $n = 20$ (pensamos en una muestra con 20 observaciones) y nivel simultáneo 0,05, entonces en vez de comparar con el percentil 0,975 de una t_{17} que es 2,11, la comparación correcta es con el percentil $1-\frac{\alpha}{2n} = 1-\frac{0{,}05}{2\cdot 20} = 0{,}99875$ de una t_{17} que es 3,543.

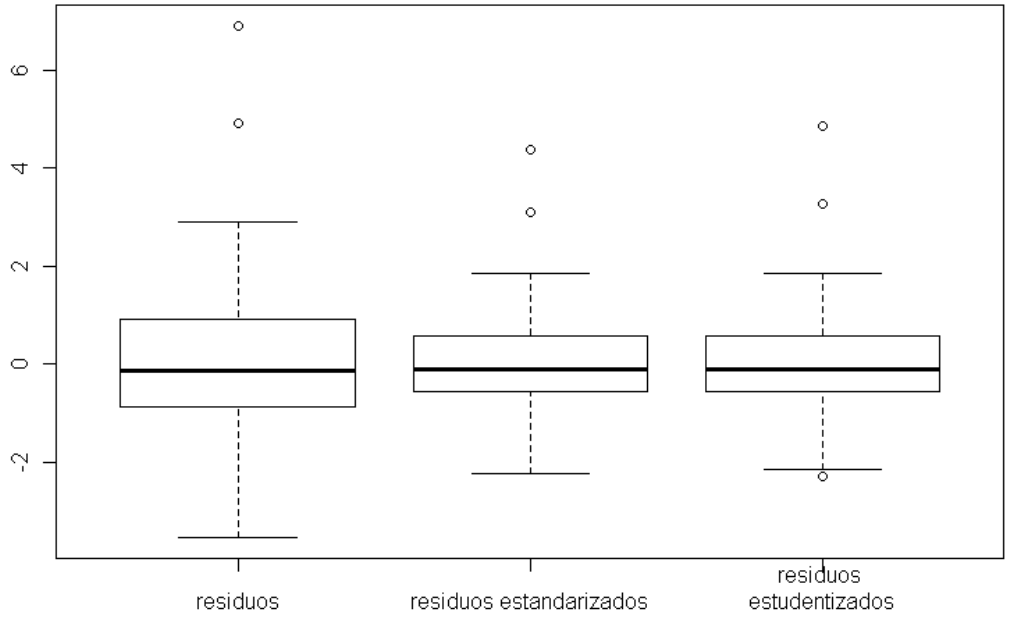
Apliquemos este test al ejemplo de los bebés de bajo peso.

Ejemplo 3.1 *En el caso de los 100 bebés, para detectar outliers a nivel* 0,05 *debemos computar el residuo estudentizado para cada caso, y compararlo con el percentil*

$$1-\frac{\alpha}{2n} = 1-\frac{0{,}05}{2\cdot 100} = 0{,}99975$$

de una t_{97}, que resulta ser 3,602. *El único residuo estudentizado cuyo valor absoluto sobrepasa este punto de corte es el correspondiente a la observación 31, que es* 4,857. *En la Figura 28 pueden verse los boxplots de los residuos, los residuos estandarizados y los residuos estudentizados para el ajuste de perímetro cefálico en función de la edad gestacional.*

Figura 28: Los boxplots de los residuos, los residuos estandarizados y los residuos estudentizados para el ajuste de perímetro cefálico en función de la edad gestacional en el ejemplo.



Este test ubica un outlier, pero no nos dice qué hacer con él. Cuando detectamos un outlier, sobre todo si es severo, es importante investigarlo. Puede tratarse de un dato mal registrado, o que fue mal transcripto a la base de datos. En tal caso podremos eliminar el outlier (o corregirlo) y analizar los casos restantes. Pero si el dato es correcto, quizás sea diferente de las otras observaciones y encontrar las causas de este fenómeno puede llegar a ser la parte más interesante del análisis. Todo esto depende del contexto del problema que uno esté estudiando. Si el dato es

correcto y no hay razones para excluirlo del análisis entonces la estimación de los parámetros debería hacerse con un método robusto, que a diferencia de mínimos cuadrados, no es tan sensible a observaciones alejadas de los demás datos.

Antes de terminar, correspondería hacer un alerta. Los residuos estudentizados son una herramienta más robusta que los residuos estandarizados para evaluar si una observación tiene un residuo inusualmente grande. Éste método para detectar outliers parece una estrategia muy apropiada. Y lo es... siempre que en la muestra haya a lo sumo un outlier. Pero, como todos los procedimientos de *leave one out,* puede conducirnos a conclusiones erradas si en la muestra hubiera más de un dato atípico, pues en tal caso, al calcular el residuo estudentizado de la observación i-ésima, la otra (u otras) observaciones atípicas aún presentes en la muestra podrían tergiversar el ajuste del modelo $\widehat{Y}_{i(i)}$ o la estimación del desvío estándar $\widehat{\sigma}_{(i)}$, alterando la distribución de los residuos estudentizados resultantes. Por eso, una estrategia todavía mejor para detectar la presencia de outliers que el estudio de los residuos estudentizados, es comparar el ajuste obtenido por cuadrados mínimos con el ajuste al modelo lineal que proporciona un método robusto, como describiremos en la Sección 3.2.4.

3.2.3. Observaciones influyentes

Estudiar la influencia de las observaciones es, de alguna manera, estudiar los cambios en el análisis cuando se omiten uno o más datos (siempre una pequeña porción de los datos disponibles). La idea es descubrir los efectos o la influencia que tiene cada caso en particular comparando el ajuste obtenido con toda la muestra con el ajuste obtenido sin ese caso particular (o sin esos pocos casos particulares). Una observación se denomina **influyente** si al excluirla de nuestro conjunto de datos la recta de regresión estimada cambia notablemente. Ejemplificaremos los conceptos en forma gráfica.

En la Figura 29 se observan scatter plots de cuatro conjuntos de 18 datos cada uno. En el gráfico (1), el conjunto de datos no presenta ni puntos influyentes ni outliers, ya que todas las observaciones siguen el mismo patrón. En los restantes tres gráficos se conservaron 17 de las observaciones del gráfico (1) y se intercambió una de ellas por los puntos que aparecen indicados como A, B y C en los respectivos scatter plots, y que son puntos atípicos en algun sentido, es decir, puntos que no siguen el patrón general de los datos. No todos los casos atípicos tendrán una fuerte influencia en el ajuste de la recta de regresión.

En la Figura 29 (2), entre las observaciones figura una que rotulamos con la letra A. El caso A puede no ser muy influyente, ya que hay muchos otros datos en la muestra con valores similares de X que evitarán que la función de regresión se desplace demasiado lejos siguiendo al caso A. Por otro lado, los casos B y C ejercerán una influencia muy grande en el ajuste, ya que como vimos en las Sec-

ciones 3.1.1 y 3.1.2 el leverage de ambas será bastante alto. Mientras mayor sea el leverage de la observación, menor será la variabilidad del residuo, esto quiere decir que para observaciones con gran leverage, el valor predicho tendrá que estar cerca del valor observado. Por eso se dice que tienen un alto grado de apalancamiento, o que cada uno de ellos es un punto de alta palanca. Luego la recta ajustada se verá matemáticamente obligada a acercarse a dichas observaciones, alejándose para ello, de los demás datos.

En la Figura 29 (3) aparece una observación indicada con B. Esta observación será influyente en el ajuste, pero como sigue el patrón lineal de los datos (o sea, sigue la estructura de esperanza condicional de Y cuando X es conocida que tienen el resto de los datos) no hará que la recta estimada cuando el punto está en la muestra varíe mucho respecto de la recta estimada en la situación en la que no está, pero reforzará (quizá artificialmente) la fuerza del ajuste observado: reforzará la significatividad de los tests que se hagan sobre los parámetros.

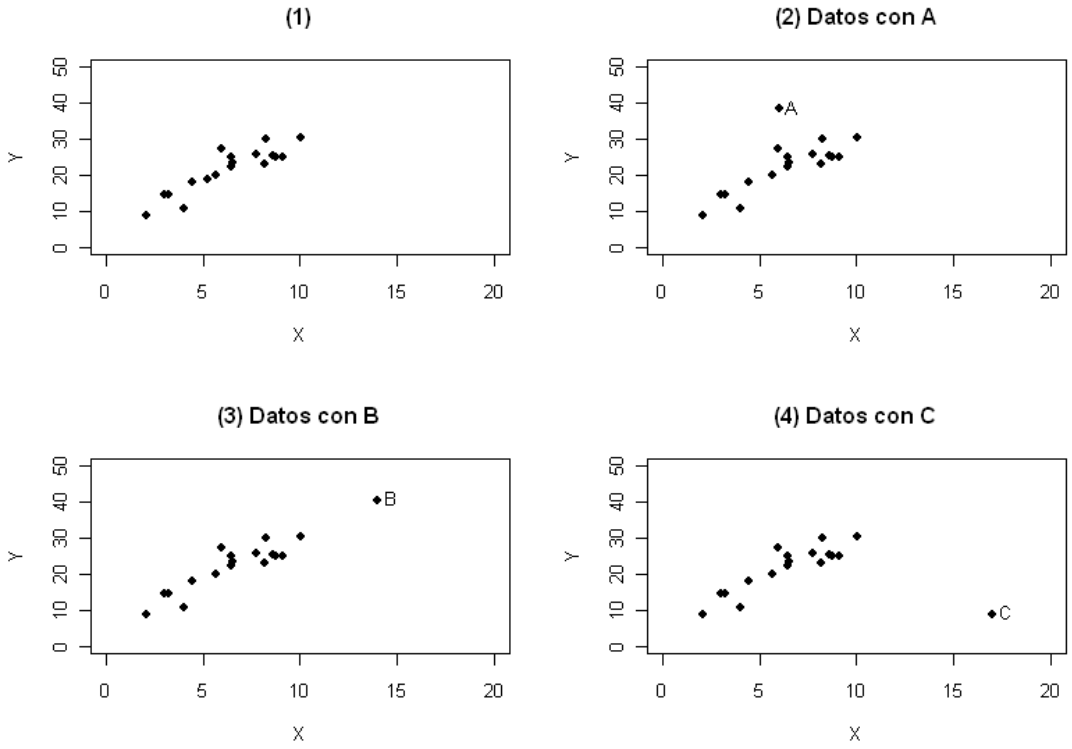
La Figura 29 (4) presenta la observación C. Esta observación será muy influyente en el ajuste, arrastrando a la recta estimada a acercarse a ella. Como no sigue la misma estructura de esperanza condicional que el resto de las observaciones, la recta ajustada en este caso diferirá mucho de la que se ajusta a los datos de la Figura 29 (1). Sin embargo, si una vez realizado el ajuste intentamos identificar este punto mirando las observaciones de mayores residuos (o residuos estandarizados) es posible que no la detectemos (dependerá de cuán extrema sea) ya que al arrastrar la recta hacia ella, tendrá un residuo mucho menor que el que tendría si usáramos la recta que ajusta a los datos del gráfico (1).

Constatemos que lo afirmado antes es cierto, buscando la recta que mejor ajusta a cada conjunto de datos, por mínimos cuadrados. A continuación figuran las salidas del `R` a los cuatro ajustes, y en la Figura 30 nuevamente los scatter plots de los cuatro conjuntos de datos, con las rectas ajustadas superpuestas.

Ajuste de mínimos cuadrados a los datos de la Figura 29 (1)

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.4063     2.0364   3.146  0.00625
xx                2.3987     0.3038   7.895 6.58e-07


Residual standard error: 2.899 on 16 degrees of freedom
Multiple R-squared: 0.7957, Adjusted R-squared: 0.783
F-statistic: 62.33 on 1 and 16 DF,  p-value: 6.579e-07

> confint(lm(yy~xx))
```

Figura 29: Scatter plot de 4 conjuntos de datos (hay 18 observaciones en cada uno): El gráfico (1) no presenta ni puntos influyentes ni outliers, (2) entre las observaciones figura la indicada con A, que es un outlier, no muy influyente, (3) en este gráfico figura la observación B, influyente pero no outlier, (4) este gráfico muestra la observación C, simultáneamente influyente y atípica.




```
               2.5 %    97.5 %
(Intercept) 2.089294 10.723305
xx          1.754633  3.042795
```

Ajuste de mínimos cuadrados a los datos de la Figura 29 (2)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.8387     3.6856   2.127 0.049338
xx            2.3281     0.5469   4.257 0.000602
```

```
Residual standard error: 5.184 on 16 degrees of freedom
Multiple R-squared: 0.5311, Adjusted R-squared: 0.5018
F-statistic: 18.12 on 1 and 16 DF,  p-value: 0.000602

> confint(lm(yy~xx))
                2.5 %    97.5 %
(Intercept) 0.02561661 15.651834
xx          1.16881319  3.487375
```

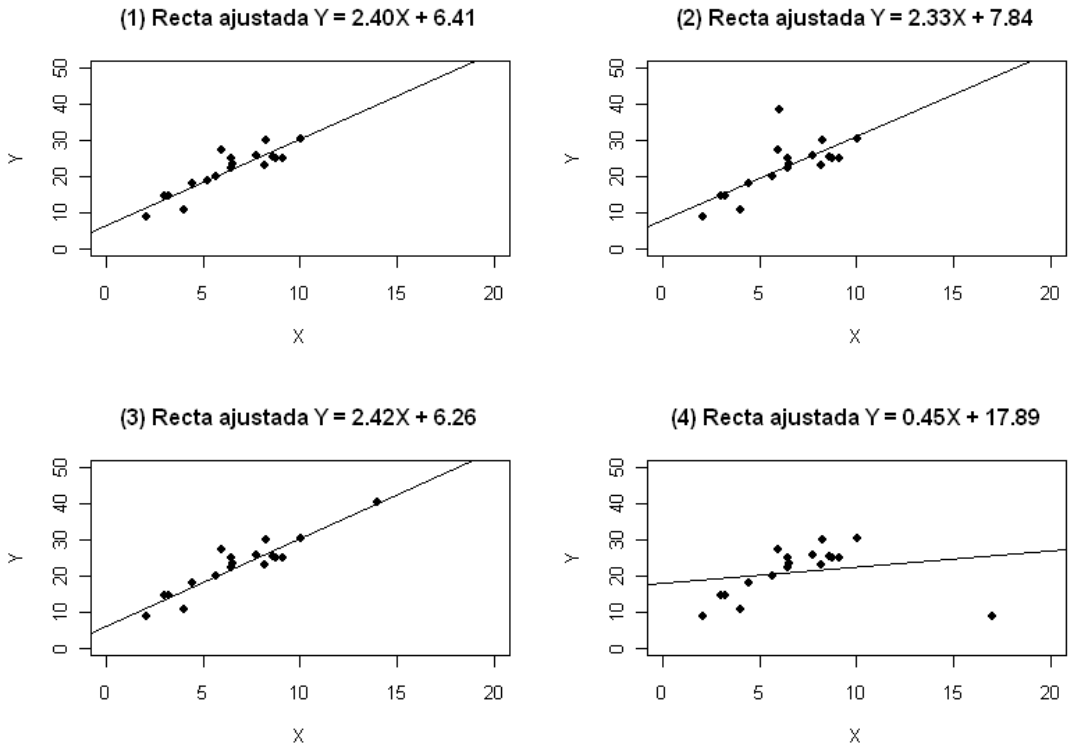
Ajuste de mínimos cuadrados a los datos de la Figura 29 (3)

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.2614     1.7778   3.522  0.00283
xx                2.4242     0.2412  10.049 2.57e-08

Residual standard error: 2.9 on 16 degrees of freedom
Multiple R-squared: 0.8632, Adjusted R-squared: 0.8547
F-statistic:   101 on 1 and 16 DF,  p-value: 2.566e-08

> confint(lm(yy~xx))
              2.5 %   97.5 %
(Intercept) 2.492573 10.03017
xx          1.912797  2.93559
```

Ajuste de mínimos cuadrados a los datos de la Figura 29 (4)

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      17.8872     3.8042   4.702  0.00024
xx                0.4471     0.4933   0.906  0.37823

Residual standard error: 6.91 on 16 degrees of freedom
Multiple R-squared: 0.04883, Adjusted R-squared: -0.01062
F-statistic: 0.8214 on 1 and 16 DF,  p-value: 0.3782

> confint(lm(yy~xx))
                2.5 %    97.5 %
(Intercept)  9.8226420 25.951836
xx          -0.5986414  1.492747
```

Figura 30: Nuevamente los scatter plots de los 4 conjunto de datos, esta vez con las rectas ajustadas.



Una vez realizado el ajuste vemos que se verifica lo anticipado. Las pendientes de las rectas estimadas en los 3 primeros gráficos no difieren demasiado entre sí, en el gráfico (2) la ordenada al origen es mayor ya que la observación A está ubicada muy por encima de los datos. La recta estimada en (3) pasa casi exactamente por el dato B y la significatividad del test para la pendiente aumenta en este caso, comparada con la del gráfico (1). Además también se incrementa el R cuadrado, que pasa de 0,79 en (1) a 0,86 en (3). En el gráfico (4) vemos que la recta ajustada difiere completamente de la recta estimada para el conjunto (1), de hecho la pendiente que era significativa para los datos del gráfico (1) deja de serlo en este caso. Vemos que la observación C arrastró la recta hacia ella. La observación C es la que más tergiversó las conclusiones del ajuste lineal.

Un comentario más que habría que hacer con respecto a la influencia es que

en este caso hemos presentado un ejemplo muy sencillo donde para cada conjunto de datos hay un sólo dato sospechoso. En las situaciones prácticas, cuando hay más de un dato anómalo en un conjunto de datos, esta presencia simultánea puede enmascararse: la técnica de sacar las observaciones de a una muchas veces no logra detectar los problemas. En regresión simple nos salva un poco el hecho de que podemos graficar muy bien los datos. No será esta la situación en regresión múltiple, por lo que se vuelve importante tener medidas cuantitativas que permitan medir el grado de influencia (al menos potencial) que tiene cada dato en un conjunto de datos.

3.2.4. Alternativa: comparación con un ajuste robusto

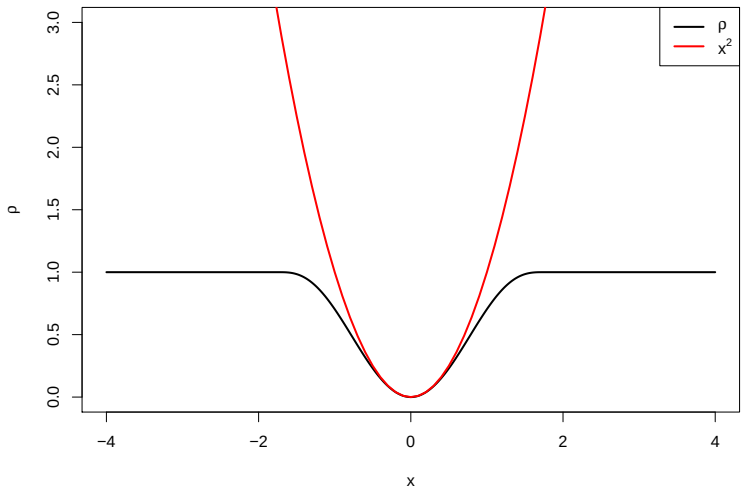
Como hemos dicho antes, el método de cuadrados mínimos como estrategia para encontrar estimadores de los parámetros del modelo lineal, resulta ser muy sensible a observaciones alejadas del resto de los datos. También vimos a través de un ejemplo sencillo, cómo una sola observación apartada del resto puede modificar sustancialmente la estimación realizada. Los métodos de estimación que son más resistentes a la influencia de observaciones apartadas del resto se denominan *métodos robustos de estimación*. De hecho, puede probarse matemáticamente, que la influencia que una sola observación atípica puede tener sobre los estimadores de mínimos cuadrados es (potencialmente) ilimitada. Este déficit no está en el modelo lineal, sino en el método de estimación elegido para ajustarlo: el método de mínimos cuadrados que propone como función para medir el desajuste (se denomina *función de pérdida*) a la suma del cuadrado de los residuos. Si en vez de tomar esa función de pérdida, eligiéramos otra, podríamos subsanar este déficit que tienen los estimadores de mínimos cuadrados, de volverse ilimitadamente sensibles a una observación atípica. ¿Qué forma debería tener la función de pérdida propuesta para que el estimador resultara robusto? Hay varias alternativas. Por un lado, la función debiera ser insensible a observaciones extremadamente grandes (o residuos enormes), y por lo tanto crecer mucho menos que el cuadrado cuando la miramos suficientemente lejos del cero. Por otro lado, si los datos de la muestra siguieran el modelo de regresión con errores normales, querríamos que el estimador que el método robusto calcula se parezca al de mínimos cuadrados, (esta propiedad en estadística se denomina *alta eficiencia*) por lo que la función de pérdida debiera parecerse al cuadrado para valores muy cercanos a cero. Es por esto que en vez de usarse el cuadrado se suele utilizar una función de pérdida del estilo (comparar con la función exhibida en (8))

$$g\left(a,b\right)=\sum_{i=1}^{n}\rho\left(\frac{Y_i-\left(a+bX_i\right)}{s_n}\right) \tag{40}$$

donde $\rho : \mathbb{R} \to \mathbb{R}$ es una función acotada, creciente y simétrica alrededor del cero, y s_n es un estimador de escala que juega el papel de σ en el modelo clásico de regresión.

Una posibilidad es ajustar una recta usando un procedimiento de ajuste robusto, por ejemplo un MM-estimador de regresión, propuesto por Yohai [1987]. En `R`, esto está programado dentro de la rutina `lmrob` en el paquete `robustbase` de `R`. La estimación se hace en tres etapas, se propone un estimador inicial de los parámetros, a partir de él se estima a s_n y finalmente se obtienen los estimadores de los parámetros a partir de ellos, minimizando la función objetivo (40). La Figura 31 muestra una posible selección de la función ρ.

Figura 31: Ejemplo de una función ρ en la familia bicuadrada (en negro) comparada con la cuadrática (en rojo).



Existen muchas otras propuestas de estimadores robustos para regresión, por ejemplo LMS (*least median of squares*), LTS (*least trimmed squares*), $\tau-$estimadores de regresión, y casi todas están implementadas en `R`. Entre ellas, otra buena opción para tener un ajuste robusto altamente robusto y eficiente implementado en `R`, que no comentaremos aquí, es la rutina `lmRob` del paquete `robust`. Una funte completa para consultarlas es el libro de Maronna, Martin, y Yohai [2006]. La dificultad con los métodos robustos de ajuste radica en dos cuestiones. En primer lugar ya no es tan fácil (y en algunos casos no es posible) exhibir una fórmula cerrada que los compute, ni encontrar los mínimos absolutos de la función objetivo,

es decir los estimadores. Y por otro, no es sencillo dar la distribución de dichos estimadores, que nos permitirá calcular los p-valores para medir la significatividad de los tests. Sin embargo, utilizando algoritmos diseñados para hallar óptimos de funciones (el IRWLS, por ejemplo) y métodos de remuestreo apropiados, que explotan la capacidad de cómputo de las computadoras actuales, sí pueden obtenerse tests e intervalos de confianza, como veremos en las salidas que presentamos a continuación.

La salida del ajuste con `lmrob` a los datos de la Figura 29 (4) aparece en la Tabla 16. En ella vemos que los valores de la pendiente y ordenada al origen estimados resultan ser muy parecidos a los que se obtienen al ajustar por el método de mínimos cuadrados a los datos (1), que no están contaminados con outliers. Vemos que en lo que a la estimación de los parámetros del modelo lineal se refiere, el método robusto prácticamente ignora a la observación C que estaba distorsionando el ajuste clásico. Y que esto lo hace automáticamente, sin que tengamos que informarle que se trata de una observación potencialmente problemática. Vemos también que el ajuste robusto da p-valores e intervalos de confianza muy parecidos a los que proporciona el ajuste clásico; lo mismo sucede con el cálculo de R^2.

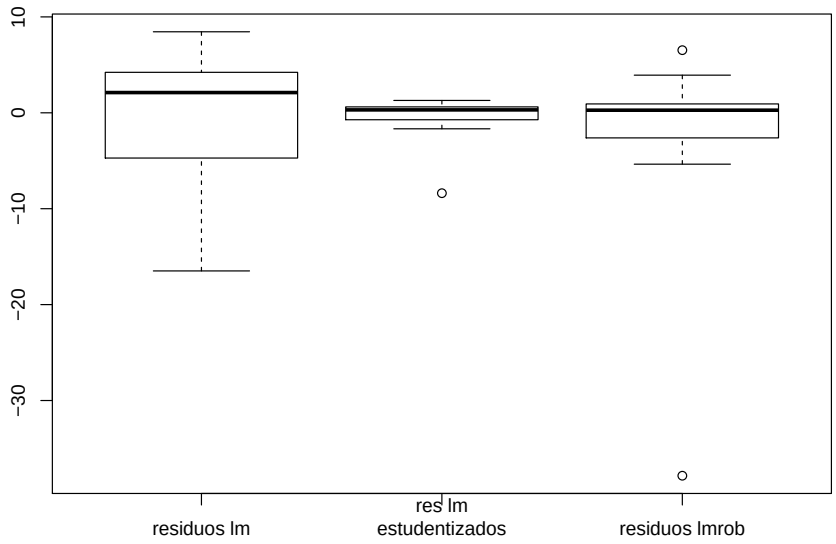
A posteriori del ajuste robusto, analizando los residuos identificamos rápidamente a la observación C como outlier, ya que el ajuste robusto no arrastra a la recta estimada y la magnitud del residuo refleja la distancia entre el valor observado y el predicho por el modelo, como puede verse en la Figura 32 donde aparecen los boxplots de los residuos y residuos estudentizados del ajuste por mínimos cuadrados, y los residuos del ajuste robusto. En los residuos del ajuste por mínimos cuadrados, no identificamos ninguna observación como outlier. El boxplot de residuos estudentizados muestra la presencia de una observación atípica. Sin embargo, vemos en el boxplot de los residuos del ajuste robusto que el outlier aparece mucho más extremo.

Tabla 16: Ajuste robusto dado por la función `lmrob` del paquete `robustbase`, a los datos de la Figura 29 (4)

```
> library(robustbase)
>  summary(lmrob(yy~xx))
 \--> method = "MM"

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5147     1.8381   3.544   0.0027 **
xx            2.3721     0.2676   8.866 1.43e-07 ***
---

Robust residual standard error: 3.149
```

```
Multiple R-squared:  0.7874,        Adjusted R-squared:  0.7741
Convergence in 8 IRWLS iterations

Robustness weights:
 observation 18 is an outlier with |weight| = 0 ( < 0.0056);
 3 weights are ~= 1. The remaining 14 ones are summarized as
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.6463  0.9142  0.9431  0.9167  0.9897  0.9959

>  confint(lmrob(yy~xx))
               2.5 %    97.5 %
(Intercept) 2.618108 10.411366
xx          1.804946  2.939334

> boxplot(residuals(lm(yy~xx)),studres(lm(yy~xx)),
residuals(lmrob(yy~xx)),names=c("residuos lm","res lm
 estudentizados" ,"residuos lmrob"))
```

Figura 32: Boxplot de los residuos para los datos de la Figura 29 (4), a la izquierda los residuos del ajuste por regresión lineal (`lm`), en el centro los residuos estudentizados del ajuste lineal (`lm`) y a la derecha los residuos del ajuste robusto propuesto (`lmrob`). El ajuste de `lm` arrastra la recta hacia el dato C enmascarando la presencia del outlier. El ajuste del `lmrob`, al no dejarse influenciar por una observación atípica permite identificar un outlier severo al estudiar los residuos.



Una propiedad interesante que tienen los MM estimadores de regresión, es que como parte del ajuste, se computan pesos para las observaciones. Si uno corre el ajuste de *mínimos cuadrados ponderados*, como describimos en la Sección con esos pesos en las observaciones, obtenemos los mismos estimadores robustos, como puede verse comparando las Tablas 16 y Tabla 17. En esta última tabla puede observarse que el peso que el ajuste robusto otorga a cada observación es prácticamente el mismo y casi uno, excepto para la última observación (la C) que recibe peso cero, es decir no interviene en el ajuste. Esta posibilidad de detectar que la observación es atípica de manera automática es muy útil, y lo será aún más cuando en vez de trabajar con una sola variable explicativa, lo hagamos con muchas, y el scatterplot se vuelva una herramienta incompleta.
Aún cuando uno esté interesado solamente en la recta de mínimos cuadrados, de todas formas conviene hacer un ajuste robusto a los datos. Si se observara una fuerte diferencia entre las conclusiones del método clásico (el ajuste de mínimos cuadrados) y el robusto, ésto sólo es señal de que existen observaciones influyentes y outliers entre los datos considerados. El estudio de los residuos del ajuste robusto permitirá la identificación de observaciones atípicas.

Tabla 17: Ajuste de mínimos cuadrados pesados a los datos de la Figura 29 (4), con los pesos calculados por el `lmrob`.

```
> ajusro<-(lmrob(yy~xx))
> robpesos<-ajusro$rweights
```

```
> robpesos
        1         2         3         4         5         6
0.9921401 0.9231724 0.9959484 0.9133037 0.9738549 0.9381117
        7         8         9        10        11        12
0.9935094 0.9992255 0.9170598 0.6463435 0.9825457 0.7538439
       13        14        15        16        17        18
0.9921665 0.9994115 0.9481115 0.9996075 0.8634395 0.0000000

> summary(lm(yy~xx,weights=robpesos))

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-4.6502 -2.1647  0.2717  0.9219  5.2523

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5147     1.9697   3.307  0.00479 **
xx            2.3721     0.2896   8.191 6.44e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.673 on 15 degrees of freedom
Multiple R-squared:  0.8173,        Adjusted R-squared:  0.8051
F-statistic: 67.09 on 1 and 15 DF,  p-value: 6.435e-07
```

3.2.5. ¿Cómo medir la influencia de una observación?

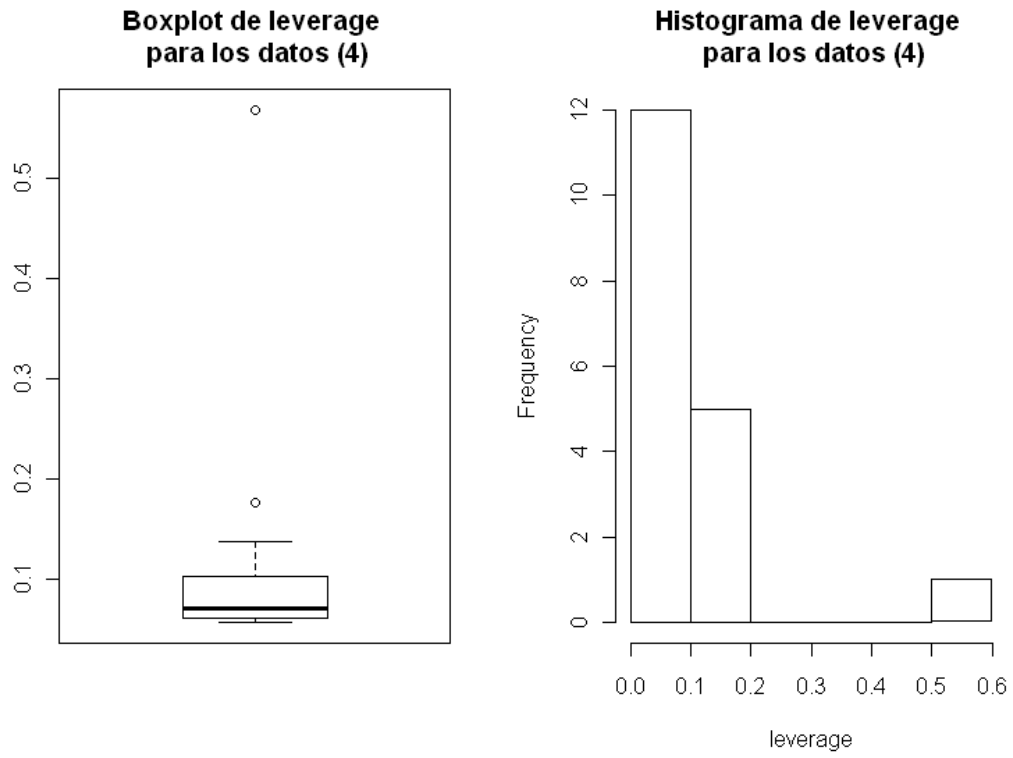
Tenemos dos medidas de influencia: el leverage y las distancias de Cook. El leverage lo definimos en la Sección 3.1.1. Pero cabe preguntarse cuan grande debe ser el leverage de una observación para declararla influyente. Se han sugerido diferentes criterios:

- Teniendo en cuenta que $\sum_{i=1}^{n} h_{ii} = 2$ y por lo tanto el promedio $\overline{h} = \frac{2}{n}$, un criterio es considerar potencialmente influyentes las observaciones con $h_{ii} > \frac{4}{n}$.
- Otro criterio es declarar potencialmente influyentes a aquellas observaciones cuyos leverages h_{ii} cumplen $h_{ii} > 0{,}5$ y evaluar o inspeccionar además los casos en que $0{,}2 < h_{ii} \leq 0{,}5$.
- Otro criterio es mirar la distribución de los h_{ii} en la muestra, en especial si

existen saltos en los valores de leverage de las observaciones. El modo más simple de hacerlo es a través de un box-plot o un histograma.

En la Figura 33 se exhiben el boxplot y el histograma de los leverage calculados para los datos de la Figura 29 (4). Hay un único dato con un leverage alto. Observemos que si hiciéramos lo mismo para los datos (3) obtendríamos algo muy parecido, ya que el leverage sólo depende de los valores de la covariable (y no de la variable respuesta). En ese sentido es una medida de influencia potencial de los datos. Los leverages para los restantes conjuntos de datos pueden verse en la Figura 34.

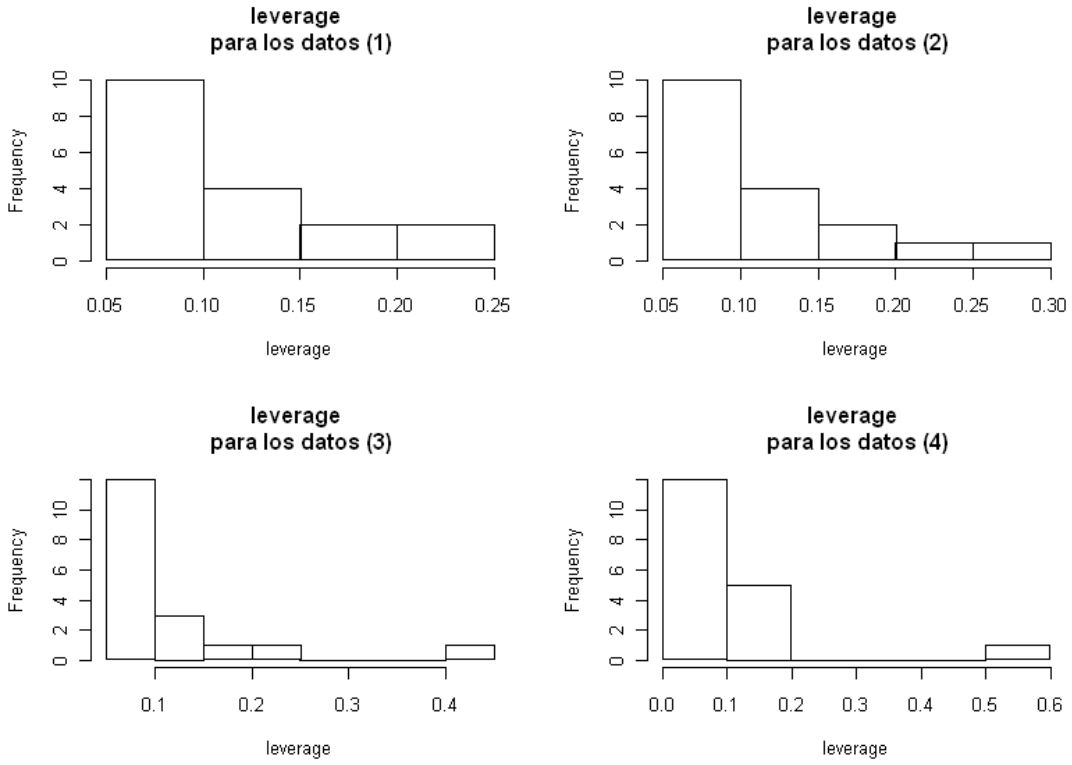
Figura 33: Boxplot e histograma para los leverage de los datos (4) graficados en la Figura 29.



La influencia de una observación depende de dos factores:

1. Cuán lejos cae el valor de Y de la tendencia general en la muestra para ese valor de X.

Figura 34: Histogramas de los leverage para los cuatro conjuntos de datos graficados en la Figura 29.



2. Cuán lejos se encuentra el valor de la variable explicativa de su media.

El leverage sólo recaba información de la situación descripta en 2). Una medida que toma en cuenta ambas facetas de una observación es la *Distancia de Cook*, definida por

$$D_i = \frac{\left(\widehat{Y}_{(i)i} - \widehat{Y}_i\right)^2}{2\widehat{\sigma}^2},$$

donde $\widehat{Y}_{(i)}$ corresponde al valor predicho para la i-ésima observación si se usaron las $n-1$ restantes observaciones para hacer el ajuste, como lo habíamos definido en la Sección 3.2.2 y $\widehat{Y}_i$ es el valor predicho para la i-ésima observación en el modelo ajustado con las n observaciones. Como en el caso de los residuos estudentizados, no es necesario recalcular el ajuste por mínimos cuadrados para calcular los D_i,

ya que otra expresión para ellos es la siguiente

$$D_i = \frac{1}{2}\,(rest_i)^2\,\frac{h_{ii}}{1-h_{ii}}.$$

La distancia de Cook se compara con los percentiles de la distribución F de Fisher con 2 y $n-2$ grados de libertad en el numerador y denominador, respectivamente (2 porque estamos estimando dos parámetros beta). El criterio para decidir si una observación es influyente es el siguiente:

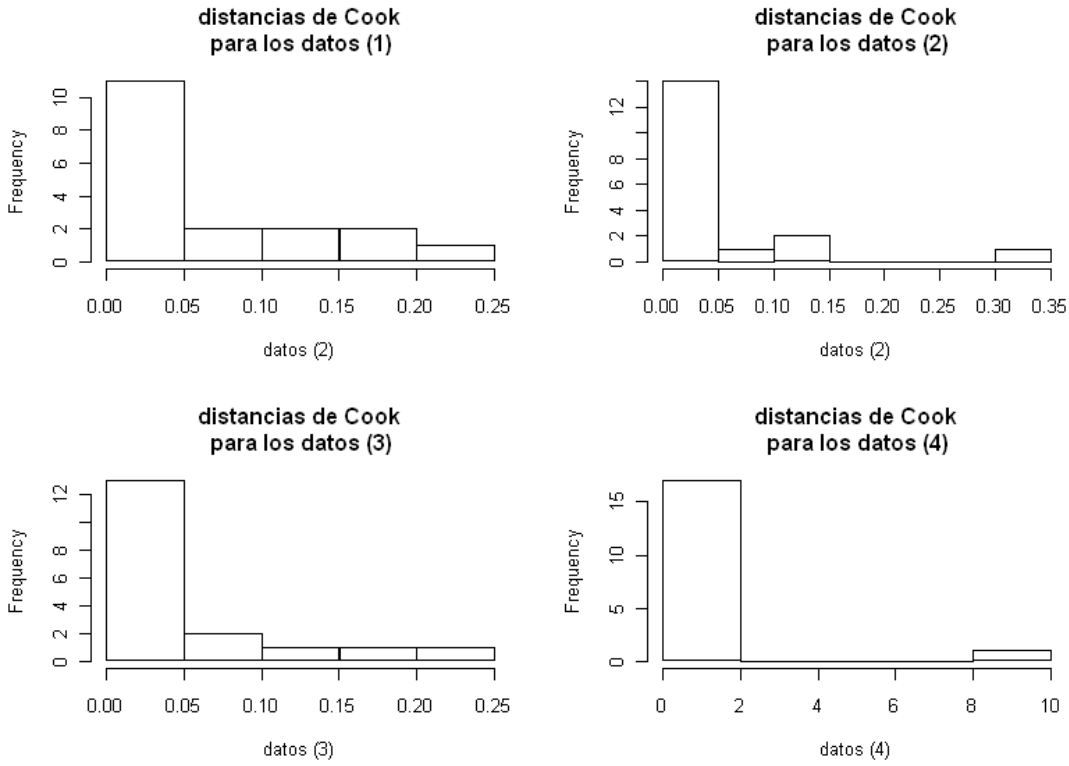
- Si $D_i <$ percentil 0,20 de la distribución $F_{2,n-2}$ entonces la observación no es influyente.
- Si $D_i >$ percentil 0,50 de la distribución $F_{2,n-2}$ entonces la observación es muy influyente y requerirá tomar alguna medida.
- Si D_i se encuentra entre el percentil 0,20 y el percentil 0,50 de la distribución $F_{2,n-2}$ se sugiere mirar además otros estadísticos.

Volviendo a los datos de la Figura 29, el percentil 0,20 de la distribución $F_{2,16}$ es 0,226 y el percentil 0,50 de la distribución $F_{2,16}$ es 0,724. Los histogramas de las distancias de Cook calculadas en este caso están en la Figura 35. Vemos que sólo en el caso de los datos (4) aparece una observación (la C) cuya distancia de Cook supera al percentil 0,50 de la distribución de Fisher, indicando que hay una observación muy influyente.
Existen otras medidas de influencia. Los DFfits y los DFbetas son medidas bastante estudiadas. Una referencia para leer sobre ellos es el libro de Kutner et al. [2005]. Los gráficos de variables agregadas (en el caso de regresión múltiple) pueden servir también para identificar observaciones influyentes, pueden verse en Weisberg [2005] secciones 3.1 y 9.2.4 o Kutner et al. [2005] sección 10.

3.2.6. Instrucciones de R para diagnóstico

```
ajuste <-lm(yy ~ xx)

#residuos
rr1<-residuals(ajuste)

#residuos estandarizados
rr2<-rstandard(ajuste)

#residuos estudentizados
rr3<-rstudent(ajuste)
```

Figura 35: Histogramas de las distancias de Cook para los datos de la Figura 29




```
# predichos o valores ajustados
ff1<-fitted(ajuste)

# grafico de residuos estudentizados vs predichos
plot(ff1,rr3,xlab="predichos",ylab="residuos")
abline(0,0)
title("Residuos estudentizados vs predichos")

##########################################
# test para encontrar outliers
ene<-length(rr3)
corte<-qt(1-0.05/(2*ene),df=ene-3)
(rr3 > corte)
sum(rr3 > corte)
```

```
#cuales son?
(1:ene)[rr3 > corte]

#p-valor, sin corregir
2*(1-pt(abs(rr3[3]),df=ene-3))
########################################

#leverage
hatvalues(ajuste)
hist(hatvalues(ajuste))

#para ver cuales son mayores que 0.2
lev <- hatvalues(ajuste)

#un criterio
(1:ene)[lev > 0.2]

#un criterio mas exigente
(1:ene)[lev > 4/ene]

#distancias de cook
dcook <- cooks.distance(ajuste)
hist(dcook)

#punto de corte
corted <- qf(0.5,2,ene-2)
(1:nn)[dcook > corted]

# qq plot de los residuos estandarizados, con la normal
qqnorm(rr2)
qqline(rr2)  #le agrega una recta que pasa por cuartil 1 y 3

plot(ajuste)  #hace varios graficos, entre ellos el QQplot de los
      #residuos estandarizados

library(car)
outlierTest(ajuste) # da pval de Bonferoni para obs mas extremas
qqPlot(ajuste, main="QQ Plot") #qq plot de los resid studentizados,
                               #con la t de student apropiada
```

```
##############################################
# ajuste robusto

library(robustbase)
ajusterob <- lmrob(yy ~ xx)
summary(ajusterob)

#residuos
resrob <- residuals(ajusterob)

boxplot(resrob,residuals(ajuste))

#pesos
hist(ajusterob$rweights)
boxplot(ajusterob$rweights)

plot(ajusterob)    # hace varios graficos
```

3.3. Ejercicios

Estos ejercicios se resuelven con el archivo `script_diagnostico.R`

Ejercicio 3.1 *Madres e hijas II. Archivo de datos* `heights.txt` *del paquete* `alr3`*. Continuando con el ejercicio 2.6, en el que proponemos ajustar el modelo lineal simple para explicar la altura de la hija,* `Dheight`*, a partir de la altura de la madre, llamada* `Mheight`*, como la variable predictora.*

(a) Hacer gráficos para evaluar la adecuación del modelo linel para explicar los datos.

(b) Compare el ajuste clásico con el ajuste robusto propuesto.

(c) Concluya respecto de la adecuación del modelo lineal en este caso.

Ejercicio 3.2 *Medidas del cuerpo V. Base de datos* `bdims` *del paquete* `openintro`*.*

(a) Realice gráficos de que le permitan evaluar los ajustes realizados en los ejercicios 2.1 y 2.2 con esta base de datos, tanto para explicar el peso por el contorno de cintura como el ajuste para explicar el peso por la altura. ¿Lo conforman estos modelos ajustados?

(b) Compare el ajuste clásico del modelo lineal con el ajuste robusto. ¿Cambian mucho los modelos ajustados? ¿Qué indica esto? No se desanime, este ejercicio sigue en el capítulo próximo.

Ejercicio 3.3 *Mamíferos, Parte V. Base de datos* `mammals` *del paquete* `openintro`*.*

(a) En el ejercicio 1.7 observamos que el scatter plot del peso del cerebro de un mamífero (`BrainWt`*) en función de su peso corporal (*`BodyWt`*) no se podía describir como una pelota de rugby más o menos achatada. Supongamos que no hubiéramos hecho el gráfico de dispersión, e intentemos ajustar un modelo lineal a los datos. Ajuste el modelo lineal simple que explica* `BrainWt` *en función de* `BodyWt`*. Luego realice el gráfico de residuos versus valores predichos. El gráfico de residuos estandarizados versus valores predichos. El de residuos estudentizados versus valores predichos. ¿Difieren mucho entre sí?*

(b) Use el test de outliers basado en los residuos estudentizados. Indique cuáles son las observaciones candidatas a outliers.

(c) Calcule los leverages. Identifique las observaciones candidatas a más influyentes según este criterio. Calcule las distancias de Cook, vea cuáles son las observaciones influyentes.

(d) Compare con el ajuste robusto.

(e) Finalmente, para el modelo de regresión propuesto en el ejercicio 2.9 para vincular los logaritmos en base 10 de ambas variables, haga un gráfico de residuos versus valores predichos, y algunos otros gráficos de diagnóstico. ¿Le parece que este modelo ajusta mejor a los datos?

Ejercicio 3.4 *Hacer un ajuste robusto a los datos de perímetro cefálico y edad gestacional. Comparar con el ajuste clásico. Identificar la presencia de outliers. ¿Son muy influyentes en el ajuste? Recordar que de todos modos este no es el último modelo que probaremos sobre estos datos.*

Ejercicio 3.5 *Resuelva el ejercicio domiciliario que figura en el Apéndice A.*

Ejercicio 3.6 *Resuelva el Taller 2 que figura en el Apéndice A.*

4. Regresión Lineal Múltiple

El modelo de regresión lineal múltiple es uno de los modelos más utilizados entre todos los modelos estadísticos.
En la mayoría de las situaciones prácticas en las que se quiere explicar una variable continua Y se dispone de muchas potenciales variables predictoras. Usualmente, el modelo de regresión lineal simple (es decir, con una sola variable predictora) provee una descripción inadecuada de la respuesta ya que suele suceder que son muchas las variables que ayudan a explicar la respuesta y la afectan de formas distintas e importantes. Más aún, en general estos modelos suelen ser muy imprecisos como para ser útiles (tienen mucha variabilidad). Entonces es necesario trabajar con modelos más complejos, que contengan variables predictoras adicionales, para proporcionar predicciones más precisas y colaborar en la cuantificación del vínculo entre ellas. En este sentido, el modelo de regresión múltiple es una extensión natural del modelo de regresión lineal simple, aunque presenta características propias que es de interés estudiar en detalle.
El modelo de regresión múltiple se puede utilizar tanto para datos observacionales como para estudios controlados a partir de ensayos aleatorizados o experimentales.

4.1. El modelo

La regresión múltiple es un modelo para la esperanza de una variable continua Y cuando se conocen variables explicativas o predictoras que denotaremos $X_1, X_2, \ldots, X_{p-1}$. Antes de formularlo en general, describiremos a modo ilustrativo la situación en la que se tienen dos variables predictoras (i.e. $p = 3$). En este caso, proponemos el siguiente modelo para la esperanza condicional de Y dado X_1 y X_2

$$E(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{41}$$

donde $\beta_0, \beta_1, \beta_2$ son constantes desconocidas que se denominan *parámetros* del modelo, o *coeficientes* de la ecuación. Muchas veces, por simplicidad, escribiremos $E(Y)$ en vez de $E(Y \mid X_1, X_2)$. El modelo se denomina "lineal" puesto que la esperanza de Y condicional a las X's depende linealmente de las covariables X_1 y X_2. Los coeficientes del modelo se estiman a partir de una muestra aleatoria de n observaciones (X_{i1}, X_{i2}, Y_i) con $1 \leq i \leq n$, donde Y_i es la variable respuesta medida en el i-ésimo individuo (o i-ésima repetición o i-ésima unidad experimental, según el caso), X_{i1} y X_{i2} son los valores de las variables predictoras en el i-ésimo individuo (o i-ésima repetición o i-ésima unidad experimental, según el caso). Una manera alternativa de escribir el modelo (41) en términos de las variables (en vez de sus valores esperados) es la siguiente

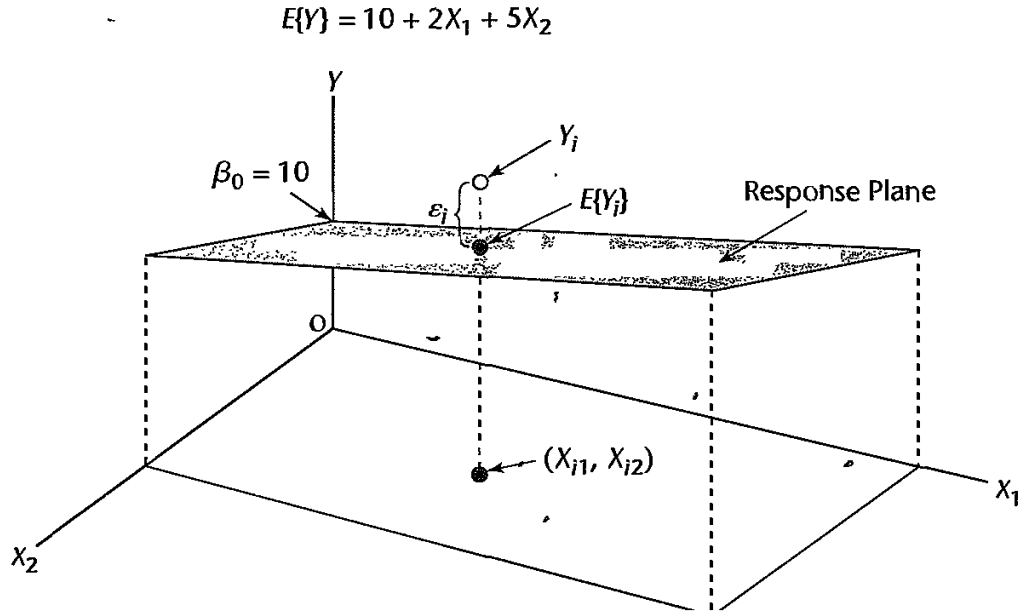
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \tag{42}$$

donde ε_i es el término del error para el individuo i-ésimo, que no es observable. A la ecuación (41) se la suele llamar función de respuesta. En analogía con la regresión lineal simple donde la función $E(Y \mid X) = \beta_0 + \beta_1 X_1$ es una recta, la función de regresión (41) es un plano. En la Figura 36 se representa una porción de la función de respuesta

$$E(Y \mid X_1, X_2) = 10 + 2X_1 + 5X_2. \tag{43}$$

Por supuesto, la única situación en la que podemos graficar es cuando $p \leq 3$ (dos o menos variables explicativas), es por eso que hemos comenzado con este caso.

Figura 36: En regresión lineal con dos variables explicativas la función de respuesta es un plano. Fuente Kutner et al. [2005], pág. 215.



Observemos que cualquier punto de la Figura 36 corresponde a una respuesta media $E(Y)$ para una combinación dada de X_1 y X_2. La Figura 36 también muestra una observación Y_i correspondiente a los niveles (X_{i1}, X_{i2}) de las dos variables predictoras. El segmento vertical entre Y_i y el gráfico de la función (el plano) de respuesta representa la diferencia entre Y_i y la media $E(Y_i) = E(Y_i \mid X_{i1}, X_{i2})$ de la distribución de probabilidad de Y para la combinación de (X_{i1}, X_{i2}). Por lo tanto, la distancia vertical entre Y_i y el plano de respuesta representa el término

de error $\varepsilon_i = Y_i - E(Y_i)$. En regresión lineal múltiple, a la función de respuesta también suele llamársela *superficie de regresión* o *superficie de respuesta.*

4.2. Significado de los coeficientes de regresión

Consideremos ahora el significado de los coeficientes en la función de regresión múltiple (42). El parámetro β_0 es el intercept u ordenada al origen del plano. Si dentro de los valores que estamos ajustando el modelo, se encuentra incluido el punto $X_1 = 0, X_2 = 0$, el origen de coordenadas, entonces β_0 representa la respuesta media $E(Y)$ en $X_1 = 0, X_2 = 0$. De lo contrario, β_0 no tiene ningún significado en particular como un término separado del modelo de regresión.
El parámetro β_1 indica el cambio en la respuesta media $E(Y)$ cuando aumentamos a X_1 en una unidad, manteniendo a X_2 constante (en cualquier valor). Del mismo modo, β_2 indica el cambio en la respuesta media $E(Y)$ cuando aumentamos a X_2 en una unidad, manteniendo a X_1 constante. En el ejemplo (43) graficado, supongamos que fijamos X_2 en el nivel $X_2 = 3$. La función de regresión (43) ahora es la siguiente:

$$E(Y) = 10 + 2X_1 + 5(3) = 25 + 2X_1, \qquad X_2 = 3.$$

Notemos que esta función de respuesta es una línea recta con pendiente $\beta_1 = 2$. Lo mismo es cierto para **cualquier otro valor de** X_2; sólo el intercept de la función de respuesta será diferente. Por lo tanto, $\beta_1 = 2$ indica que la respuesta media $E(Y)$ aumenta en 2 unidades, cuando se produce un incremento unitario en X_1, cuando X_2 se mantiene constante, sin importar el nivel de X_2.
Del mismo modo, $\beta_1 = 5$, en la función de regresión (43) indica que la respuesta media $E(Y)$ se incrementa en 5 unidades, cuando se produce un incremento unitario en X_2, siempre que X_1 se mantenga constante.
Cuando el efecto de X_1 en la respuesta media no depende del nivel de X_2, y además el efecto de X_2 no depende del nivel de X_1, se dice que las dos variables predictoras tienen *efectos aditivos* o *no interactuan.* Por lo tanto, el modelo de regresión tal como está propuesto en (41) está diseñado para las variables predictoras cuyos efectos sobre la respuesta media son aditivos.
Los parámetros β_1 y β_2 a veces se llaman *coeficientes de regresión parcial* porque reflejan el efecto parcial de una variable de predicción cuando la otra variable predictora es incluida en el modelo y se mantiene constante.

Observación 4.1 *El modelo de regresión para el que la superficie de respuesta es un plano puede ser utilizado tanto porque se crea que modela la verdadera relación entre las variables, o como una aproximación a una superficie de respuesta más compleja. Muchas superficies de respuesta complejas se pueden aproximar razona-*

blemente bien por planos para valores limitados (o acotados) de las covariables X_1 *y* X_2.

4.3. Modelo de Regresión Lineal Múltiple

El modelo de regresión lineal múltiple es un modelo para la variable aleatoria Y cuando se conocen $X_1, X_2, \ldots, X_{p-1}$ las variables regresoras. El modelo es

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1} + \varepsilon_i, \tag{44}$$

donde $\beta_0, \beta_1, \ldots, \beta_{p-1}$ son parámetros (es decir, números) desconocidos, X_{i1}, X_{i2}, $\ldots, X_{ip-1}$ son los valores de las variables predictoras medidas en el i-ésimo individuo (o i-ésima repetición del experimento o i-ésima unidad experimental, según el caso) con $1 \leq i \leq n$, n es el tamaño de muestra, Y_i es la variable respuesta medida en el i-ésimo individuo (observado) y ε_i es el error para el individuo i-ésimo, que no es observable. Haremos supuestos sobre ellos:

$$\varepsilon_i \sim N\left(0, \sigma^2\right), 1 \leq i \leq n, \quad \text{independientes entre sí.} \tag{45}$$

Es decir,

- los ε_i tienen media cero, $E\left(\varepsilon_i\right) = 0$.
- los ε_i tienen todos la misma varianza desconocida que llamaremos σ^2 y que es el otro parámetro del modelo, $Var\left(\varepsilon_i\right) = \sigma^2$.
- los ε_i tienen distribución normal.
- los ε_i son independientes entre sí, e independientes de las covariables X_{i1}, X_{i2}, $\ldots, X_{ip-1}$.

Si definimos $X_{i0} = 1$ para todo i, podemos escribir a (44) de la siguiente forma equivalente

$$\begin{aligned} Y_i &= \beta_0 X_{i0} + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1} + \varepsilon_i \\ &= \sum_{j=0}^{p-1} \beta_j X_{ij} + \varepsilon_i \end{aligned}$$

Observemos que del hecho de que los ε_i son independientes y tienen distribución $N\left(0, \sigma^2\right)$ y de (44) se deduce que, condicional a $X_1, \ldots, X_{p-1}$, $Y_i \sim$

$N\left(\sum_{j=0}^{p-1}\beta_j X_{ij}, \sigma^2\right)$ independientes entre sí. Tomando esperanza (condicional) en (44) obtenemos

$$E\left(Y \mid X_1, ..., X_{p-1}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1},$$

que es una manera alternativa de escribir el modelo (44). Las variables predictoras pueden ser acomodadas para contemplar una serie de situaciones cuyo tratamiento iremos desarrollando a lo largo del curso. Esencialmente pueden ser

- variables continuas, y todas distintas. En la Sección 4.7 veremos un ejemplo de dos continuas.
- variables categóricas o cualitativas, en la Sección 4.13 veremos varios ejemplos donde aparecerán categóricas de dos categorías, que se suelen denominar binarias o dicotómicas o dummies, o de más de dos categorías.
- variables continuas, algunas representando potencias de otras. A esta situación se le suele llamar regresión polinomial.
- variables continuas, pero aparecen en el modelo transformaciones de las originales.
- variables modelando efectos de interacción entre dos o más variables, continuas o categóricas (ver Secciones 4.16 y 4.18).
- combinaciones de algunos o de todos los casos anteriores.

Observación 4.2 *Como ya dijimos en el caso $p = 3$, el término **lineal** en modelo lineal se refiere al hecho de que el modelo (44) es lineal tanto en los parámetros $\beta_0, \ldots, \beta_{p-1}$ como en las covariables $X_1, \ldots, X_{p-1}$ que no tienen porqué ser las variables originalmente observadas para cada individuo o para cada repetición del experimento, pudiendo ser una transformación o recodificación o combinación de ellas. En este sentido, el modelo*

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

*se estudia y ajusta como un modelo de regresión lineal en dos variables: X_i y X_i^2, (aunque matemáticamente se trate de una función cuadrática en una sola variable). Un ejemplo de modelo **no lineal** es el siguiente*

$$Y_i = \beta_0 \exp\left(\beta_1 X_i\right) + \varepsilon_i$$

puesto que no puede expresarse de la forma (44). Varios libros tratan el tema de regresión no lineal, por ejemplo Kutner et al. [2005], parte III.

4.4. Modelo de Regresión Lineal en notación matricial

Ahora presentaremos el modelo (44) en notación matricial. Es una notable propiedad del álgebra de matrices el hecho de que tanto la presentación del modelo como los resultados del ajuste del modelo de regresión lineal múltiple (44) escrito en forma matricial tienen el mismo aspecto (la misma forma) que los que ya vimos para regresión lineal simple. Sólo cambian algunos grados de libertad y algunas constantes.
Enfatizamos en la notación matricial puesto que éste es el tratamiento estándar del tema, y además porque refleja los conceptos esenciales en el ajuste del modelo. Nosotros no calcularemos nada, las cuentas las hace la computadora.
Para expresar el modelo (44) de forma matricial definimos las siguientes matrices

$$\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \underset{n\times p}{\mathbf{X}} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} \tag{46}$$

$$\underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \underset{n\times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Observemos que los vectores $\mathbf{Y}$ y $\boldsymbol{\varepsilon}$ son los mismos que para la regresión lineal simple. El vector $\boldsymbol{\beta}$ contiene los parámetros de regresión adicionales. Cada fila de la matriz $\mathbf{X}$ corresponde a las observaciones correspondientes a cada individuo (la fila i-ésima contiene las observaciones del individuo i-ésimo) y las columnas identifican a las variables.
El modelo (44) se escribe matricialmente en la siguiente forma

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\,\underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

donde
$\mathbf{Y}$ es un vector de respuestas
$\boldsymbol{\beta}$ es un vector de parámetros
$\mathbf{X}$ es una matriz de covariables
$\boldsymbol{\varepsilon}$ es un vector de variables aleatorias normales independientes con esperanza $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ y matriz de varianzas y covarianzas

$$Var(\boldsymbol{\varepsilon}) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}.$$

Entonces tomando a las variables equis como fijas, o, lo que es lo mismo, condicional a las variables equis, la esperanza de $\mathbf{Y}$ resulta ser

$$E(\mathbf{Y} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$$

y la matriz de covarianza de las $\mathbf{Y}$ resulta ser la misma que la de $\boldsymbol{\varepsilon}$

$$Var(\mathbf{Y} \mid \mathbf{X}) = \sigma^2\mathbf{I}.$$

Al igual que hicimos con el modelo de regresión simple, muchas veces omitiremos la condicionalidad a las equis en la notación, es decir, como es bastante habitual en la literatura, escribiremos $E(\mathbf{Y})$ en vez de $E(\mathbf{Y} \mid \mathbf{X})$.

4.5. Estimación de los Parámetros (Ajuste del modelo)

Usamos el método de mínimos cuadrados para ajustar el modelo. O sea, definimos la siguiente función

$$g(b_0, b_1, \ldots, b_{p-1}) = \sum_{i=1}^{n} (Y_i - b_0X_{i0} - b_1X_{i1} - b_2X_{i2} - \cdots - b_{p-1}X_{ip-1})^2 \quad (47)$$

y los estimadores $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_{p-1}$ serán aquellos valores de $b_0, b_1, \ldots, b_{p-1}$ que minimicen a g. Los denominamos estimadores de mínimos cuadrados. Denotaremos al vector de coeficientes estimados por $\widehat{\boldsymbol{\beta}}$.

$$\underset{p\times 1}{\widehat{\boldsymbol{\beta}}} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_{p-1} \end{bmatrix}$$

Las ecuaciones de mínimos cuadrados normales para el modelo de regresión lineal general son

$$\mathbf{X}^t\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^t\mathbf{Y}$$

donde $\mathbf{X}^t$ quiere decir la matriz traspuesta. Algunos autores lo notan $\mathbf{X}'$ (recordemos que la matriz traspuesta es aquella matriz $p \times n$ que tiene por filas a las columnas de $\mathbf{X}$). Los estimadores de mínimos cuadrados son

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y}$$

Observación 4.3 *Para encontrar los estimadores de $\boldsymbol{\beta}$ no se necesita que los errores sean normales.*

Observación 4.4 *En el caso de la regresión lineal, los estimadores de mínimos cuadrados de los betas coinciden también con los estimadores de máxima verosimilitud para el modelo antes descripto, es decir, cuando se asume normalidad de los errores.*

4.6. Valores Ajustados y Residuos

Denotemos al vector de valores ajustados (*fitted values*, en inglés) $\widehat{Y}_i$ por $\widehat{\mathbf{Y}}$ y al vector de residuos $e_i = Y_i - \widehat{Y}_i$ lo denotamos por $\mathbf{e}$

$$\underset{n\times 1}{\widehat{\mathbf{Y}}} = \begin{bmatrix} \widehat{Y}_1 \\ \widehat{Y}_2 \\ \vdots \\ \widehat{Y}_n \end{bmatrix} \qquad \underset{n\times 1}{\mathbf{e}} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Los valores ajustados se calculan del siguiente modo

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y}$$

que son los valores que están en la *superficie de respuesta ajustada* (o sea, en el plano ajustado en el caso $p = 3$). Los residuos se escriben matricialmente como

$$\begin{aligned} \mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} &= \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \\ &= \mathbf{Y} - \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y} \\ &= \left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\right)\mathbf{Y} \end{aligned}$$

Llamando

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t \in \mathbb{R}^{n\times n} \tag{48}$$

a la "*hat matrix*" (la matriz que "sombrerea") tenemos que

$$\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$$

y

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\,\mathbf{Y}.$$

La matriz de varianzas de los residuos es

$$Var\,(\mathbf{e}) = \sigma^2\,(\mathbf{I} - \mathbf{H})\,. \tag{49}$$

Observación 4.5 (residuos) *El modelo de regresión lineal impone que los errores ε_i sean independientes, normales y tengan todos la misma varianza. Como ya hemos dicho, los errores no son observables. Los residuos e_i, que son el correlato empírico de los errores, son observables. Sin embargo, los residuos no son independientes entre sí y sus varianzas no son iguales. Veámoslo.*
Por (49), la varianza de e_i es el elemento que ocupa el lugar ii de la matriz $\sigma^2(\mathbf{I}-\mathbf{H})$. Si la matriz $\mathbf{H}$ fuera igual a cero (que no tendría sentido para el modelo de regresión lineal), todos los residuos tendrían la misma varianza σ^2 (igual que la varianza de los errores). Sin embargo esto no sucede. Calculemos el elemento que ocupa el lugar ii de la matriz $\sigma^2(\mathbf{I}-\mathbf{H})$.

$$Var(e_i) = \sigma^2(1-h_{ii})$$

donde h_{ii} representa el elemento que ocupa el lugar ii de la matriz $\mathbf{H}$. Pero sabemos que

$$\begin{aligned} h_{ij} &= \left(\mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\right)_{ij} = [\textit{fila } i \textit{ de } X]\left(\mathbf{X}^t\mathbf{X}\right)^{-1}[\textit{fila } j \textit{ de } X]^t \\ &= \mathbf{x}_i^t\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{x}_j \end{aligned}$$

donde $\mathbf{x}_i^t$ representa la $i-$ésima fila de $\mathbf{X}$. Luego,

$$Var(e_i) = \sigma^2(1-h_{ii}) = \sigma^2\left(1-\mathbf{x}_i\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{x}_i^t\right)$$

Como en el caso de regresión lineal simple, al elemento ii de la matriz $\mathbf{H}$, es decir, a h_{ii}, se lo denominará el ***leverage o palanca de la observación i-ésima****. Esta cantidad servirá para detectar observaciones atípicas o potencialmente influyentes. Nos ocuparemos de esto en la Sección 5.2.*
En cuanto a la independencia, los residuos no son independientes entre sí ya que la $cov(e_i, e_j)$ ocupa el lugar $ij-$ésimo de la matriz $\sigma^2(\mathbf{I}-\mathbf{H})$. Nuevamente, si la matriz $\mathbf{H}$ fuera igual a cero (que no tendría sentido), entonces dichas covarianzas valdrían cero. Pero

$$cov(e_i, e_j) = \sigma^2(-h_{ij}) = \sigma^2\left(-\mathbf{x}_i\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{x}_j^t\right)$$

donde h_{ij} representa el elemento que ocupa el lugar ij de la matriz $\mathbf{H}$.

Observación 4.6 (teórica) $\mathbf{H}$, *y por lo tanto* $\mathbf{I}-\mathbf{H}$, *son matrices de proyección (es decir que $\mathbf{H}^2 = \mathbf{H}$ y lo mismo ocurre con $\mathbf{I}-\mathbf{H}$).* $\mathbf{H}$ *proyecta al subespacio de $\mathbb{R}^n$ generado por las columnas de $\mathbf{X}$. Algunos textos la notan con la letra $\mathbf{P}$.*

4.7. Dos predictoras continuas

Antes de seguir con las sumas de cuadrados, las estimaciones de los intervalos de confianza para los coeficientes y el test F, veamos un ejemplo numérico con $p = 3$. Consideremos los datos correspondientes a mediciones de 100 niños nacidos con bajo peso en Boston, Massachusetts presentados en el artículo de Leviton et al. [1991], tratados en el libro de Pagano et al. [2000]. Al estudiar el modelo de regresión lineal simple encontramos una relación lineal significativa entre el perímetro cefálico y la edad gestacional para la población de niños nacidos con bajo peso. La recta ajustada a esos datos era

$$\widehat{Y} = 3{,}9143 + 0{,}7801 X_1$$

Nos preguntamos ahora si el perímetro cefálico también dependerá del peso del niño al nacer. Veamos un scatter plot (gráfico de dispersión) del perímetro cefálico versus el peso al nacer, para los 100 niños. El scatter plot de la Figura 37 sugiere que el perímetro cefálico aumenta al aumentar el peso. Pero una vez que hayamos ajustado por la edad gestacional, ¿será que el conocimiento del peso al nacer mejorará nuestra habilidad para predecir el perímetro cefálico de un bebé?
Para responder a esta pregunta ajustamos un modelo de regresión lineal múltiple con dos variables predictoras. Sean

$$\begin{aligned} Y_i &= \text{ perímetro cefálico del i-ésimo niño, en centímetros (}\texttt{headcirc}\text{)} \\ X_{i1} &= \text{ edad gestacional del i-ésimo niño, en semanas (}\texttt{gestage}\text{)} \\ X_{i2} &= \text{ peso al nacer del i-ésimo niño, en gramos (}\texttt{birthwt}\text{)} \end{aligned}$$

Proponemos el modelo (42), o lo que es lo mismo, el modelo (44) con $p = 3$, o sea dos covariables. Lo reescribimos

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i.$$

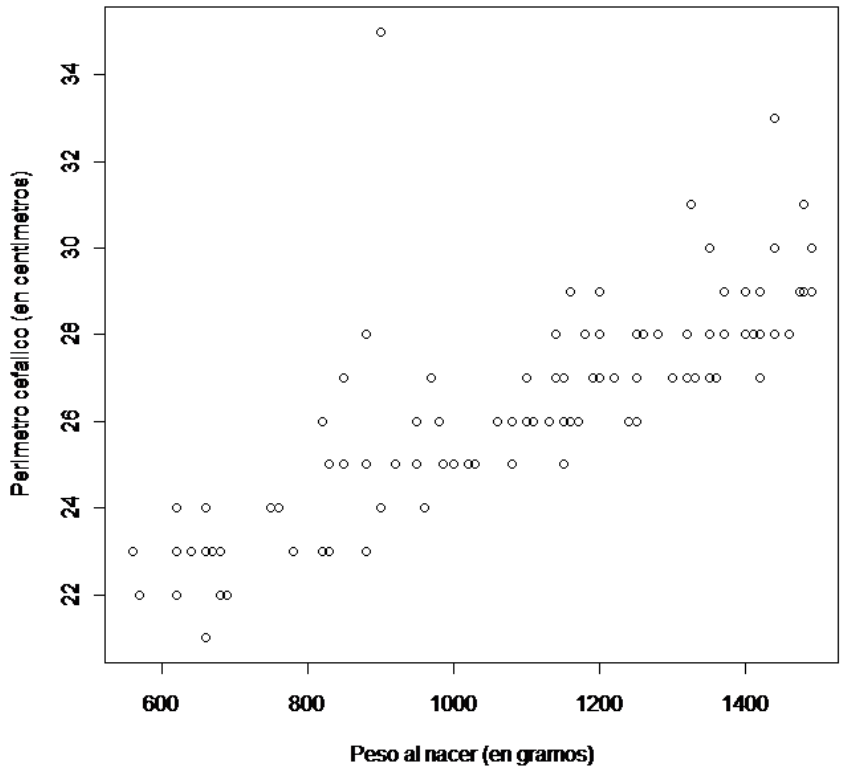
Para darnos una idea de las herramientas con las que trabaja la computadora que ajustará el modelo, listamos los primeros siete datos en la Tabla 18.

El modelo ajustado y las instrucciones para hacerlo en `R`, figuran en la Tabla 19. La superficie ajustada resulta ser

$$\widehat{Y} = 8{,}3080 + 0{,}4487 X_1 + 0{,}0047 X_2.$$

La ordenada al origen, que es 8,3080 es, en teoría, el valor medio del perímetro cefálico para bebés de bajo peso con edad gestacional de 0 semanas y peso al nacer de 0 gramos, y por lo tanto carece de sentido. El coeficiente estimado de edad gestacional (0,4487) no es el mismo que cuando la edad gestacional era la

Figura 37: Perímetro cefálico versus peso al nacer para la muestra de 100 bebés de bajo peso.



única variable explicativa en el modelo; su valor descendió de 0,7801 a 0,4487. Esto implica que, si mantenemos el peso al nacer de un niño constante, cada incremento de una semana en la edad gestacional corresponde a un aumento de 0,4487 centímetros en su perímetro cefálico, en promedio. Una manera equivalente de decirlo es que dados dos bebés con el mismo peso al nacer pero tales que la edad gestacional del segundo de ellos es una semana más grande que la del primero, el perímetro cefálico esperado para el segundo bebé será 0,4487 centímetros mayor que el primero.

De forma similar, el coeficiente del peso al nacer indica que si la edad gestacional de un bebé no cambia, cada incremento de un gramo en el peso al nacer redunda en un aumento de 0,0047 centímetros en el perímetro cefálico, en promedio. En este

Tabla 18: Primeros siete datos de bebés de bajo peso

Niño i	$Y_i =$ `headcirc`	$X_{i1} =$ `gestage`	$X_{i2} =$ `birthwt`
1	27	29	1360
2	29	31	1490
3	30	33	1490
4	28	31	1180
5	29	30	1200
6	23	25	680
7	22	27	620

caso en el que el valor del coeficiente estimado es tan pequeño, puede tener más sentido expresar el resultado aumentando las unidades involucradas, por ejemplo decir: si la edad gestacional no cambia, cada incremento de 10 g. en el peso al nacer redunda en un aumento de 0,047 cm. en el perímetro cefálico, en promedio.

4.8. Resultados de Análisis de la Varianza (y estimación de σ^2)

4.8.1. Sumas de cuadrados y cuadrados medios (SS y MS)

Las sumas de cuadrados para el análisis de la varianza son,

$$\begin{aligned} \text{SSTo} &= \text{suma de cuadrados total} \\ &= \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 \\ &= \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 \end{aligned}$$

que en términos matriciales puede escribirse como

$$\text{SSTo} = \mathbf{Y}^t\mathbf{Y} - \frac{1}{n}\mathbf{Y}^t\mathbf{J}\mathbf{Y} = \mathbf{Y}^t\left[\mathbf{I} - \frac{1}{n}\mathbf{J}\right]\mathbf{Y},$$

Tabla 19: Ajuste del modelo lineal para los datos de bebés de bajo peso, `headcirc` con dos explicativas continuas: `gestage` y `birthwt`

```
> ajuste2<-lm(headcirc~gestage+birthwt)
>
> summary(ajuste2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.3080154  1.5789429   5.262 8.54e-07
gestage     0.4487328  0.0672460   6.673 1.56e-09
birthwt     0.0047123  0.0006312   7.466 3.60e-11
---

Residual standard error: 1.274 on 97 degrees of freedom
Multiple R-squared: 0.752,        Adjusted R-squared: 0.7469
F-statistic: 147.1 on 2 and 97 DF,  p-value: < 2.2e-16
```

donde $\mathbf{J}$ es una matriz $n \times n$ toda de unos. De igual modo,

$$\begin{aligned}\text{SSRes} &= \text{suma de cuadrados de los residuos (SSE)} \\ &= \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 = \sum_{i=1}^{n} e_i^2 \\ &= \mathbf{e}^t\mathbf{e} = \left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^t \left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \\ &= \mathbf{Y}^t\mathbf{Y} - \widehat{\boldsymbol{\beta}}^t\mathbf{X}^t\mathbf{Y} = \mathbf{Y}^t\left[\mathbf{I} - \mathbf{H}\right]\mathbf{Y}\end{aligned}$$

que en términos matriciales se escribe

$$\text{SSRes} = \mathbf{Y}^t\mathbf{Y} - \widehat{\boldsymbol{\beta}}^t\mathbf{X}^t\mathbf{Y} = \mathbf{Y}^t\left[\mathbf{I} - \mathbf{H}\right]\mathbf{Y}$$

y

$$\begin{aligned}\text{SSReg} &= \text{suma de cuadrados de la regresión o del modelo (SSM)} \\ &= \sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2,\end{aligned}$$

y vale

$$\text{SSReg} = \widehat{\boldsymbol{\beta}}^t\mathbf{X}^t\mathbf{Y} - \frac{1}{n}\mathbf{Y}^t\mathbf{J}\mathbf{Y} = \mathbf{Y}^t\left[\mathbf{H} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{J}\right]\mathbf{Y}.$$

Más allá de las expresiones matemáticas que permiten calcularlas, en la regresión múltiple se cumple la misma propiedad que en la regresión simple en cuanto a las sumas de cuadrados. Volvamos sobre ellas. Recordemos que como los estimadores $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_{p-1}$ se eligen como aquellos valores de $b_0, b_1, \ldots, b_{p-1}$ que minimicen a g dada en (47), luego los parámetros elegidos hacen que la suma de los cuadrados de los residuos (SSRes) sea lo más chica posible. Pero, aunque esta superficie sea la mejor superficie disponible, todavía cabe preguntarse cuan bueno es el ajuste encontrado, es decir, cuan bien ajusta el modelo a los datos observados. Para ello, una manera es comparar el ajuste que proporciona el modelo de regresión lineal con algo, y el algo que siempre podemos elegir es el modelo más básico que podemos encontrar. Entonces usamos las sumas de cuadrados para calcular el ajuste del modelo más básico (un solo parámetro que ajuste a todas las observaciones). Es decir, elegimos el valor de μ tal que minimice

$$\sum_{i=1}^{n} (Y_i - \mu)^2 ,$$

sin tener en cuenta para nada los valores de las covariables $(X_1, \ldots, X_{p-1})$. Es un resultado de un curso inicial de estadística que el valor de μ que minimiza dicha suma es el promedio de las Ys es decir, $\mu = \overline{Y}$. Esencialmente, estamos tomando como medida de cuan bien ajusta un modelo, a la suma de los cuadrados; en general

$$\Delta_{\text{modelo}} = \sum (\text{observados} - \text{modelo})^2 \tag{50}$$

donde el modelo es la superficie de respuesta (44) en regresión lineal múltiple y un sólo parámetro en el modelo más básico. Para cada modelo usamos la ecuación (50) para ajustar ambos modelos, es decir, encontramos los valores de los parámetros que minimizan (50) entre todos los valores posibles y, luego, básicamente si el modelo lineal es razonablemente bueno ajustará a los datos significativamente mejor que el modelo básico. Es decir, la resta

$$\Delta_{\text{modelo básico}} - \Delta_{\text{regresión lineal}} = \text{SSTo} - \text{SSRes}$$

será pequeña comparada con lo que era la SSTo. Esto es un poco abstracto así que mejor lo miramos en un ejemplo.

Imaginemos que nos interesa predecir el perímetro cefálico de un niño al nacer (Y) a partir de la edad gestacional del bebé (X_1) y de su peso al nacer (X_2). ¿Cuánto será el perímetro cefálico de un bebé con 33 semanas de edad gestacional y que pesa 1490 gramos al nacer? Si no tuviéramos un modelo preciso de la relación entre las tres variables en niños nacidos con bajo peso, ¿cuál podría ser nuestro mejor pronóstico? Bueno, posiblemente la mejor respuesta sea dar el número promedio de perímetros cefálicos en nuestra base de datos, que resulta ser 26,45 cm. Observemos

que la respuesta sería la misma si ahora la pregunta fuera: ¿cuánto será el perímetro cefálico de un niño con 25 semanas de gestación y que pesó 680 g. al nacer? Nuevamente, en ausencia de un vínculo preciso, nuestro mejor pronóstico sería dar el promedio observado de perímetros cefálicos, o sea 26,45 cm. Claramente hay un problema: no importa cual es la edad gestacional o el peso al nacer del niño, siempre predecimos el mismo valor de perímetro cefálico. Debería ser claro que la media es poco útil como modelo de la relación entre dos variables, pero es el modelo más básico del que se dispone.
Repasemos entonces los pasos a seguir. Para ajustar el modelo más básico, predecimos el outcome Y por $\overline{Y}$, luego calculamos las diferencias entre los valores observados y los valores que da el modelo ($\overline{Y}$ siempre para el modelo básico) y la ecuación (50) se convierte en la SSTo (es decir, SSTo es la cantidad total de diferencias presentes cuando aplicamos el modelo básico a los datos). La SSTo representa una medida del desajuste que surge de usar el promedio como único resumen de los datos observados. En un segundo paso ajustamos el modelo más sofisticado a los datos (el modelo de regresión lineal múltiple con dos predictores). Este modelo permite pronosticar un valor distinto para cada combinación de covariables. A este valor lo hemos llamado valor predicho y resulta ser

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{i1} + \widehat{\beta}_2 X_{i2}.$$

En el ejemplo, para la primer pregunta nuestra respuesta sería

$$\widehat{\beta}_0 + \widehat{\beta}_1 33 + \widehat{\beta}_2 1490 = 8{,}3080 + 0{,}4487 \cdot 33 + 0{,}0047 \cdot 1490 = 30{,}118$$

y para la segunda pregunta tendríamos

$$\widehat{\beta}_0 + \widehat{\beta}_1 25 + \widehat{\beta}_2 680 = 8{,}3080 + 0{,}4487 \cdot 25 + 0{,}0047 \cdot 680 = 22{,}722.$$

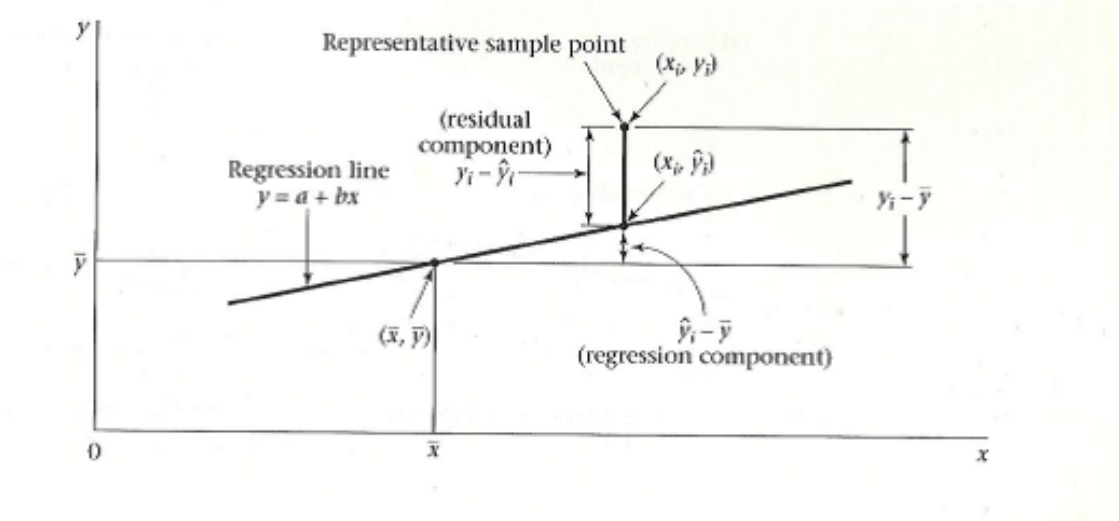
Hemos visto que el modelo de regresión lineal múltiple encuentra los valores de $\widehat{\beta}_0, \widehat{\beta}_1$ y $\widehat{\beta}_2$ por el método de mínimos cuadrados, es decir minimizando las diferencias entre el modelo ajustado a los datos y los propios datos. Sin embargo, aun en este modelo optimizado hay todavía imprecisiones que se representan por las diferencias entre cada valor observado (Y_i) y cada valor predicho por la regresión $\left(\widehat{Y}_i\right)$. Como antes, calculamos esas diferencias, elevamos al cuadrado cada una de ellas y las sumamos (si las sumáramos sin elevarlas al cuadrado la suma terminaría dando cero). El resultado se conoce como la suma de los cuadrados de los residuos (SSRes). Este valor representa el grado de imprecisión del modelo lineal con estas dos covariables ajustado a los datos. Podemos usar estos dos valores para calcular cuanto mejor es usar la superficie de respuesta estimada en vez de la media como modelo (es decir, ¿cuánto mejor es el mejor modelo posible comparado con el peor?) La mejora en predicción resultante al usar el mejor modelo en vez de la

media se calcula al hacer la resta entre SSTo y SSRes. Esta diferencia nos muestra la reducción en la imprecisión que se obtiene por usar un modelo de regresión lineal. Como en el caso de regresión lineal simple, puede verse que esta resta da SSReg, es decir

$$\text{SSTo} - \text{SSRes} = \text{SSReg}.$$

La Figura 38 muestra ambas distancias para una misma observación, en el caso de regresión lineal simple.

Figura 38: Distancias que intervienen en las sumas de cuadrados para una observación. Fuente: Rosner [2006], pág. 473.



Si el valor de SSReg es grande, entonces usar el modelo de regresión lineal es muy distinto a usar la media para predecir el outcome. Esto implica que el modelo de regresión ha hecho una gran mejora en la calidad de la predicción de la variable respuesta. Por otro lado, si SSReg es chico, entonces el hecho de usar el modelo de regresión es sólo un poco mejor que usar la media. (Observemos de paso que SSTo siempre será mayor que SSRes ya que tomando $b_1 = b_2 = \cdots = b_{p-1} = 0$ y $b_0 = \overline{Y}$ que son valores posibles para los parámetros de la regresión lineal múltiple recuperamos al modelo básico, es decir, el modelo básico está contenido entre todos los modelos posibles bajo la regresión lineal múltiple). Pero ahora, por supuesto, aparece la natural pregunta de cuándo decimos que un valor de SSReg es "grande" o "pequeño".

4.8.2. Coeficiente de Determinación Múltiple (R^2 y R^2 ajustado)

Una primera manera de zanjar esto es calcular la proporción de mejora debida al modelo. Esto es fácil de hacer dividiendo la suma de cuadrados de la regresión por la suma de cuadrados total. Es lo que hacíamos también en regresión lineal simple. El resultado se denomina R^2, el **coeficiente de determinación múltiple**. Para expresar este valor como un porcentaje hay que multiplicarlo por 100. Luego, como en el caso de regresión lineal simple, R^2 representa la proporción de variabilidad de la variable respuesta que queda explicada por el modelo de regresión relativa a cuánta variabilidad había para ser explicada antes de aplicar el modelo. Luego, como porcentaje, representa el porcentaje de variación de la variable respuesta que puede ser explicada por el modelo

$$R^2 = \frac{\text{SSReg}}{\text{SSTo}} = 1 - \frac{\text{SSRes}}{\text{SSTo}}.$$

De igual modo que para el modelo de regresión lineal simple, R (la raíz cuadrada de R^2) resulta ser la correlación de Pearson entre los valores observados de (Y_i) y los valores predichos $\left(\widehat{Y}_i\right)$ sin tener en cuenta el signo. Por lo tanto los valores grandes de R múltiple (al que se lo suele llamar *coeficiente de correlación múltiple*) representan una alta correlación entre los valores observados y predichos del outcome. Un R múltiple igual a uno representa una situación en la que el modelo predice perfectamente a los valores observados.

Observación 4.7 *El hecho de agregar variables explicativas X al modelo de regresión sólo puede aumentar el R^2 y nunca reducirlo, puesto que la suma de cuadrados de los residuos SSReg nunca puede aumentar con más covariables X y la suma de cuadrados total SSTo siempre vale lo mismo para un conjunto fijo de respuestas Y_i. Por este hecho, de que la inclusión de más covariables siempre aumenta el R^2, sean estas importantes o no, se sugiere que cuando se quieran comparar modelos de regresión con distinto número de covariables en vez de usarse el R^2 se utilice una medida modificada que ajusta por el número de covariables explicativas incluidas en el modelo. El* ***coeficiente de determinación múltiple ajustado****, que se suele denominar R_a^2 , ajusta a R^2 dividiendo cada suma de cuadrados por sus correspondientes grados de libertad, de la siguiente forma*

$$R_a^2 = 1 - \frac{\frac{SSRes}{n-p}}{\frac{SSTo}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSRes}{SSTo}$$

Este coeficiente de determinación múltiple puede, de hecho, disminuir cuando se agrega una covariable al modelo, ya que cualquier disminución de la SSRes puede ser más que compensada por la pérdida de un grado de libertad en el denominador

$n-p$. Si al comparar un modelo con las covariables $X_1, \ldots, X_k$ para explicar a Y con un modelo que tiene las mismas $X_1, \ldots, X_k$ y además a X_{k+1} como covariables vemos un aumento de los R_a^2, esto es una indicación de que la covariable X_{k+1} es importante para predecir a Y, aún cuando las covariables $X_1, \ldots, X_k$ ya están incluidas en el modelo. Si en cambio, el R_a^2 no aumenta o incluso disminuye al incorporar a X_{k+1} al modelo, esto es señal de que una vez que las variables $X_1, \ldots, X_k$ se utilizan para predecir a Y, la variable X_{k+1} no contribuye a explicarla y de debe incluirse en el modelo.

Observación 4.8 *Hemos dicho que en el modelo lineal múltiple, el R^2 representa el cuadrado del coeficiente de correlación muestral de Pearson entre los valores Y_i observados y los valores $\widehat{Y}_i$ predichos. Esto también sucede en regresión lineal simple. Es decir,*

$$r = \frac{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)\left(\widehat{Y}_i - \overline{\widehat{Y}}\right)}{\sqrt{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 \sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{\widehat{Y}}\right)^2}}$$

es tal que el valor absoluto de r es la raíz de R^2, $|r| = \sqrt{R^2}$. En este caso, el signo de r es positivo ya que los valores observados y los predichos están positivamente correlacionados. Entonces, ¿cómo juega la raíz cuadrada? Como R^2 es un número comprendido entre 0 y 1, la raíz cuadrada es en dicho intervalo una función creciente que es la inversa de la función elevar al cuadrado. Por lo tanto, como puede verse en la Figura 39, $r = \sqrt{R^2}$ será ***mayor*** *que R^2.*

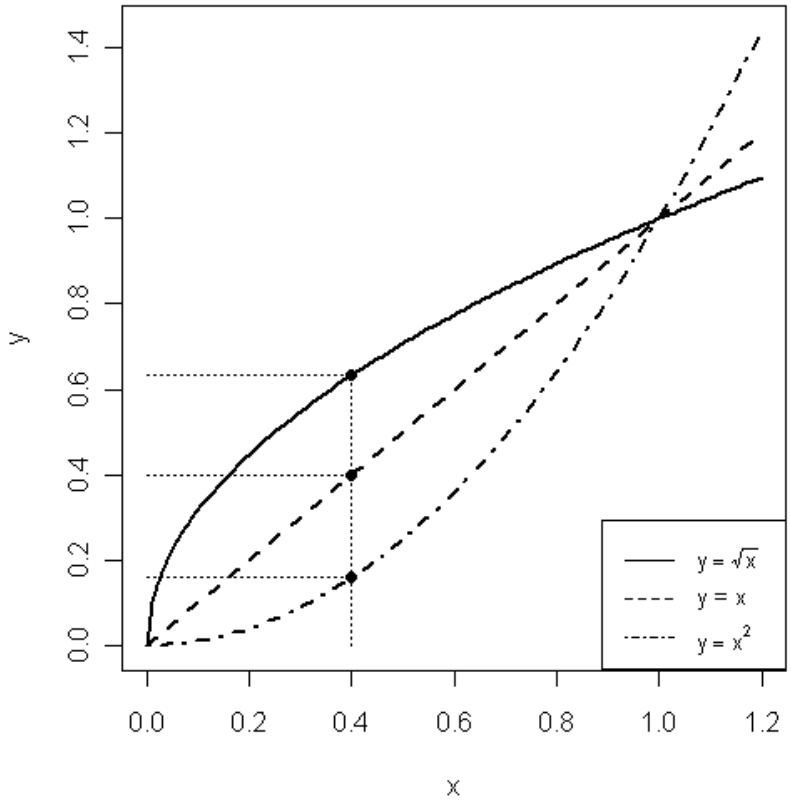
Para ver cómo funciona este vínculo entre r y R^2 inspeccionamos un par de ejemplos numéricos, que exhibimos en la Tabla 20.

Tabla 20: Algunos valores del coeficiente de determinación múltiple R^2 con el respectivo valor del coeficiente de correlación muestral de Pearson, r entre valores predichos y valores observados.

R^2	r
0,1	0,316
0,4	0,632
0,6	0,775
0,7	0,837
0,9	0,949
0,99	0,995

Desde esta óptica, otra interpretación del R^2 es pensar que un buen modelo debería producir valores predichos altamente correlacionados con los valores observados.

Figura 39: Función raíz cuadrada comparada con la función elevar al cuadrado y la identidad en el intervalo $(0, 1)$. Están graficadas las imágenes del $x = 0{,}4$, con tres puntos cuyas alturas son (en orden ascendente) $0{,}4^2 = 0{,}16; 0{,}4$ y $\sqrt{0{,}4} = 0{,}632$.



Esta es otra manera de visualizar por qué un R^2 alto es, en general, una buena señal de ajuste.

4.8.3. Test F

Como en el modelo de regresión lineal simple, una segunda forma de usar las sumas de cuadrados para evaluar la bondad de ajuste del modelo de regresión lineal múltiple a los datos es a través de un test F. Este test se basa en el cociente de la mejora debida al modelo (SSReg) y la diferencia entre el modelo y los datos observados (SSRes). La Tabla 21 resume la información que involucra a la construcción del test F. De hecho, en vez de utilizar las sumas de cuadrados por sí mismas,

tomamos lo que se denominan los cuadrados medios (MS *mean squares* o sumas medias de cuadrados o cuadrados medios). Para trabajar con ellos, es necesario primero dividir a las sumas de cuadrados por sus respectivos grados de libertad. Para la SSReg, los grados de libertad son simplemente el número de covariables en el modelo, es decir, $p-1$. Del mismo modo que sucedía con la regresión lineal simple, las diferencias $\left(\widehat{Y}_i - \overline{Y}\right)$ quedan determinadas al fijar los $p-1$ coeficientes que acompañan a las $p-1$ covariables, luego las diferencias $\left(\widehat{Y}_i - \overline{Y}\right)$ tienen $p-1$ grados de libertad.

Tabla 21: Tabla de ANOVA para el modelo de Regresión Lineal General (44)

Fuente de variación	SS	g.l.	MS
Regresión	$\text{SSReg} = \sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2$	$p-1$	$\text{MSReg} = \frac{\text{SSReg}}{p-1}$
Residuos	$\text{SSRes} = \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$	$n-p$	$\text{MSRes} = \frac{\text{SSRes}}{n-p}$
Total	$\text{SSTo} = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$	$n-1$	

Para la SSRes los grados de libertad son el número de observaciones menos el número de parámetros que se estiman (es decir, el número de coeficientes beta incluyendo el β_0), en este caso $n-p$. Esto proviene, al igual que en el caso de regresión lineal simple, del hecho de que los residuos satisfacen p ecuaciones normales. Las p ecuaciones que se obtienen al igualar a cero las p derivadas. Luego, si conocemos $n-p$ de ellos, podemos hallar los restantes p a partir de despejarlos de las p ecuaciones lineales.
Los resultados son, respectivamente, el cuadrado medio de regresión (que notaremos MSReg o MSM, es decir *regression mean square* o *model mean square*) y el cuadrado medio de residuos (MSRes o MSE, es decir, *residual mean square* o *mean square error*). El estadístico F es una medida de cuánto mejora el modelo la predicción de la variable respuesta comparada con el nivel de imprecisión de los datos originales. Si el modelo es bueno, esperamos que la mejora en la predicción debida al modelo sea grande (de manera que MSReg sea grande) y que la diferencia entre el modelo y los datos observados sea pequeña (o sea, MSRes pequeña). Por eso, un buen modelo debe tener un estadístico F grande (al menos mayor a 1 porque en tal caso el numerador, de decir, la mitad superior de (51) será mayor que el denominador -la mitad inferior de (51)). El estadístico F es

$$F = \frac{\text{MSReg}}{\text{MSRes}} = \frac{\frac{\text{SSReg}}{p-1}}{\frac{\text{SSRes}}{n-p}} = \frac{\text{SSReg}}{\text{SSRes}}\frac{(n-p)}{(p-1)}. \tag{51}$$

Se construye para testear las hipótesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$
$$H_1 : \text{ no todos los } \beta_k \ (k = 1, 2, \ldots, p-1) \text{ son iguales a } 0$$

Observemos que H_0 dice que no hay vínculo entre la variable respuesta y las regresoras. En cambio, H_1 dice que al menos una de las variables regresoras sirve para predecir a Y. La distribución de F cuando H_0 es cierta es la distribución F (de Snedecor o de Fisher) con $p-1$ grados de libertad en el numerador y $n-p$ grados de libertad en el denominador[4]. Esto es porque bajo el supuesto de normalidad de los errores, se tiene que

$$\text{SSRes} \sim \chi^2_{n-p}$$

y si además H_0 es verdadera, entonces

$$\text{SSReg} \sim \chi^2_{p-1}$$

y además SSRes y SSReg son independientes. El test rechaza H_0 cuando $F > F_{p-1,n-p,1-\alpha}$, el $1-\alpha$ percentil de la distribución válida cuando H_0 es verdadera. Para valores grandes de F (es decir, p-valores pequeños) el test rechaza H_0 y concluye que no todos los coeficientes que acompañan a las covariables del modelo de regresión lineal son nulos.

Observación 4.9 *Cuando $p-1=1$, este test se reduce al test F visto en el modelo de regresión lineal simple para testear si β_1 es 0 o no.*

Observación 4.10 *La existencia de una relación de regresión lineal, por supuesto, no asegura que puedan hacerse predicciones útiles a partir de ella.*

Usualmente, como ya hemos visto en el modelo lineal simple, estos valores aparecen en la salida de cualquier paquete estadístico en lo que se conoce como tabla de ANOVA (*Analysis of Variance table*, que presentamos en la Tabla 21.
Usualmente la tabla se completa con dos últimas columnas que se denominan F y p-valor. La columna F tiene un único casillero completo (el correspondiente a la primer fila) con el valor del estadístico, es decir

$$F_{obs} = \frac{\text{MSReg}}{\text{MSRes}}.$$

La columna p-valor tiene también un único casillero con el p-valor del test, que es la probabilidad, calculada asumiendo que H_0 es verdadera, de observar un valor del estadístico F tan alejado de lo esperado como el observado en la muestra, o más alejado aún, o sea

$$p - valor = P\left(F_{p-1,n-p} > F_{obs}\right).$$

[4]Por definición, si U es una variable aleatoria con distribución χ^2_k y V es otra variable aleatoria independiente de U con distribución χ^2_m, entonces la variable $W = \frac{U/k}{V/m}$ se denomina *F de Fisher con k grados de libertad en el numerador y m grados de libertad en el denominador.*

4.8.4. Estimación de σ^2

El modelo de regresión lineal dado en (44) y (45) impone que los errores $\varepsilon_1, \ldots, \varepsilon_n$ sean variables aleatorias independientes con esperanza cero y $Var(\varepsilon_i) = \sigma^2$. Si tuviéramos los errores, sabemos que un estimador insesgado de σ^2 es

$$\frac{1}{n-1}\sum_{i=1}^{n}(\varepsilon_i - \overline{\varepsilon})^2 .$$

El problema es que en el modelo de regresión lineal múltiple, al igual que en el caso de regresión lineal simple, los errores **no son observables.** Para estimar a σ^2 los podemos reemplazar por sus correlatos empíricos, los residuos $e_1, \ldots, e_n$. Pero, como ya vimos en la Observación 4.5 los residuos **no** son independientes. En el caso del modelo lineal simple habíamos visto que los residuos están ligados entre sí ya que satisfacen dos ecuaciones lineales (las dos ecuaciones normales):

- la suma de los residuos $e_1, \ldots, e_n$ es cero.

- la correlación muestral entre $e_1, \ldots, e_n$ y $X_1, \ldots, X_n$ es cero, o equivalentemente, el coeficiente de correlación de Pearson calculado para $(X_1, e_1), \ldots, (X_n, e_n)$ es cero.

En el caso de regresión lineal múltiple con $p-1$ variables predictoras, los residuos están ligados entre sí de una manera más estrecha, ya que satisfacen p ecuaciones lineales (linealmente independientes): como $\mathbf{e} = (\mathbf{I} - \mathbf{H})\,\mathbf{Y}$ y $\mathbf{H}$ es una matriz de proyección de rango p resulta que $\mathbf{He} = \mathbf{0}$. Una de ellas es, también, que la suma de los residuos vale cero. Informalmente se dice que los residuos tienen $n-p$ grados de libertad. Esto quiere decir que conociendo $n-p$ de ellos, podemos deducir cuánto valen los p restantes despejándolos de las ecuaciones normales. Luego, el estimador de σ^2 se basará en los residuos de la siguiente forma

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{1}{n-p}\sum_{i=1}^{n}(e_i - \overline{e})^2 = \frac{1}{n-p}\sum_{i=1}^{n}(e_i)^2 \\ &= \frac{1}{n-p}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 = \frac{\text{SSRes}}{n-p} \\ &= \text{MSRes.} \end{aligned} \tag{52}$$

Es decir, el cuadrado medio de los residuos es el estimador de σ^2 dado por el modelo de regresión. En la salida de un paquete estadístico se puede encontrar en el casillero correspondiente en la tabla de ANOVA.

4.9. Inferencias sobre los parámetros de la regresión

Los estimadores de mínimos cuadrados $\widehat{\beta}_k$ son insesgados, es decir,

$$E\left(\widehat{\beta}_k\right) = \beta_k.$$

La matriz de covarianza de dichos estimadores $Var\left(\widehat{\boldsymbol{\beta}}\right)$ está dada por una matriz $p \times p$ que en la coordenada jk tiene la covarianza entre $\widehat{\beta}_j$ y $\widehat{\beta}_k$ y que resulta ser

$$Var\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2 \left(X^t X\right)^{-1}.$$

Como vimos en la Sección 4.8.4, MSRes es el estimador de σ^2, por lo que la estimación de dicha matriz está dada por

$$\widehat{Var}\left(\widehat{\boldsymbol{\beta}}\right) = \widehat{\sigma}^2 \left(X^t X\right)^{-1} = \text{MSRes}\left(X^t X\right)^{-1}.$$

4.9.1. Intervalos de confianza para β_k

Para el modelo de errores normales dado por (44) y (45) tenemos que

$$\frac{\widehat{\beta}_k - \beta_k}{\sqrt{\widehat{Var}\left(\widehat{\beta}_k\right)}} \sim t_{n-p} \text{ para } k = 0, 1, \ldots, p-1.$$

Recordemos que $n-p$ es el número de observaciones menos el número de covariables del modelo menos uno. Muchas veces al denominador $\sqrt{\widehat{Var}\left(\widehat{\beta}_k\right)}$ se lo llama $s\left(\widehat{\beta}_k\right)$. Luego, el intervalo de confianza de nivel $1-\alpha$ para cada β_k es

$$\widehat{\beta}_k \pm t_{n-p,1-\frac{\alpha}{2}} \sqrt{\widehat{Var}\left(\widehat{\beta}_k\right)}. \tag{53}$$

4.9.2. Tests para β_k

Los tests para β_k se llevan a cabo de la forma usual. Para testear

$$H_0 : \beta_k = 0$$
$$H_1 : \beta_k \neq 0$$

usamos el estadístico

$$T = \frac{\widehat{\beta}_k}{\sqrt{\widehat{Var}\left(\widehat{\beta}_k\right)}}$$

y rechazamos H_0 cuando $|T| \geq t_{n-p,1-\frac{\alpha}{2}}$. El p-valor, a su vez, se calcula como

$$p - valor = P\left(|t_{n-p}| \geq |T_{obs}|\right).$$

Observemos que cuando realizamos este test asumimos que en el modelo aparecen todas las restantes covariables. Se puede calcular la potencia de este test.

4.9.3. Inferencias conjuntas

El objetivo de los intervalos de confianza y tests presentados en las secciones 4.9.1 y 4.9.2 es proveer conclusiones con un nivel prefijado de confianza sobre cada uno de los parámetros $\beta_0, \beta_1, \ldots, \beta_{p-1}$ por separado. La dificultad es que éstos no proporcionan el 95 por ciento de confianza de que las conclusiones de los p intervalos son correctas. Si las inferencias fueran independientes, la probabilidad de que los p intervalos construidos cada uno a nivel 0,95, contengan al verdadero parámetro sería $(0{,}95)^p$, o sea, solamente 0,857 si p fuese 3. Sin embargo, las inferencias no son independientes, ya que son calculadas a partir de un mismo conjunto de datos de la muestra, lo que hace que la determinación de la probabilidad de que todas las inferencias sean correctas sea mucho más difícil.
En esta sección propondremos intervalos de confianza de **nivel conjunto** 0,95. Esto quiere decir que nos gustaría construir una serie de intervalos (o tests) para los cuales tengamos una garantía sobre la exactitud de todo el conjunto de intervalos de confianza (o tests). Al conjunto de intervalos de confianza (o tests) de interés lo llamaremos familias de intervalos de confianza de nivel conjunto o simultáneo (o regiones de confianza de nivel simultáneo o tests o inferencias conjuntas). En nuestro ejemplo, la familia se compone de p estimaciones, para $\beta_0, \beta_1, \ldots, \beta_{p-1}$. Podríamos estar interesados en construir regiones de confianza para una cantidad g entre 1 y p de estos parámetros, con g prefijado. Distingamos entre un intervalo de confianza de nivel 0,95 para un parámetro, y una familia de intervalos de nivel simultáneo 0,95 para g parámetros. En el primer caso, 0,95 es la proporción de intervalos construido con el método en cuestión que cubren al verdadero parámetro de interés cuando se seleccionan repetidamente muestras de la población de interés y se construyen los intervalos de confianza para cada una de ellas. Por otro lado, cuando construimos una familia de regiones o intervalos de confianza de nivel simultáneo 0,95 para g parámetros: $\theta_1, \ldots, \theta_g$ el valor 0,95 indica la proporción de familias de g intervalos que están enteramente correctas (cubren a los g parámetros de interés, simultáneamente) cuando se seleccionan repetidamente muestras de la población de interés y se construyen los intervalos de confianza específicos para los g parámetros en cuestión, o sea

$$P\left(\{\theta_1 \in I_1\} \cap \{\theta_2 \in I_2\} \cap \cdots \cap \{\theta_g \in I_g\}\right) = 0{,}95,$$

si $I_1, \ldots, I_g$ son los g intervalos construidos usando los mismos datos. Luego, el **nivel simultáneo** de una familia de regiones o intervalos de confianza corresponde a la probabilidad, calculada previa al muestreo, de que la familia entera de afirmaciones sea correcta.
Ilustremos esto en el caso del ejemplo de los 100 bebés de bajo peso. Si nos interesara construir intervalos de confianza de nivel simultáneo 0,95 para β_1 y β_2, una familia de intervalos de confianza simultáneos para estos datos consistiría en dos intervalos de confianza de modo tal que si tomáramos muestras de 100 bebés de bajo peso, les midiéramos la edad gestacional, el perímetro cefálico y el peso al nacer, y luego construyéramos para cada muestra los dos intervalos de confianza para β_1 y β_2, para el 95 % de las muestras ambos intervalos construidos con este método cubrirían tanto al verdadero β_1 como al verdadero β_2. Para el 5 % restante de las muestras, resultaría que uno o ambos intervalos de confianza sería incorrecto.
En general es sumamente deseable contar con un procedimiento que provea una familia de intervalos de confianza de nivel simultáneo cuando se estiman varios parámetros con una misma muestra de datos, ya que le permite al analista entrelazar varios resultados juntos en un conjunto integrado de conclusiones con la seguridad de que todo el conjunto de inferencias es correcto. Para obtenerlos hay básicamente dos herramientas estadísticas disponibles. Una de ellas es el estudio matemático en detalle del fenómeno en cuestión, en este caso, estudiar matemáticamente las propiedades de los estimadores $\widehat{\beta}_0, \ldots, \widehat{\beta}_{p-1}$ de manera de poder obtener la distribución exacta de alguna medida numérica que los resuma, como el $\text{máx}_{0 \leq k \leq p-1} \left|\widehat{\beta}_k\right|$ o las descripciones matemáticas del elipsoide p dimensional más pequeño que los contenga, con probabilidad 0,95, para contar un par de ejemplos que son utilizados en distintas áreas de la estadística para construir regiones de confianza de nivel simultáneo. Veremos otro en la Sección 4.10.2. La otra herramienta consiste en construir intervalos de confianza con nivel simultáneo a partir de ajustar el nivel de confianza de cada intervalo individual a un valor más alto, de modo de poder asegurar el nivel simultáneo de la construcción. Esto es lo que se conoce como el método de Bonferroni para la construcción de intervalos de nivel simultáneo. Una descripción detallada de este método puede consultarse en Kutner et al. [2005], pág. 155 a 157. Este procedimiento es de aplicación bastante general en la estadística. En vez de usar el percentil de la t propuesto en la Sección 4.9.1 para cada intervalo de confianza para β_k se usa el percentil correspondiente a un nivel mayor. Cuando se quieren construir intervalos de confianza de nivel simultáneo $1-\alpha$ para g coeficientes de la regresión, el percentil que se utiliza es el correspondiente a un nivel $1-\frac{\alpha}{2g}$ en cada intervalo en particular. Resultan ser intervalos más anchos que los presentados en la Sección 4.9.1. Una observación importante es que el procedimiento de Bonferroni es conservativo, es decir, el nivel conjunto de los intervalos así construidos resulta ser mayor o igual a $1-\alpha$.

Así, se pueden construir los intervalos de confianza simultáneos de Bonferroni para estimar varios coeficientes de regresión de manera simultánea. Si se desean estimar simultáneamente g parámetros (donde $g \leq p$), los intervalos de confianza con nivel simultáneo $1-\alpha$ son los siguientes

$$\widehat{\beta}_k \pm t_{n-p,1-\frac{\alpha}{2g}} \sqrt{\widehat{Var}\left(\widehat{\beta}_k\right)}.$$

Más adelante discutiremos tests que conciernan varios parámetros de regresión en forma simultánea.

4.9.4. Aplicación al ejemplo

Antes de seguir presentando teoría, veamos cómo se calculan e interpretan estas cuestiones en el ejemplo de los 100 bebés de bajo peso. Para dicho ejemplo, cuyo modelo contenía a la edad gestacional y el peso al nacer como variables explicativas, $p-1$ resulta ser igual a 2 (luego $p=3$). La distribución t involucrada en la construcción de intervalos de confianza o tests para los β_k tiene en este caso $n-p = 100-3 = 97$ grados de libertad. En la Tabla 19 que figura en la página 125 exhibimos los coeficientes estimados. Los recordamos a continuación

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.3080154  1.5789429   5.262 8.54e-07
gestage     0.4487328  0.0672460   6.673 1.56e-09
birthwt     0.0047123  0.0006312   7.466 3.60e-11
```

Luego,

$$\widehat{\beta}_0 = 8{,}3080 \quad \widehat{\beta}_1 = 0{,}4487 \quad \widehat{\beta}_2 = 0{,}0047$$

y sus errores estándares respectivos resultan ser

$$\sqrt{\widehat{Var}\left(\widehat{\beta}_0\right)} = s\left(\widehat{\beta}_0\right) = 1{,}5789 \quad s\left(\widehat{\beta}_1\right) = 0{,}0672 \quad s\left(\widehat{\beta}_2\right) = 0{,}00063$$

Luego, los respectivos estadísticos t observados en cada caso son

$$T = \frac{\widehat{\beta}_1 - 0}{\sqrt{\widehat{Var}\left(\widehat{\beta}_1\right)}} = \frac{0{,}4487}{0{,}0672} = 6{,}67$$

cuando $k=1$ y

$$T = \frac{\widehat{\beta}_2 - 0}{\sqrt{\widehat{Var}\left(\widehat{\beta}_2\right)}} = \frac{0{,}0047}{0{,}00063} = 7{,}46$$

cuando $k = 2$. En ambos casos, los p-valores resultan ser menores que 0,001. Observemos que en la salida de cualquier paquete estadístico figuran tanto las estimaciones de los betas, como sus desvíos estándares estimados, los valores de t observados y los p-valores respectivos. En ambos casos rechazamos las hipótesis nulas a nivel 0,05 y concluimos que β_1 es distinta de cero cuando en el modelo aparece X_2 como explicativa (en el primer test) y que β_2 es distinta de cero cuando en el modelo aparece X_1 como explicativa (en el segundo test). Como además ambos estimadores son positivos, concluimos que el perímetro cefálico aumenta cuando aumenta tanto la edad gestacional como cuando aumenta el peso al nacer. Debemos tener presente, sin embargo, que varios tests de hipótesis basados en los mismos datos no son independientes; si cada test se realiza a nivel de significación α, la probabilidad global de cometer un error de tipo I –o rechazar la hipótesis nula cuando es verdadera– es, de hecho, mayor que α. Para eso se pueden realizar los tests simultáneos presentados, como los de Bonferroni.
Los intervalos de confianza para ambos parámetros de la regresión resultan ser

$$\begin{aligned}
&\widehat{\beta}_1 \pm t_{97,0,975}\sqrt{\widehat{Var}\left(\widehat{\beta}_1\right)} \\
&= [0{,}4487 - 1{,}9847 \cdot 0{,}06724; \quad 0{,}4487 + 1{,}9847 \cdot 0{,}06724] \\
&= [0{,}315\,25; \quad 0{,}582\,15]
\end{aligned}$$

y

$$\begin{aligned}
&\widehat{\beta}_2 \pm t_{97,0,975}\sqrt{\widehat{Var}\left(\widehat{\beta}_2\right)} \\
&= [0{,}004712 - 1{,}9847 \cdot 0{,}00063; \quad 0{,}004712 + 1{,}9847 \cdot 0{,}00063] \\
&= [0{,}00346; \quad 0{,}00596]
\end{aligned}$$

o, calculados con el `R`, como figuran en la Tabla 22.
Si usáramos el procedimiento de Bonferroni para construir los intervalos, tendríamos que usar el percentil

$$1 - \frac{\alpha}{2g} = 1 - \frac{0{,}05}{2 \cdot 3} = 0{,}99167$$

de una $t_{97,}$ es decir, $t_{97,0,9917} = 2{,}43636$ en vez de $t_{97,0,975} = 1{,}9847$, que nos dará intervalos más anchos, como puede observarse comparando los intervalos de confianza de las Tablas 22 y 23, la primera contiene a los intervalos de confianza de nivel 0,95 cada uno, y la segunda contiene los intervalos de confianza de nivel simultáneo 0,95.
Si calculamos el R^2 para este modelo (que figura en la Tabla 19) vemos que es $R^2 = 0{,}752$, luego el modelo que contiene a la edad gestacional y el peso al nacer

Tabla 22: Intervalos de confianza de nivel 0,95 para β_0, β_1 y β_2 para los datos de niños de bajo peso al nacer

```
> confint(ajuste2)
                 2.5 %       97.5 %
(Intercept) 5.174250734 11.441780042
gestage     0.315268189  0.582197507
birthwt     0.003459568  0.005964999
```

Tabla 23: Intervalos de confianza de nivel simultáneo 0,95 para β_0, β_1 y β_2 para los datos de niños de bajo peso al nacer, construidos con el método de Bonferroni

```
> confint(ajuste2,level=(1-(0.05/3)))
                0.833 %     99.167 %
(Intercept) 4.461384677 12.154646098
gestage     0.284907765  0.612557932
birthwt     0.003174601  0.006249966
> 0.05/(2*3)
[1] 0.008333333
```

como variables explicativas explica el 75,20 % de la variabilidad en los datos observados de perímetro cefálico; el modelo que tenía solamente a la edad gestacional explicaba el 60,95 %. Este aumento en el R^2 sugiere que agregar la variable peso al modelo mejora nuestra habilidad para predecir el perímetro cefálico para la población de bebés nacidos con bajo peso. Pero, como ya vimos, debemos ser muy cuidadosos al comparar coeficientes de determinación de dos modelos diferentes. Ya dijimos que la inclusión de una nueva covariable al modelo nunca puede hacer que el R^2 decrezca; el conocimiento de la edad gestacional y el peso al nacer, por ejemplo, nunca puede explicar menos de la variabilidad observada en los perímetros cefálicos que el conocimiento de la edad gestacional sola (aun si la variable peso no contribuyera en la explicación). Para sortear este problema podemos usar una segunda medida (cuando el interés sea comparar el ajuste que producen dos o más modelos entre sí), el R^2 ajustado (que notaremos R^2_a), que compensa por la complejidad extra que se le agrega al modelo. El R^2 ajustado aumenta cuando la inclusión de una variable mejora nuestra habilidad para predecir la variable y disminuye cuando no lo hace. Consecuentemente, el R^2 ajustado nos permite hacer

una comparación más justa entre modelos que contienen diferente número de covariables. Como el coeficiente de determinación, el R^2 ajustado es una estimación del coeficiente de correlación poblacional ρ; a diferencia del R^2, sin embargo, no puede ser directamente interpretado como la proporción de la variabilidad de los valores Y que queda explicada por el modelo de regresión. En este ejemplo, el R^2 ajustado resulta ser 0,7469 (ver nuevamente la Tabla 19) que al ser mayor que el R^2 ajustado del modelo con sólo una variable explicativa, la edad gestacional (era $R_a^2 = 0{,}6055$) indica que la inclusión del peso al nacer en el modelo, mejora nuestra capacidad para predecir el perímetro cefálico del niño.
Finalmente, la tabla ANOVA para estos datos aparece en la Figura 40 con el SPSS y en la Tabla 24 con el R.

Figura 40: Tabla de ANOVA para los datos de niños de bajo peso al nacer

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	477,327	2	238,663	147,058	,000[a]
	Residual	157,423	97	1,623		
	Total	634,750	99			

a. Variables predictoras: (Constante), birthwt, Edad gestacional (semanas)
b. Variable dependiente: Perímetro cefálico (centímetros)

Tabla 24: Tabla de Anova para los datos de bebés de bajo peso, en R.

```
> ajuste2<-lm(headcirc~gestage+birthwt)
> ajuste1<-lm(headcirc~1)
> anova(ajuste1,ajuste2)
Analysis of Variance Table

Model 1: headcirc ~ 1
Model 2: headcirc ~ gestage + birthwt
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     99 634.75
2     97 157.42  2    477.33 147.06 < 2.2e-16 ***
---
```

Observemos que el estimador de σ^2 que surge del modelo de regresión es

$$\text{MSRes} = \frac{\text{SSRes}}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 = 1{,}62, \tag{54}$$

por la Sección 4.8.4. Si comparamos el valor observado del estimador $\widehat{\sigma}^2 = 1{,}62$ para este modelo con el estimador de la varianza no explicada por el modelo de regresión lineal simple que sólo tiene a la edad gestacional como explicativa, que era 2,529 (ver Tabla 2.8) observamos que con la inclusión del peso hemos reducido la variabilidad no explicada por el modelo, mejorando la calidad del ajuste obtenido (y de las predicciones que pueden hacerse con él).

4.10. Estimación de la Respuesta Media

4.10.1. Intervalo de confianza para $E(Y_h)$

Nos interesa estimar la respuesta media o esperada cuando $(X_1, \ldots, X_{p-1})$ toma el valor dado $(X_{h1}, \ldots, X_{h,p-1})$. Notamos a esta respuesta media por $E(Y_h)$ o bien $E(Y_h \mid (X_{h1}, \ldots, X_{h,p-1}))$. Como en regresión lineal simple estos valores $(X_{h1}, \ldots, X_{h,p-1})$ pueden ser valores que hayan ocurrido en la muestra considerada o pueden ser algunos otros valores de las variables predictoras dentro del alcance (*scope*) del modelo. Definimos el vector

$$\mathbf{X}_h = \begin{bmatrix} 1 \\ X_{h1} \\ \vdots \\ X_{h,p-1} \end{bmatrix}$$

de modo que la respuesta a ser estimada es

$$E(Y_h) = E(Y_h \mid \mathbf{X}_h) = \mathbf{X}_h^t \boldsymbol{\beta}.$$

La respuesta media estimada correspondiente a $\mathbf{X}_h$, que denotamos por $\widehat{Y}_h$ es la variable aleatoria que se calcula del siguiente modo

$$\widehat{Y}_h = \mathbf{X}_h^t \widehat{\boldsymbol{\beta}} = \widehat{\beta}_0 + \widehat{\beta}_1 X_{h1} + \widehat{\beta}_2 X_{h2} + \cdots + \widehat{\beta}_{p-1} X_{h,p-1}.$$

Para el modelo de errores normales (45) la distribución de $\widehat{Y}_h$ será normal, con media

$$E\left(\widehat{Y}_h\right) = \mathbf{X}_h^t \boldsymbol{\beta} = E(Y_h) \tag{55}$$

y varianza

$$Var\left(\widehat{Y}_h\right) = \sigma^2 \mathbf{X}_h^t \left(\mathbf{X}^t \mathbf{X}\right)^{-1} \mathbf{X}_h = \mathbf{X}_h^t Var\left(\widehat{\boldsymbol{\beta}}\right) \mathbf{X}_h.$$

Como la esperanza del predicho es igual a lo que queremos estimar, es decir, $E\left(\widehat{Y}_h\right) = E\left(Y_h\right)$, el estimador resulta ser insesgado. La varianza estimada resulta ser

$$\widehat{Var}\left(\widehat{Y}_h\right) = \text{MSRes} \cdot \mathbf{X}_h^t \left(\mathbf{X}^t\mathbf{X}\right)^{-1} \mathbf{X}_h = \mathbf{X}_h^t \widehat{Var}\left(\widehat{\boldsymbol{\beta}}\right) \mathbf{X}_h. \tag{56}$$

A partir de (55) y (56) puede obtenerse un intervalo de confianza de nivel $1-\alpha$ para $E\left(Y_h\right)$, la respuesta media esperada cuando las covariables son $\mathbf{X}_h$, que viene dado por

$$\widehat{Y}_h \pm t_{n-p,1-\alpha/2} \cdot \sqrt{\widehat{Var}\left(\widehat{Y}_h\right)}. \tag{57}$$

En general, estos intervalos serán calculados usando un paquete estadístico.

4.10.2. Región de Confianza para la Superficie de Regresión

La región de confianza para toda la superficie de regresión es una extensión de la banda de confianza de Hotelling o Working-Hotelling para una recta de regresión (cuando hay una sola variable predictora). Los puntos de la frontera de la región de confianza en $\mathbf{X}_h$, se obtienen a partir de

$$\widehat{Y}_h \pm W \cdot \sqrt{\widehat{Var}\left(\widehat{Y}_h\right)}.$$

donde

$$W^2 = pF_{p,n-p;1-\alpha}. \tag{58}$$

Puede probarse que eligiendo este percentil, la región resultante cubrirá a la superficie de regresión **para todas las combinaciones posibles de las variables** $\mathbf{X}$ (dentro de los límites observados), con nivel $1-\alpha$. Es por eso que esta región de confianza tiene nivel simultáneo o global $1-\alpha$, como discutimos en la Sección 4.9.3.

4.10.3. Intervalos de Confianza Simultáneos para Varias Respuestas Medias

Para estimar un número de respuestas medias $E\left(Y_h\right)$ correspondientes a distintos vectores $\mathbf{X}_h$ con coeficiente de confianza global $1-\alpha$ podemos emplear dos enfoques diferentes:

1. Usar las regiones de confianza para la superficie de regresión basadas en la distribución de Hotelling (58) para varios vectores $\mathbf{X}_h$ de interés

$$\widehat{Y}_h \pm W \cdot \sqrt{\widehat{Var}\left(\widehat{Y}_h\right)}.$$

donde $\widehat{Y}_h, W$ y $\widehat{Var}\left(\widehat{Y}_h\right)$ están definidos respectivamente en (55), (58) y (56). Como la región de confianza para la superficie de regresión basada en la distribución de Hotelling cubre la respuesta media para todos los vectores $\mathbf{X}_h$ posibles con nivel conjunto $1-\alpha$, los valores de frontera seleccionados cubrirán las respuestas medias para los vectores $\mathbf{X}_h$ de interés con nivel de confianza global mayor a $1-\alpha$.

2. Usar intervalos de confianza simultáneos de Bonferroni. Cuando se quieren hallar g intervalos de confianza simultáneos, los límites serán

$$\widehat{Y}_h \pm B \cdot \sqrt{\widehat{Var}\left(\widehat{Y}_h\right)}.$$

donde

$$B = t_{n-p,1-\frac{\alpha}{2g}}.$$

Para una aplicación en particular, podemos comparar los valores de W y B para ver cuál procedimiento conduce a tener los intervalos de confianza más angostos. Si los niveles $\mathbf{X}_h$ no son conocidos antes de aplicar el modelo, sino que surgen del análisis, es mejor usar los intervalos basados en la distribución de Hotelling, puesto que la familia de estos intervalos incluye a todos los posibles valores de $\mathbf{X}_h$.

4.11. Intervalos de Predicción para una Nueva Observación $Y_{h(\mathbf{nueva})}$

Como en el caso de regresión lineal simple, estamos interesados ahora en predecir una nueva observación Y correspondiente a un nivel dado de las covariables $\mathbf{X}_h$. La nueva observación Y a ser predicha se puede ver como el resultado de una nueva repetición del experimento u observación, independiente de los resultados anteriores en los que se basa el análisis de regresión. Denotamos el nivel de $\mathbf{X}$ para la nueva observación por $\mathbf{X}_h$ y a la nueva observación de Y como $Y_{h(\text{nueva})}$. Por supuesto, asumimos que el modelo de regresión subyacente aplicable a los datos con los que contamos sigue siendo apropiado para la nueva observación.
La diferencia entre la estimación de la respuesta media $E(Y_h)$, tratado en la sección anterior, y la predicción de una nueva respuesta $Y_{h(\text{nueva})}$, que discutimos en esta, es básica. En el primer caso, se estima la media de la distribución de Y. En el segundo caso, queremos predecir un *resultado individual* surgido a partir de la distribución de Y. Por supuesto, la gran mayoría de los resultados individuales se desvían de la respuesta media, y esto debe ser tenido en cuenta por el procedimiento para la predicción de la $Y_{h(\text{nueva})}$.

4.11.1. Intervalo de predicción para $Y_{h(\text{nueva})}$ cuando los parámetros son conocidos

Para ilustrar la naturaleza de un intervalo de predicción para una nueva observación de la $Y_{h(\text{nueva})}$ de la manera más simple posible, en primer lugar supondremos que todos los parámetros de regresión son conocidos. Más adelante abandonaremos este supuesto para tener el enfoque realista y haremos las modificaciones pertinentes.
Consideremos el ejemplo de los niños con bajo peso al nacer. Supongamos que supiéramos que los parámetros del modelo son

$$\begin{aligned} \beta_0 = 8 \quad \beta_1 = 0{,}5 \quad \beta_2 = 0{,}004 \quad \sigma = 1{,}25 \\ E(Y) = 8 + 0{,}5X_1 + 0{,}004X_2 \end{aligned}$$

El analista considera ahora un bebé de 30 semanas de edad gestacional y que pesó 1360g. al nacer. El perímetro cefálico medio para $X_{h1} = 30$ y $X_{h2} = 1360$ es

$$E(Y) = 8 + 0{,}5 \cdot 30 + 0{,}004 \cdot 1360 = 28{,}44$$

En la Figura 41 se muestra la distribución para Y_h para $\mathbf{X}_h^t = (1, 30, 1360)$. Su media es $E(Y_h) = 28{,}44$ y su desvío estándar es $\sigma = 1{,}25$. La distribución es normal debido al modelo de regresión (44) y (45).
Supongamos que fuéramos a predecir el perímetro cefálico de un bebé con estos valores de las covariables, diríamos que está entre

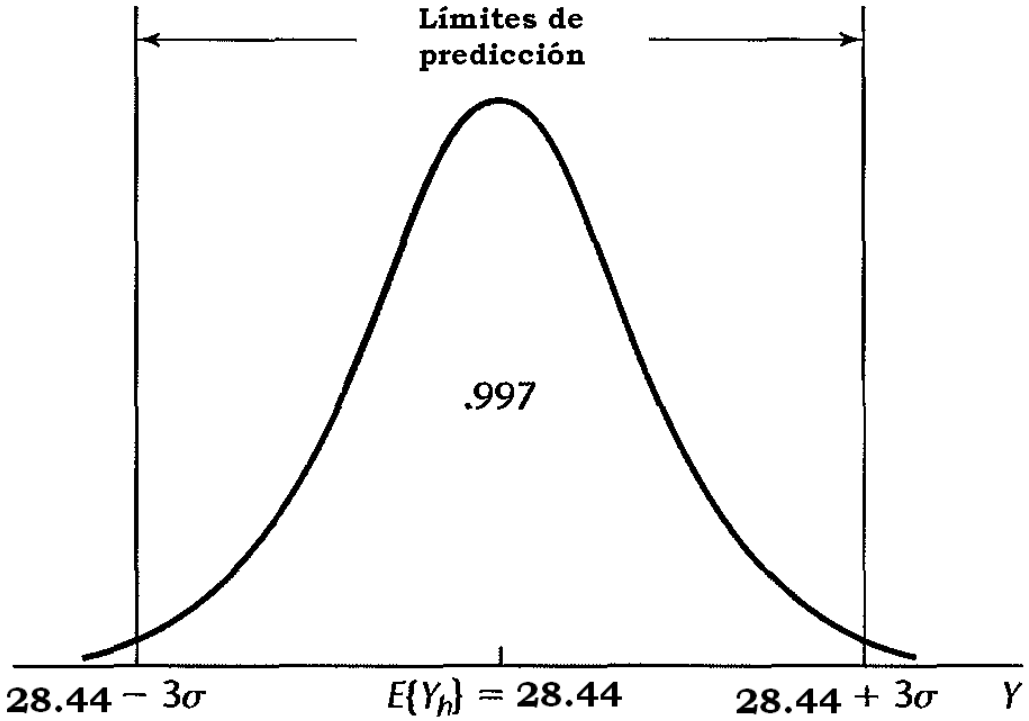
$$\begin{aligned} E(Y_h) \pm 3\sigma \\ 28{,}44 \pm 3 \cdot 1{,}25 \end{aligned}$$

de modo que el intervalo de predicción sería

$$24{,}69 \leq Y_{h(\text{nueva})} \leq 32{,}19$$

Como el 99,7 por ciento del área en una distribución de probabilidad normal cae dentro de los tres desvíos estándares de la media, hay una probabilidad de 0,997 de que este intervalo de predicción dé una predicción correcta para el perímetro cefálico del bebé en cuestión, con 30 semanas de gestación y que pesó 1360g. al nacer. Los límites de predicción en este caso son bastante amplios, por lo que la predicción no es muy precisa, sin embargo, el intervalo de predicción indica que el bebé tendrá un perímetro cefálico mayor a 24 cm., por ejemplo.
La idea básica de un intervalo de predicción es, pues, elegir un rango en la distribución de Y en donde la mayoría de las observaciones caerá, y luego, declarar que la observación siguiente caerá en este rango. La utilidad del intervalo de predicción

Figura 41: Distribución de Y_h cuando $\mathbf{X}_h^t = (1, 30, 1360)$. Fuente: Kutner et al. [2005], pág. 57.



depende, como siempre, del ancho del intervalo y de la necesidad de precisión por parte del usuario.
En general, cuando los parámetros del modelo de regresión con errores normales son conocidos, los límites de la predicción de la $Y_{h(\text{nueva})}$ son

$$E(Y_h) \pm z_{1-\frac{\alpha}{2}} \cdot \sigma \tag{59}$$

4.11.2. Intervalo de predicción para $Y_{h(\text{nueva})}$ cuando los parámetros son desconocidos

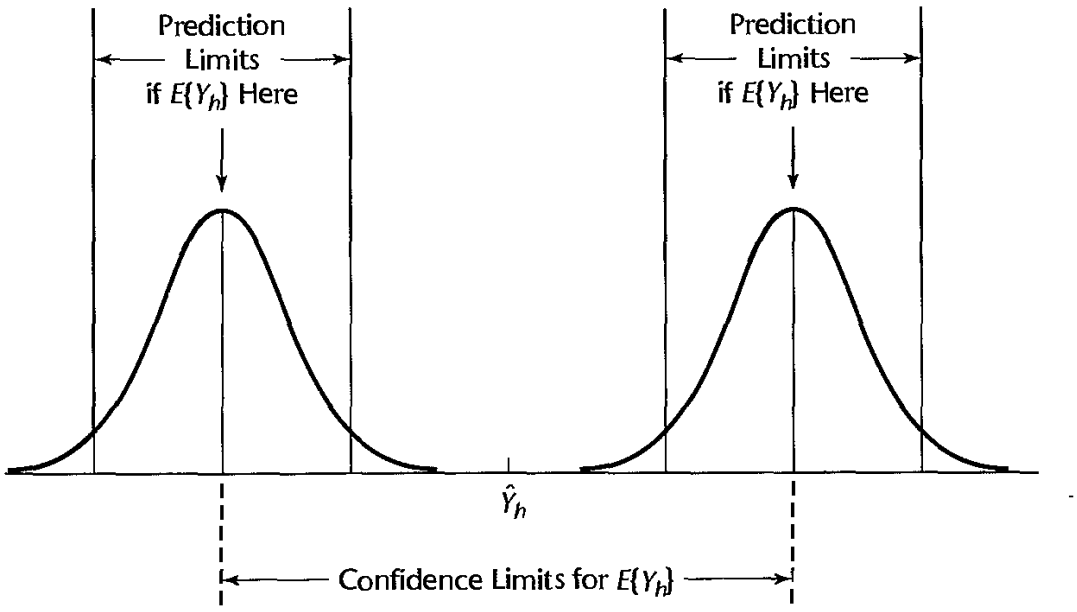
Cuando los parámetros de regresión son desconocidos, deben ser estimados. La media de la distribución de Y se estima por $\widehat{Y}_h$, como de costumbre, y la varianza de la distribución de Y se estima por la MSRes. No podemos, sin embargo sólo utilizar los límites de la predicción de (59) con los parámetros reemplazados por los estimadores puntuales correspondientes. La razón de ello es ilustrada de manera

intuitiva en la Figura 42. En ella se muestran dos distribuciones de probabilidad de Y, que corresponde a los límites superior e inferior de un intervalo de confianza para $E(Y_h)$. En otras palabras, la distribución de Y puede ser ubicada tan a la izquierda como la distribución que se exhibe a la extrema izquierda, o tan a la derecha como la distribución que se exhibe a la extrema derecha, o en cualquier lugar en el medio. Dado que no sabemos la media $E(Y_h)$ y sólo la podemos estimar por un intervalo de confianza, no podemos estar seguros de la localización de la distribución de Y.
La Figura 42 también muestra los límites de predicción para cada una de las dos distribuciones de probabilidad de Y allí presentadas. Ya que no podemos estar seguros de la localización del centro de la distribución de Y, los límites de la predicción de $Y_{h(\text{nueva})}$ claramente deben tener en cuenta dos elementos, como se muestra en la Figura 42:

1. La variación en la posible ubicación de la (esperanza o centro de la) distribución de Y.

2. La variación dentro de la distribución de probabilidad de Y.

Figura 42: Predicción de $Y_{h(\text{nueva})}$ cuando los parámetros son desconocidos. Fuente: Kutner et al. [2005], pág 58.



Los límites de predicción para una nueva observación $Y_{h(\text{nueva})}$ en un determinado nivel $\mathbf{X}_h$ se obtienen por medio del siguiente resultado

$$\frac{Y_{h(\text{nueva})} - \widehat{Y_h}}{s\,(\text{pred})} \sim t_{n-p} \tag{60}$$

Observemos que en el estadístico de Student utilizamos el estimador puntual $\widehat{Y_h}$ en el numerador y no la verdadera media $E\,(Y_h)$ porque la media real se desconoce y no puede ser utilizada al hacer la predicción. El desvío estándar estimado de la predicción, $s\,(\text{pred})$, en el denominador se define por

$$\begin{aligned} s^2\,(\text{pred}) &= \text{MSRes} \; + \widehat{Var}\left(\widehat{Y}_h\right) \\ &= \text{MSRes} \cdot \left(1 + \mathbf{X}_h'\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}_h\right), \end{aligned}$$

de manera análoga a lo que habíamos calculado para el modelo de regresión lineal simple. A partir de dicho resultado, el intervalo de predicción de la $Y_{h(\text{nueva})}$ correspondiente a $\mathbf{X}_h$ de nivel $1-\alpha$ es

$$\begin{aligned} &\widehat{Y}_h \pm t_{n-p,1-\alpha/2} \cdot s\,(\text{pred}) \\ &\widehat{Y}_h \pm t_{n-p,1-\alpha/2} \cdot \sqrt{\text{MSRes} \cdot \left(1 + \mathbf{X}_h'\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}_h\right)} \end{aligned}$$

Observemos que el numerador del estadístico de Student (60) representa cuán lejos se desviará la nueva observación $Y_{h(\text{nueva})}$ de la media estimada $\widehat{Y}_h$ basada en los n casos originales en el estudio. Esta diferencia puede ser vista como el error de predicción, con $\widehat{Y}_h$ jugando el papel de la mejor estimación puntual del valor de la nueva observación $Y_{h(\text{nueva})}$. La varianza de este error de predicción puede ser fácilmente obtenida mediante la utilización de la independencia de la nueva observación, $Y_{h(\text{nueva})}$ y los n casos originales de la muestra en la que se basa $\widehat{Y}_h$.

$$\begin{aligned} Var\,(\text{pred}) &= Var\left(Y_{h(\text{nueva})} - \widehat{Y_h}\right) \\ &= Var\left(Y_{h(\text{nueva})}\right) + Var\left(\widehat{Y_h}\right) \\ &= \sigma^2 + Var\left(\widehat{Y_h}\right) \end{aligned}$$

Luego, la varianza del error de predicción $Var\,(\text{pred})$ tiene dos componentes:

1. La varianza de la distribución de Y en $\mathbf{X} = \mathbf{X}_h$, es decir, σ^2.
2. La varianza de la distribución muestral de $\widehat{Y_h}$, es decir, $Var\left(\widehat{Y_h}\right)$.

Un estimador insesgado de $Var\,(\text{pred})$ es

$$s^2\,(\text{pred}) = \text{MSRes } + \widehat{Var}\left(\widehat{Y}_h\right).$$

Por supuesto, como este estimador es siempre mayor que $\widehat{Var}\left(\widehat{Y}_h\right)$, que aparece en el intervalo de confianza (57), el intervalo de predicción de la $Y_{h(\text{nueva})}$ correspondiente a $\mathbf{X}_h$ de nivel $1-\alpha$ siempre será más largo que el intervalo de confianza de nivel $1-\alpha$ para $E\,(Y_h)$, la respuesta media esperada cuando las covariables son $\mathbf{X}_h$.

4.11.3. Ejemplo de cálculo de Intervalo de Confianza para $E\,(Y_h)$ y de un Intervalo de Predicción para $Y_{h(\textbf{nueva})}$

Apliquemos estos dos resultados (cálculo de intervalo de confianza e intervalo de predicción) a un caso particular, usando los datos de bebés de bajo peso. Buscamos un intervalo de confianza para la media del perímetro cefálico de un bebé con 30 semanas de gestación y que pesó 1360g. al nacer, de nivel 0,95. El intervalo de confianza resulta ser

Tabla 25: Intervalos de confianza y predicción de nivel 0,95 para los datos de niños de bajo peso al nacer, para edad gestacional de 30 semanas y peso al nacer de 1360g.

```
> new<-data.frame(gestage=30, birthwt= 1360)
> predict.lm(ajuste2,new,interval="confidence")
       fit      lwr      upr
1 28.17871 27.81963 28.53778
> predict.lm(ajuste2,new,interval="prediction")
       fit      lwr      upr
1 28.17871 25.62492 30.73249
```

O, bien, operando a mano, la matriz de varianzas de los coeficientes beta da

```
> vcov(sal2)
             (Intercept)       gestage       birthwt
(Intercept)  2.4930607944 -9.986181e-02  3.714576e-04
gestage     -0.0998618122  4.522022e-03 -2.801056e-05
birthwt      0.0003714576 -2.801056e-05  3.983870e-07
```

Recordemos que $\widehat{Var}\left(\widehat{Y}_h\right)$ está definida en (56), luego

$$\begin{aligned}
&\widehat{Var}\left(\widehat{Y}_h\right) \\
&= \mathbf{X}_h^t \widehat{Var}\left(\widehat{\boldsymbol{\beta}}\right) \mathbf{X}_h \\
&= \begin{bmatrix}1 & 30 & 1360\end{bmatrix} \begin{bmatrix} 2{,}4930607944 & -9{,}986181 \times 10^{-2} & 3{,}714576 \times 10^{-4} \\ -0{,}0998618122 & 4{,}522022 \times 10^{-3} & -2{,}801056 \times 10^{-5} \\ 0{,}0003714576 & -2{,}801056 \times 10^{-5} & 3{,}983870 \times 10^{-7} \end{bmatrix} \begin{bmatrix} 1 \\ 30 \\ 1360 \end{bmatrix} \\
&= 0{,}032731
\end{aligned}$$

Como

$$\begin{aligned}
t_{n-p,1-\alpha/2} &= t_{97,0,975} = 1{,}984723 \\
\widehat{Y}_h &= 8{,}3080 + 0{,}4487 \cdot 30 + 0{,}0047122 \cdot 1360 = 28{,}178
\end{aligned}$$

resulta que el intervalo de confianza de nivel $1-\alpha = 0{,}95$ para $E\left(Y_h\right)$, la respuesta media esperada cuando las covariables son $\mathbf{X}_h$, es

$$\begin{aligned}
&\widehat{Y}_h \pm t_{n-p,1-\alpha/2} \cdot \sqrt{\widehat{Var}\left(\widehat{Y}_h\right)} \\
&28{,}178 \pm 1{,}984723 \cdot \sqrt{0{,}032731} \\
&28{,}178 \pm 0{,}359\,07
\end{aligned}$$

es decir

$$[27{,}819; \quad 28{,}537]$$

Por otro lado, el intervalo de predicción de la $Y_{h(\text{nueva})}$ correspondiente a $\mathbf{X}_h$ de nivel $1-\alpha = 0{,}95$ es

$$\begin{aligned}
&\widehat{Y}_h \pm t_{n-p,1-\alpha/2} \cdot s\,(\text{pred}) \\
&\widehat{Y}_h \pm t_{n-p,1-\alpha/2} \cdot \sqrt{\text{MSRes} \; + \widehat{Var}\left(\widehat{Y}_h\right)}
\end{aligned}$$

Como

$$\text{MSRes} \; = 1{,}62,$$

el intervalo de predicción de la $Y_{h(\text{nueva})}$ resulta ser

$$\begin{aligned}
&28{,}178 \pm 1{,}984723 \cdot \sqrt{1{,}62 + 0{,}032731} \\
&28{,}178 \pm 2{,}551\,5
\end{aligned}$$

es decir,

$$[25{,}62; \quad 30{,}730]\,.$$

4.11.4. Precaución Respecto de Extrapolaciones Ocultas

Al estimar una respuesta media o al predecir una nueva observación en la regresión múltiple, hay que tener especial cuidado de que la estimación o predicción esté comprendida dentro del alcance del modelo. El peligro, por supuesto, es que el modelo puede no ser apropiado cuando se lo extiende fuera de la región de las observaciones. En regresión múltiple, es particularmente fácil perder la noción de esta región ya que los niveles de $X_1, ..., X_{p-1}$ definen a la región en *forma conjunta*. Por lo tanto, uno no puede simplemente mirar los rangos de cada variable predictora de forma individual. Para visualizar el problema, consideremos la Figura 43, donde la región sombreada es la región de las observaciones para una regresión múltiple con dos variables de predicción y el punto con un círculo alrededor representa los valores (X_{h1}, X_{h2}) para los que se desea predecir la $Y_{h(\text{nueva})}$. Dicho punto está dentro de los rangos de las variables predictoras X_1 y X_2 en forma individual, sin embargo, está bien fuera de la región conjunta de las observaciones. Cuando sólo hay dos variables de predicción es fácil descubrir que se está frente a esta extrapolación, a través de un scatterplot (o gráfico de dispersión) pero esta detección se hace mucho más difícil cuando el número de variables predictivas es muy grande. Se discute en la Sección 5.2 un procedimiento para identificar las extrapolaciones ocultas cuando hay más de dos variables predictoras.

4.12. Ejercicios (primera parte)

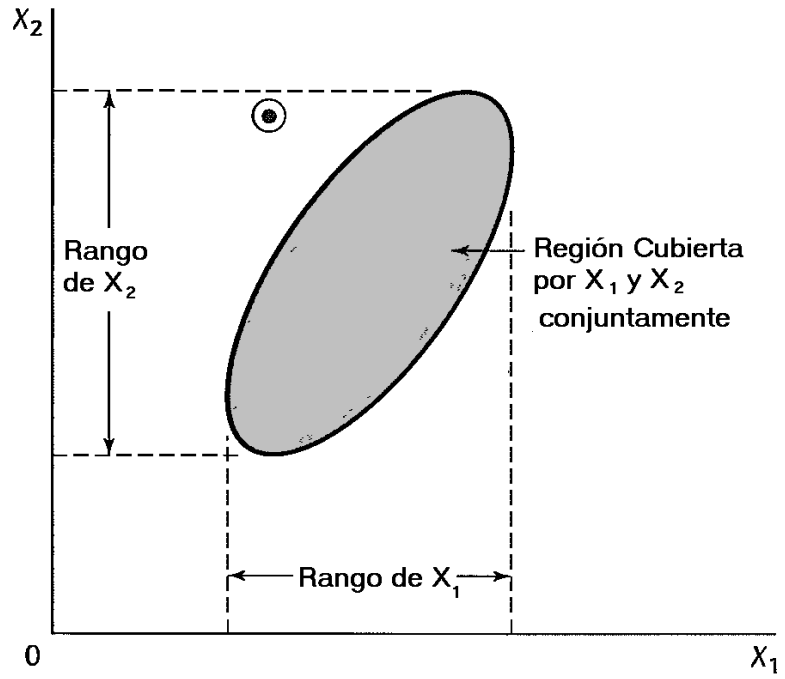
Ejercicio 4.1 *Medidas del cuerpo V. Base de datos* `bdims` *del paquete* `openintro`.

(a) En el ejercicio 2.1 explicamos el peso de las personas registradas en esta base de datos, por el contorno de la cadera y en el ejercicio 2.2 la explicamos con un modelo con la altura como covariable. Proponga un modelo de regresión múltiple que explique el peso medido en kilogramos (`wgt`) utilizando el contorno de la cadera medida en centímetros (`hip.gi`) y la altura media en centímetros (`hgt`) como covariables. Escriba el modelo que está ajustando. Realice el ajuste con el R.

(b) Interprete los coeficientes estimados. ¿Resultan significativos? Cambian sus valores respecto de los que tenían los coeficientes que acompañaban a estas variables en los modelos de regresión lineal simple?

(c) Evalúe la bondad del ajuste realizado, a través del R^2. Indique cuánto vale y qué significa. Se quiere comparar este ajuste con el que dan los dos modelos lineales simples propuestos en los ejercicios 2.1 y 2.2. ¿Es correcto comparar los R^2 de los tres ajustes? ¿Qué valores puedo comparar? ¿Es mejor este ajuste múltiple?

Figura 43: Región de observaciones en X_1 y X_2 conjuntamente, comparada con los rangos de X_1 y X_2 por separado.



(d) Estime la varianza de los errores. Compare este estimador con los obtenidos en los dos ajustes simples.

(e) Estime el peso esperado para la población de adultos cuyo contorno de cadera mide 100 cm y su altura es de 174cm. Dé un intervalo de confianza de nivel 0.95 para este valor esperado.

(f) Prediga el peso de un adulto cuyo contorno de cadera mide 100 cm y su altura es de 174cm. Dé un intervalo de predicción de nivel 0.95 para este valor. Compare las longitudes de los tres intervalos de predicción que se obtienen usando el modelo que solamente tiene al contorno de cadera como explicativa, al que solamente usa la altura y al modelo múltiple que contiene a ambas.

4.13. Predictores Categóricos

Hasta ahora hemos visto el modelo de regresión lineal simple o múltiple con uno o varios predictores continuos. Sin embargo, tanto en regresión lineal simple como múltiple los predictores pueden ser variables binarias, categóricas, numéricas discretas o bien numéricas continuas.

4.13.1. Predictores Binarios

Comencemos con un ejemplo.
Los niveles de glucosa por encima de 125 mg/dL son diagnóstico de diabetes, mientras que los niveles en el rango de 100 a 125 mg/dL señalan un aumento en el riesgo de progresar a esta condición grave. Por lo tanto, es de interés determinar si la actividad física, una característica del estilo de vida que es modificable, podría ayudar a las personas a reducir sus niveles de glucosa y, por ende, evitar la diabetes. Responder a esta pregunta de manera concluyente requeriría un ensayo clínico aleatorizado, lo cual es a la vez difícil y costoso. Por ello, preguntas como estas son con frecuencia, inicialmente respondidas utilizando datos observacionales. Pero esto es complicado por el hecho de que las personas que hacen ejercicio físico difieren en muchos aspectos de las que no lo hacen, y algunas de las otras diferencias podrían explicar cualquier asociación (no ajustada) entre el ejercicio físico y el nivel de glucosa.
Usaremos un modelo lineal simple para predecir el nivel de glucosa usando una medida de la cantidad y frecuencia de ejercicio físico que realizan. La base de datos está en el archivo `azucar.txt`, se compone de los datos de $n = 220$ personas. Corresponde a datos artificialmente creados. La pregunta que queremos responder es si el hecho de hacer actividad física puede contribuir a bajar el nivel de glucosa y ayudar a prevenir la progresión a la diabetes entre las personas personas en riesgo.

Hay muchas maneras de codificar numéricamente las clases de una variable cualitativa. Usaremos variables indicadoras que valen 0 ó 1. Estas variables indicadoras son fáciles de usar y son ampliamente utilizadas, pero de ninguna manera son la única forma de cuantificar una variable cualitativa. En la Observación 4.12 comentamos una propuesta alternativa de codificación. Para el ejemplo, definimos la variable indicadora (o binaria, o dummy) por

$$X_{i1} = \begin{cases} 1 & \text{si el iésimo paciente hace actividad física} \\ & \text{(al menos 3 veces por semana)} \\ 0 & \text{si no} \end{cases} \tag{61}$$

El modelo de regresión lineal para este caso es

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$$

La función de respuesta para este modelo de regresión es

$$E\left(Y \mid X_1\right) = \beta_0 + \beta_1 X_1. \tag{62}$$

Para entender el significado de los coeficientes de regresión en este modelo, consideremos primero el caso de una persona que no hace ejercicio. Para tal persona, $X_1 = 0$, y la función de respuesta (62) se reduce a

$$E\left(Y\right) = \beta_0 + \beta_1 0 = \beta_0 \qquad \text{no ejercita}$$

Para una persona que sí hace ejercicio, $X_1 = 1$, y la función de respuesta (62) se convierte en

$$E\left(Y\right) = \beta_0 + \beta_1 1 = \beta_0 + \beta_1 \qquad \text{ejercita}$$

Luego, el modelo de regresión lineal en este caso consiste simplemente en expresar la media del nivel de glucosa en cada población mediante dos coeficientes distintos, donde β_0 es la media de la glucosa para las personas que no ejercitan y $\beta_0 + \beta_1$ es la media de la glucosa para las personas que ejercitan; por lo tanto, β_1 es la diferencia (positiva o negativa, dependiendo del signo) en niveles medios de glucosa para las personas que ejercitan respecto de las que no. Observemos que esto es consistente con nuestra interpretación más general de β_j como el cambio en $E[Y|X_j]$ por un aumento de una unidad de X_j. En este caso, si el ejercicio estuviera asociado con menores niveles de glucosa (como se presume) β_1 debería ser negativo.
En la Tabla 26 presentamos el resultado de ajustar el modelo propuesto a los datos. Los datos están en el archivo `azucar.txt`. Las variables se denominan `glucosa` (la respuesta) y `ejercicio` la explicativa.
El coeficiente estimado para la actividad física (`ejercicio`) muestra que los niveles basales de glucosa fueron alrededor de 7,4 mg/dL más bajos para personas que hacían ejercicios al menos tres veces por semana que para las personas que ejercitaban menos. Esta diferencia es estadísticamente significativa ($t = -6{,}309$, $p - valor = 1{,}5410^{-9} < 0{,}05$).
Sin embargo, en la base de datos considerada, las personas que hacen ejercicio resultaron ser en promedio, un poco más jóvenes, un poco más propensas a consumir alcohol, y, en particular, como puede verse en la Figura 44 tienen en promedio un menor índice de masa corporal (BMI), todos factores asociados con el nivel de glucosa. Esto implica que el promedio más bajo de la glucosa que observamos entre las personas que hacen ejercicio puede deberse al menos en parte, a diferencias en estos otros predictores. En estas condiciones, es importante que nuestra estimación de la diferencia en los niveles promedio de glucosa asociados con el ejercicio se "ajuste" a los efectos de estos factores de confusión potenciales de la asociación sin ajustar. Idealmente, el ajuste de un modelo de regresión múltiple (o sea, de múltiples predictores) proporciona una estimación del efecto de ejercitar en el nivel medio de glucosa, manteniendo las demás variables constantes.

Tabla 26: Ajuste de la regresión para la variable glucosa con ejercicio como explicativa.

```
> ajuste1<-lm(glucosa ~ ejercicio, data = azucar)
> summary(ajuste1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  98.9143     0.8512 116.212  < 2e-16 ***
ejercicio    -7.4273     1.1773  -6.309 1.54e-09 ***
---

Residual standard error: 8.722 on 218 degrees of freedom
Multiple R-squared:  0.1544,        Adjusted R-squared:  0.1505
F-statistic:  39.8 on 1 and 218 DF,  p-value: 1.545e-09
```

Observación 4.11 *¿Qué pasa si ponemos dos variables binarias para modelar ejercicio? O sea, si definimos X_1 como antes,*

$$X_{i1} = \begin{cases} 1 & \textit{si la iésima persona ejercita} \\ \\ 0 & \textit{si no} \end{cases}$$

y

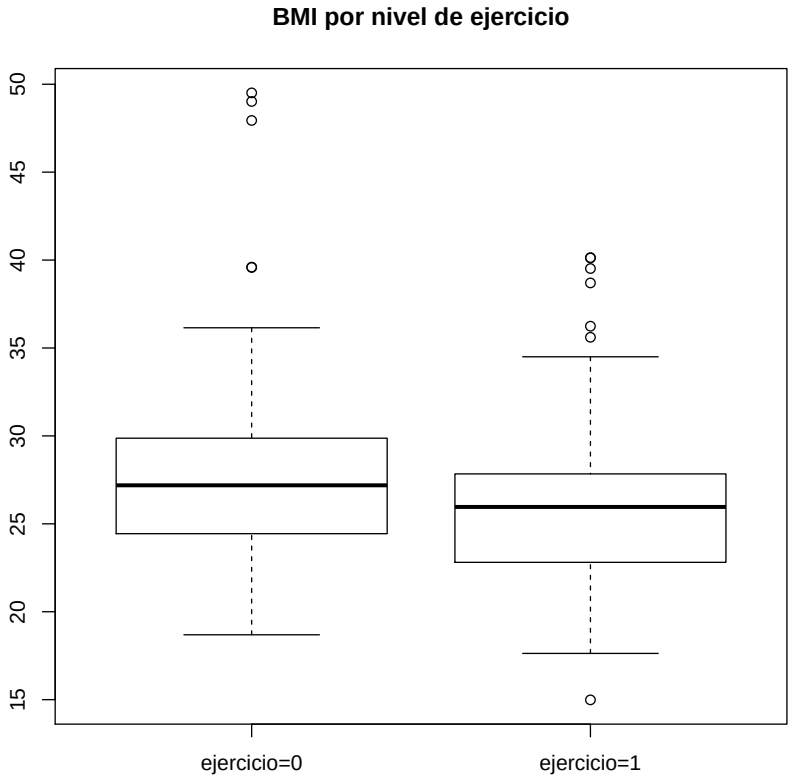
$$X_{i2} = \begin{cases} 1 & \textit{si la iésima persona no ejercita} \\ \\ 0 & \textit{si no} \end{cases}$$

Acá decimos que ejercita si hace actividad física más de tres veces por semana. Entonces el modelo sería

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \tag{63}$$

Esta manera intuitiva de incorporar una variable indicadora para cada clase de la predictora cualitativa, desafortunadamente, conduce a problemas tanto estadísticos (de identificación de parámetros) como computacionales. Para verlo, supongamos que tuviéramos $n = 4$ observaciones, las primeras dos compuestas por personas que ejercitan ($X_1 = 1,\ X_2 = 0$) y las dos segundas que no lo hacen

Figura 44: Boxplot del `bmi`, separados por niveles de la variable `ejercicio`, para los datos del archivo `azucar`.



$(X_1 = 0,\ X_2 = 1\ )$. *Entonces la matriz* X *sería*

$$X = \begin{array}{c} \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} & X_1 & X_2 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \end{array}$$

Observemos que la suma de las columnas X_1 *y* X_2 *da la primer columna, de modo que las columnas de esta matriz son linealmente dependientes. Esto tiene un efecto*

serio en la matriz X^tX.

$$X^tX = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix}$$

Vemos que la primer columna de la matriz X^tX *es igual a la suma de las últimas dos, de modo que las columnas son linealmente dependientes. Luego, la matriz* X^tX *no tiene inversa, y por lo tanto, no se pueden hallar únicos estimadores de los coeficientes de regresión. De hecho, no hay unicidad tampoco en los parámetros del modelo (lo que en estadística se conoce como identificabilidad de los parámetros) puesto que la función de respuesta para el modelo (63) es*

$$E(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = \begin{cases} \beta_0 + \beta_1 & \textit{si ejercita} \\ \\ \beta_0 + \beta_2 & \textit{si no ejercita} \end{cases}$$

En particular, tomando

$$\begin{aligned} \beta_0 &= a \\ \beta_1 &= b \\ \beta_2 &= c \end{aligned}$$

o bien

$$\begin{aligned} \beta_0 &= a - b \\ \beta_1 &= 2b \\ \beta_2 &= c \end{aligned}$$

resulta, en ambos casos

$$E(Y \mid X_1, X_2) = \begin{cases} a + b & \textit{si ejercita} \\ \\ a + c & \textit{si no ejercita} \end{cases}$$

para cualesquiera números reales a, b, c. *Una salida simple a este problema es desprenderse de una de las variables indicadoras. En nuestro ejemplo nos deshacemos de* X_2. *Esta forma de resolver el problema de identificabilidad no es la única pero, como hemos visto, permite una interpretación sencilla de los parámetros. Otra posibilidad en este caso consiste en eliminar* β_0 *y proponer el modelo*

$$E(Y \mid X_1, X_2) = \beta_1 X_1 + \beta_2 X_2 = \begin{cases} \beta_1 & \textit{si ejercita} \\ \\ \beta_2 & \textit{si no ejercita} \end{cases}$$

Sin embargo, no la exploraremos ya que nuestra propuesta anterior es, no sólo satisfactoria sino también la más utilizada en el área.

Comparemos este modelo lineal con una sola regresora dicotómica con el test t para comparar las medias de dos poblaciones, a través de dos muestras independientes. Sean $W_1, \ldots, W_{n_1}$ variables aleatorias independientes idénticamente distribuidas con $E(W_i) = \mu_0$ e independientes de $Z_1, \ldots, Z_{n_2}$ que a su vez son variables aleatorias independientes entre sí e idénticamente distribuidas con $E(Z_i) = \mu_1$. El test t permite decidir entre las hipótesis

$$H_0 : \mu_0 = \mu_1$$
$$H_1 : \mu_0 \neq \mu_1$$

donde $\mu_0 = E(Y \mid X_1 = 0)$ es decir, la esperanza de la glucosa para las personas que no ejercitan y $\mu_1 = E(Y \mid X_1 = 1)$ la esperanza de la glucosa para las personas que sí lo hacen. Recordemos que este test presupone que las observaciones de cada población tienen distribución normal con las medias μ_0 y μ_1 respectivamente, y la misma varianza (aunque desconocida). Para el conjunto de datos `azucar`, la salida de correr el test t figura en la Tabla 27.

Tabla 27: Test t para dos muestras normales independientes, datos `azucar`.

```
> t.test(glucosa ~ ejercicio, var.equal = TRUE, data = azucar)

        Two Sample t-test

data:  glucosa by ejercicio
t = 6.309, df = 218, p-value = 1.545e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 5.107071 9.747588
sample estimates:
mean in group 0 mean in group 1
       98.91429        91.48696
```

Recordemos que el estadístico del test es

$$\sqrt{n_1 + n_2}\frac{\left(\overline{W}_{n_1} - \overline{Z}_{n_2}\right)}{S_p} \underset{\text{Bajo H}_0}{\sim} t_{n_1+n_2-2}$$

donde n_1 y n_2 son los tamaños de las muestras respectivas, y

$$S_p^2 = \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1}\left(W_i - \overline{W}_{n_1}\right)^2 + \sum_{j=1}^{n_2}\left(Z_j - \overline{Z}_{n_2}\right)^2\right]$$

es la varianza *pooleada* o combinada de ambas muestras. Por otra parte, para el modelo (26), el test de $H_0 : \beta_1 = 0$ es también un test t, observemos que tanto el estadístico calculado como el p-valor son los mismos.
Finalmente podemos concluir que en el caso en el que el modelo lineal tiene una sola variable explicativa categórica, realizar el test de si el coeficiente que la acompaña es estadísticamente significativo es equivalente a utilizar un test t de comparación de medias entre dos poblaciones normales independientes, con igual varianza.
Dos observaciones con respecto a la codificación de la variable binaria dada en (61):

- Comparemos el valor de β_0 estimado en la Tabla 26 (que es $\widehat{\beta}_0 = 98{,}9143$) con el promedio de la glucosa de las personas que no ejercitan (el grupo correspondiente a `ejercicio = 0`) calculado en la Tabla 27, que es 98,914, como anticipáramos. De igual modo, recuperamos el promedio de glucosa de las personas que ejercitan (91,487 en la Tabla 27) a partir de sumar $\widehat{\beta}_0 + \widehat{\beta}_1$ de la Tabla 26
$$\widehat{\beta}_0 + \widehat{\beta}_1 = 98{,}9143 - 7{,}4273 = 91{,}487.$$

- Codificando de esta forma, el promedio de la variable `ejercicio` da la proporción de personas que hacen ejercicio en la muestra, que son el 52,27 % de la muestra, como puede comprobarse en la Tabla 28 que tiene los estadísticos descriptivos de la variable `ejercicio`.

Tabla 28: Estadísticos descriptivos de la variable `ejercicio`.

```
> summary(ejercicio)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  1.0000  0.5227  1.0000  1.0000
```

Observación 4.12 *Una alternativa comúnmente utilizada para la codificación de las variables binarias es (1 = sí, 2 = no). Si definimos la variable X_3 con este código, el modelo es*

$$E\left(Y \mid X_3\right) = \beta_0 + \beta_3 X_3.$$

Luego la función de respuesta para las personas que ejercitan $(X_3 = 1)$ *es*

$$E(Y) = \beta_0 + \beta_3,$$

y para las que no ejercitan $(X_3 = 2)$ *es*

$$E(Y) = \beta_0 + 2\beta_3.$$

Nuevamente, la diferencia entre ambas medias es el coeficiente β *correspondiente, en este caso* β_3*. Luego el coeficiente* β_3 *conserva su interpretación como la diferencia en el nivel medio de glucosa entre grupos, pero ahora entre las personas que no hacen ejercicio, comparadas con aquellas que sí lo hacen, una manera menos intuitiva de pensarlo. De hecho,* β_0 *sólo no tiene una interpretación directa, y el valor promedio de la variable binaria no es igual a la proporción de observaciones de la muestra que caen en ninguno de los dos grupos. Observar que, sin embargo, en general el ajuste del modelo, es decir, los valores ajustados, los errores estándares, y los p-valorespara evaluar la diferencia de la glucosa en ambos grupos serán iguales con cualquier codificación.*

4.13.2. Un predictor binario y otro cuantitativo

Incorporemos al modelo una variable cuantitativa. Tomaremos el índice de masa corporal que se denomina `bmi` (*body mass index*, medido en kg/m^2) en la base de datos,

$$X_{i2} = \text{BMI de la persona iésima.}$$

El índice de masa corporal (BMI) es una medida de asociación entre el peso y la talla de un individuo ideada por el estadístico belga L. A. J. Quetelet, por lo que también se conoce como índice de Quetelet. Se calcula según la expresión matemática

$$BMI = \frac{peso}{estatura^2}$$

donde la masa o peso se expresa en kilogramos y la estatura en metros, luego la unidad de medida del BMI es kg/m^2. En el caso de los adultos se ha utilizado como uno de los recursos para evaluar su estado nutricional, de acuerdo con los valores propuestos por la Organización Mundial de la Salud: a grandes rasgos se divide en tres categorías: delgadez (si $BMI < 18{,}5$), peso normal (cuando $18{,}5 \leq BMI < 25$) y sobrepeso (si $BMI \geq 25$), con subclasificaciones que contemplan los casos de infrapeso u obesidad.
Luego el modelo de regresión lineal múltiple que proponemos es

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i.$$

O, si escribimos la función de respuesta (o sea, el modelo para la esperanza de Y) obtenemos

$$E(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \tag{64}$$

Interpretemos los parámetros. Para las personas que no hacen ejercicio ($X_1 = 0$) la función de respuesta es

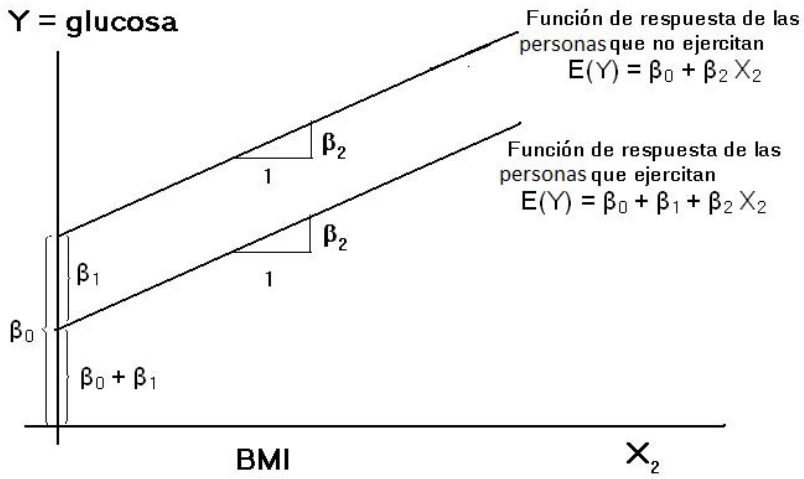
$$E(Y) = \beta_0 + \beta_1 0 + \beta_2 X_2 = \beta_0 + \beta_2 X_2 \qquad \text{no ejercita} \tag{65}$$

O sea, la función de respuesta para la glucosa media de las personas que no ejercitan es una línea recta con ordenada al origen β_0 y pendiente β_2.
Para las que sí hacen ejercicio ($X_1 = 1$) la función de respuesta (64) se convierte en

$$E(Y) = \beta_0 + \beta_1 1 + \beta_2 X_2 = (\beta_0 + \beta_1) + \beta_2 X_2 \qquad \text{ejercita} \tag{66}$$

Esta función también es una línea recta, con la misma pendiente β_2 pero con ordenada al origen $(\beta_0 + \beta_1)$. En la Figura 45 se grafican ambas funciones.

Figura 45: Significado de los coeficientes del modelo de regresión (64) con una variable indicadora X_1 de ejercicio y una variable continua $X_2 =$ `bmi` (datos `azucar`).



Enfoquémosnos en el significado de los coeficientes en la expresión (64) en el caso de las mediciones del nivel de glucosa. Vemos que el nivel medio de glucosa, $E(Y)$, es una función lineal del BMI (X_2) de la persona, con la misma pendiente β_2 para ambos tipos de personas. β_1 indica cuánto más baja (o más alta) es la función de respuesta para las personas que hacen ejercicio respecto de las que no lo hacen,

fijado el BMI. Luego β_1 mide el efecto diferencial por ejercitar. Como el ejercicio debiera reducir el nivel de glucosa, esperamos que β_1 sea menor que cero y que la recta de valores de glucosa esperados para personas que ejercitan (66) esté por debajo de las que no lo hacen (65). En general, β_1 muestra cuánto más baja (o más alta) se ubica la recta de respuesta media para la clase codificada por 1 respecto de la recta de la clase codificada por 0, para cualquier nivel fijo de X_2.

Tabla 29: Ajuste de la regresión para la variable `glucosa` con `ejercicio` y `bmi` como explicativas

```
> ajuste2<-lm(glucosa ~ ejercicio + bmi, data = azucar)
> summary(ajuste2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  84.4141     3.2336  26.105  < 2e-16 ***
ejercicio    -6.4879     1.1437  -5.673 4.46e-08 ***
bmi           0.5227     0.1128   4.633 6.20e-06 ***
---

Residual standard error: 8.339 on 217 degrees of freedom
Multiple R-squared:  0.2305,        Adjusted R-squared:  0.2234
F-statistic:  32.5 on 2 and 217 DF,  p-value: 4.491e-13
```

En la Tabla 29 figura el ajuste del modelo propuesto. La función de respuesta ajustada es

$$\widehat{Y} = 84{,}414 - 6{,}488X_1 + 0{,}523X_2.$$

Nos interesa medir el efecto de ejercitar (X_1) en el nivel de glucosa en sangre. Para ello buscamos un intervalo de confianza del 95 % para β_1. Necesitamos el percentil 0,975 de la t de Student con $n - 3 = 217$ grados de libertad. Como $t\,(0{,}975, 217) = 1{,}970956 \simeq 1{,}959964 = z_{0{,}975}$, los límites para el intervalo de confianza resultan ser

$$-6{,}488 \pm 1{,}971 \cdot 1{,}1437$$

o sea,

$$-6{,}488 - 1{,}971 \cdot 1{,}1437 \leq \beta_1 \leq -6{,}488 + 1{,}971 \cdot 1{,}1437$$
$$-8{,}742 \leq \beta_1 \leq -4.234$$

Podemos obtener este resultado directamente con `R`.

```
> confint(ajuste2)
                 2.5 %     97.5 %
(Intercept) 78.0408659 90.7873793
ejercicio   -8.7421142 -4.2336953
bmi          0.3003302  0.7449921
```

Luego, con el 95 por ciento de confianza concluimos que las personas que ejercitan tienen un nivel de glucosa entre 4,23 y 8,74 mg/dL, **más bajo** que las que no lo hacen, en promedio, para un cada nivel de `bmi` fijo. Un test formal de

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

con nivel de significatividad de 0,05 nos conduciría a rechazar H_0 y aceptar H_1, es decir, que el ejercicio tiene efecto cuando en el modelo incluimos el `bmi`, pues el intervalo de confianza del 95 % para β_1 no contiene al cero. Eso lo vemos también en la tabla de salida del paquete estadístico, en el p-valor de dicho coeficiente, que es $4{,}46 \cdot 10^{-8} < 0{,}05$.

Observación 4.13 *¿Por qué no ajustar dos regresiones lineales separadas (una para las personas que ejercitan y otra para las que no) en vez de hacer un ajuste con el total de datos? O sea, ajustar*

$$E\left(Y \mid X_2\right) = \beta_0^{(0)} + \beta_2^{(0)} X_2 \qquad \textit{no ejercitan} \tag{67}$$

para las que no ejercitan y

$$E\left(Y \mid X_2\right) = \beta_0^{(1)} + \beta_2^{(1)} X_2 \qquad \textit{ejercitan} \tag{68}$$

para las que ejercitan. Hay dos razones para esto.

- *El modelo (64) asume pendientes iguales en (67) y (68) y la misma varianza del error de para cada tipo de persona. En consecuencia, la pendiente común β_2 se puede estimar mejor usando la información en la muestra conjunta. Ojo, este modelo **no** debería usarse si no se cree que este supuesto sea correcto para los datos a analizar.*

- *Usando el modelo (64) otras inferencias, como por ejemplo las realizadas sobre β_0 y β_1 resultarán más precisas pues se dispone de más observaciones para estimarlos y estimar a σ^2 (lo que se traduce en más grados de libertad en el MSRes). De todos modos, en este ejemplo donde hay doscientas observaciones, tenemos grados de libertad suficientes para proponer dos modelos si creyéramos que el modelo (64) no describe bien a los datos.*

Observación 4.14 *Los modelos de regresión múltiple en los que todas las variables explicativas son cualitativas se suelen denominar* ***modelos de análisis de la varianza (ANOVA)****. Los modelos que contienen algunas variables explicativas cuantitativas y otras variables explicativas cualitativas, para los que la variable explicativa de interés principal es cualitativa (por ejemplo, tipo de tratamiento que recibe el paciente) y las variables cuantitativas se introducen primariamente para reducir la varianza de los términos del error, se suelen denominar* ***modelos de análisis de la covarianza (ANCOVA)****.*

4.14. Predictores Cualitativos con más de dos clases

4.14.1. Una sola predictora cualitativa con más de dos clases

Otra de las acciones que suelen recomendarse para evitar que las personas con niveles de glucosa alto entren en el diagnóstico de diabetes es bajar el 7 % de su peso, o más. Para los datos del archivo `azucar` se registró la variable `peso.evo` que es una variable categórica que mide la evolución del peso del paciente en el último año. Tiene tres categorías: "*bajó de peso*", si su peso disminuyó un 7 % o más respecto del control médico anual anterior, "*su peso se mantuvo igual*", cuando la diferencia entre ambos pesos difiere en menos de un 7 % respecto del peso anterior, o bien "*aumentó de peso*", si su peso actual aumentó un 7 % o más respecto del registrado en el control médico anual anterior. La evolución del peso fue codificada en orden de 1 a 3, según las categorías recién definidas. Este es un ejemplo de una variable *ordinal* (con valores o categorías cuyo orden relativo es relevante, pero separados por incrementos que pueden no estar reflejados en forma precisa en la codificación numérica asignada). Las categorías de la variable `peso.evo` figuran en la Tabla 30.

Tabla 30: Niveles de la variable `peso.evo`, que codifica la evolución del peso en el último año.

Categorías de peso.evo	codificación original
Disminuyó de peso en el último año (7 % o más)	1
Pesa igual que hace un año (menos de 7 % de diferencia)	2
Aumentó de peso en el último año (7 % o más)	3

Las variables categóricas de más de dos niveles también puede ser *nominales*, en el sentido que no haya un orden intrínseco en las categorías. Etnia, estado civil, ocupación y región geográfica son ejemplos de variables nominales. Con las variables nominales es aún más claro que la codificación numérica usada habitualmente

para representar a la variable en la base de datos no puede ser tratada como los valores de una variable numérica como nivel de glucosa en sangre.
Las categorías se suelen crear para ser mutuamente excluyentes y exhaustivas, por lo que que cada miembro de la población se encuentra en una y sólo una categoría. En este sentido, tanto las categorías ordinales como las nominales definen subgrupos de la población.
Es secillo acomodar ambos tipos de variables tanto en la regresión lineal múltiple como en otros modelos de regresión, usando variables indicadoras o *dummies*. Como en las variables binarias, donde dos categorías se representan en el modelo con una sola variable indicadora, las variables categóricas con $K \geq 2$ niveles se representan por $K - 1$ indicadoras, una para cada nivel de la variable, excepto el nivel de referencia o basal. Supongamos que elegimos el nivel 1 como nivel de referencia. Entonces para $k = 2, 3, ..., K$, la k-ésima variable indicadora toma el valor 1 para las observaciones que pertenecen a la categoría k, y 0 para las observaciones que pertenecen a cualquier otra categoría. Observemos que para $K = 2$ esto también describe el caso binario, en el cual la respuesta "*no*" define el nivel basal o de referencia y la variable indicadora toma el valor 1 sólo para el grupo "*sí*".
Traduzcamos todo al ejemplo. Como la variable ordinal `peso.evo` tiene 3 categorías, necesitamos definir 2 variables dummies. Las llamamos `Ievo2` e `Ievo3`. En la Tabla 31, observamos los valores para las dos variables indicadoras correspondientes a la variable categórica `peso.evo`. Cada nivel de `peso.evo` queda definidio por una combinación única de las dos variables indicadoras.

Tabla 31: Codificación de las variables indicadoras para una variable categórica multinivel

	Variables indicadoras		
peso.evo	Ievo2	Ievo3	Categoría
1	0	0	bajó de peso
2	1	0	mantuvo su peso
3	0	1	aumentó de peso

Por el momento consideremos un modelo simple en el cual los tres niveles de `peso.evo` sean los únicos predictores. Entonces

$$E\left(Y \mid X\right) = \beta_0 + \beta_2 \texttt{Ievo2} + \beta_3 \texttt{Ievo3} \tag{69}$$

donde X representa las dos variables dummies recién definidas, es decir,

$$X = \left(\texttt{Ievo2}, \texttt{Ievo3}\right).$$

Para tener mayor claridad, en (69) hemos indexado a los $\beta's$ en concordancia con los niveles de `peso.evo`, de modo que β_1 no aparece en el modelo. Si dejamos que las dos indicadoras tomen el valor 0 ó 1 de manera de definir los tres (¿por qué no cuatro?) niveles de `peso.evo`, obtenemos

$$E(Y \mid X) = \begin{cases} \beta_0 & \text{si } \texttt{peso.evo} = 1, \text{ o sea } \texttt{Ievo2} = 0 \text{ e } \texttt{Ievo3} = 0 \\ \beta_0 + \beta_2 & \text{si } \texttt{peso.evo} = 2, \text{ o sea } \texttt{Ievo2} = 1 \text{ e } \texttt{Ievo3} = 0 \\ \beta_0 + \beta_3 & \text{si } \texttt{peso.evo} = 3, \text{ o sea } \texttt{Ievo2} = 0 \text{ e } \texttt{Ievo3} = 1 \end{cases} \quad (70)$$

De (70) es claro que β_0, la ordenada al origen, da el valor de $E(Y \mid X)$ en el grupo de referencia, el grupo "bajó de peso", o `peso.evo` $= 1$. Entonces es sólo cuestión de restarle a la segunda línea la primera línea para ver que β_2 da la diferencia en el promedio de glucosa en el grupo "mantuvo su peso" (`peso.evo` $= 2$) comparado con el grupo "bajó de peso". De acuerdo con esto, el test de $H_0 : \beta_2 = 0$ es un test para chequear si los niveles medios de glucosa son los mismos en los dos grupos "bajó de peso" y "mantuvo su peso" (`peso.evo` $= 1$ y 2). Y de manera similar para β_3.
Podemos hacer unas cuantas observaciones a partir de (70).

- Sin otros predictores, o covariables, el modelo es equivalente a un ANOVA de un factor (one-way ANOVA). También se dice que el modelo está *saturado* (es decir, no impone estructura alguna a las medias poblacionales) y las medias de cada grupo de la población se estimarán bajo el modelo (70) por el promedio de las muestras correspondientes. Con covariables, las medias estimadas para cada grupo se ajustarán a las diferencias entre grupos en las covariables incluidas en el modelo.

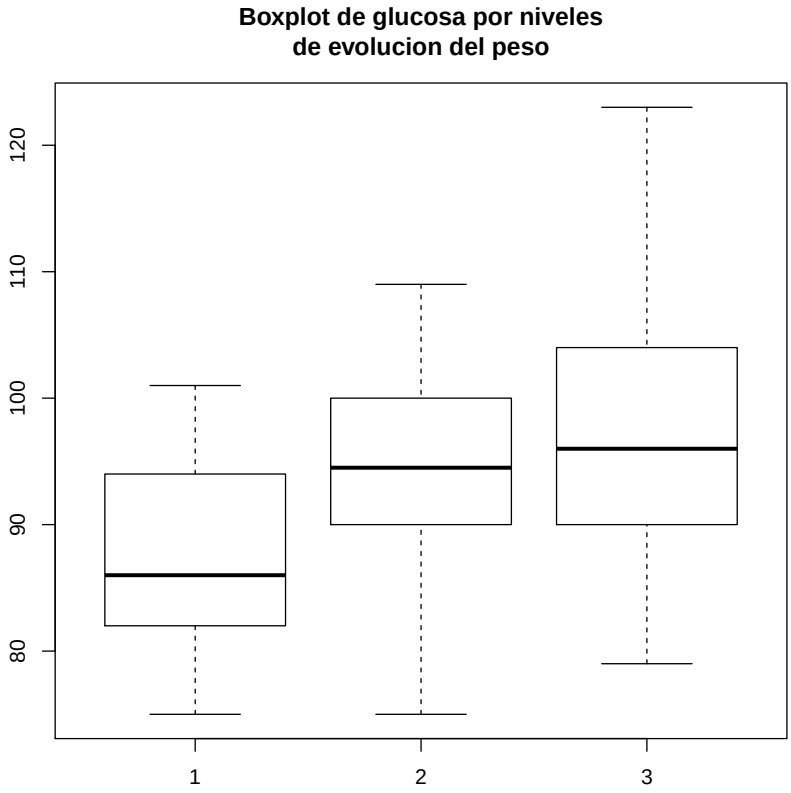
- Los parámetros del modelo (y por lo tanto las dummies que los acompañan) pueden ser definidos para que sean iguales a la media poblacional de cada grupo o, sino, para que sean las diferencias entre las medias poblacionales de dos grupos distintos, como en (70). Por ejemplo, la diferencia en los niveles medios de la variable Y entre los grupos "aumentó de peso" (`peso.evo = 3`) y "mantuvo su peso" (`peso.evo = 2`) está dada por $\beta_3 - \beta_2$ (chequearlo). Todos los paquetes estadísticos permiten calcular de manera directa estimadores y tests de hipótesis acerca de estos *contrastes lineales.* Esto implica que la elección del grupo de referencia es, en algun sentido, arbitraria. Mientras que alguna elección en particular puede ser la mejor para facilitar la presentación, posiblemente porque los contrastes con el grupo de referencia seleccionado sean los de mayor interés, cuando se toman grupos de referencia alternativos, esencialmente se está definiendo el mismo modelo.

Tabla 32: Ajuste de regresión lineal múltiple para explicar a la variable `glucosa` con la evolución del peso como categórica, `Ievo` (datos de la base `azucar`). El R produce las dos binarias de forma automática (`Ievo2` e `Ievo3`).

```
> Ievo<-factor(peso.evo)  #convierte a la variable en factor
> contrasts(Ievo)        #da la codificacion
  2 3
1 0 0
2 1 0
3 0 1
> ajuste3<-lm(glucosa ~  Ievo)
> summary(ajuste3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   88.273      1.564  56.449  < 2e-16 ***
Ievo2          6.283      1.888   3.327  0.00103 **
Ievo3          8.997      1.774   5.072 8.45e-07 ***
---
Residual standard error: 8.983 on 217 degrees of freedom
Multiple R-squared:  0.1071,        Adjusted R-squared:  0.09884
F-statistic: 13.01 on 2 and 217 DF,  p-value: 4.607e-06
```

La Tabla 32 muestra los resultados para el modelo con `peso.evo` tratada como una variable categórica, utilizando de nuevo los datos del archivo `azucar`. La estimación de $\widehat{\beta}_0$ es 88,273 mg / dL, esta es la estimación del nivel de glucosa medio para el grupo que bajó de peso (grupo de referencia). Las diferencias entre los niveles de glucosa del grupo de referencia y los otros dos grupos (de distinta evolución del peso) resultan ser estadísticamente significativas; como todas dan positivas indican que la glucosa estaría relacionada con la evolución del peso. Por ejemplo, el nivel promedio de glucosa en el grupo "subió de peso" (`Ievo3`) es 8,997 mg / dL mayor que la del grupo "bajó de peso" (`peso.evo = 1`) ($t = 5{,}072$, p-valor $= 8{,}45 \cdot 10^{-7}$). En la Figura 46 vemos un boxplot de los datos de glucosa separados según sus niveles de `peso.evo`, donde se aprecia esta diferencia.
Es de interés testear si la variable `peso.evo` sirve para explicar al nivel de glucosa. Para evaluarla en su conjunto se utiliza el test F que describiremos en la Sección 4.14.4. Antes de hacerlo discutamos otra manera de introducir a la variable `peso.evo` en el modelo.

Figura 46: Boxplot de los datos de `glucosa`, según sus niveles de `peso.evo`.



4.14.2. Variables indicadoras versus variables numéricas

Una alternativa al uso de variables indicadoras de una variable de predicción cualitativa es pensarla como numérica. En el ejemplo de la glucosa, podríamos utilizar una única variable predictora Z y asignar valores 1,2 y 3 a las clases, como se describe en la Tabla 33.
Los valores numéricos son, por supuesto, arbitrarios y podrían ser cualquier otro conjunto de números. El modelo en este caso sería

$$Y_i = \beta_0 + \beta_1 Z_i + \varepsilon_i \tag{71}$$

La principal dificultad en tomar a las variables categóricas como numéricas es que la numeración otorgada a las categorías distintas define una métrica (una distancia) entre las clases de la variable cualitativa que puede no resultar razonable para

Tabla 33: Variable categórica mirada como numérica

peso.evo	Z
Bajó de peso	1
Mantuvo su peso	2
Aumentó de peso	3

modelizar. Veámoslo en el ejemplo. Escribimos la función de respuesta media con el modelo (71) para las tres clases de la variable cualitativa

$$E\left(Y \mid Z\right) = \begin{cases} \beta_0 + \beta_1 & \text{si } \texttt{peso.evo} = 1 \\ \beta_0 + 2\beta_1 & \text{si } \texttt{peso.evo} = 2 \\ \beta_0 + 3\beta_1 & \text{si } \texttt{peso.evo} = 3 \end{cases}$$

Notemos la implicación clave de este modelo:

$$\begin{aligned} & E\left(Y \mid \texttt{peso.evo} = 2\right) - E\left(Y \mid \texttt{peso.evo} = 1\right) \\ & = E\left(Y \mid \texttt{peso.evo} = 3\right) - E\left(Y \mid \texttt{peso.evo} = 2\right) \\ & = \beta_1 \end{aligned}$$

Luego, la codificación 1 a 5 implica que pensamos que la respuesta media cambia **en la misma cantidad** cuando pasamos de `peso.evo=1` a `peso.evo=2` o de `peso.evo=2` a `peso.evo=3`. Esto puede no estar en coincidencia con la realidad y resulta de la codificación 1 a 3 que asigna igual distancia entre los 3 tipos de evolución del peso. Por supuesto, con distintas codificaciones podemos imponer espaciamientos diferentes entre las clases de la variable cualitativa pero esto sería siempre arbitrario.

En contraposición, el uso de variables indicadoras no hace supuestos sobre el espaciamiento de las clases y descansa en los datos para mostrar los efectos diferentes que ocurren. En el caso del modelo (70) no se impone ningun patrón o vínculo entre sí a las cinco medias de los grupos definidos por la variable categórica, tanto en el modelo sin covariables como si las tuviera. Aquí β_2 da la diferencia en el promedio de glucosa en el grupo `peso.evo=2` comparado con el grupo `peso.evo=1`, y β_3 da la diferencia en el promedio de glucosa en el grupo `peso.evo=3` comparado con el grupo `peso.evo=1` y $\beta_3 - \beta_2$ da la diferencia en el promedio de glucosa en el grupo `peso.evo=3` comparado con el grupo `peso.evo=2`. Observemos que no hay restricciones arbitrarias que deban cumplir estos tres efectos. En cambio, si la variable `peso.evo` fuera tratada como una variable numérica que toma valores de 1 a 3, las esperanzas poblacionales de cada grupo se verían obligadas a yacer en una línea recta. En síntesis: para variables categóricas es preferible usar la codificación que proporcionan las variables dummies.

4.14.3. Variables numéricas como categóricas

Algunas veces, aún cuando las variables son originalmente cuantitativas se las puede incluir en un modelo como categóricas. Por ejemplo, la variable cuantitativa edad se puede transformar agrupando las edades en las categorías: menor de 21, 21 a 34, 35 a 49, etc. En este caso, se usan variables indicadoras o dummies para las clases de este nuevo predictor. A primera vista, este enfoque parece cuestionable, ya que la información sobre la edad real se pierde. Además, se ponen parámetros adicionales en el modelo, lo que conduce a una reducción de los grados de libertad asociados con el MSRes.
Sin embargo, hay ocasiones en las que la sustitución de una variable cuantitativa por indicadoras puede ser apropiado. Por ejemplo, cuando se piensa que la relación entre la respuesta y la explicativa puede no ser lineal (en el caso en que la glucosa aumentara tanto para personas muy jóvenes o muy grandes) o en una encuesta a gran escala, donde la pérdida de 10 ó 20 grados de libertad es irrelevante. En una primera etapa exploratoria, donde se está muy en duda acerca de la forma de la función de regresión, puede ajustarse un modelo como (70) en una primera etapa exploratoria, y luego, en virtud de lo observado, incluir a la variable (o una transformación de ella) como numérica.
Esto es de hecho lo que se hizo con la variable evolución del peso: a partir de dos variables numéricas medidas sobre el mismo paciente (el peso en un año y el peso en el siguiente) se construyó una variable categórica.

4.14.4. El test F

A pesar de que todos los contrastes entre los niveles de una variable explicativa categórica están disponibles para ser estimados y comparados luego de ajustar un modelo de regresión, los test t para estas comparaciones múltiples en general no proporcionan una evaluación conjunta de la importancia de la variable categórica para predecir a la variable respuesta, o más precisamente no permiten realizar un único test de la hipótesis nula de que el nivel medio de la variable respuesta es el mismo para todos los niveles de este predictor. En el ejemplo, esto es equivalente a un test de si alguno de los dos coeficientes correspondientes a `Ievo2` o `Ievo2` difieren de cero. El resultado que aparece en la Tabla 32 ($F_{\text{obs}} = 13{,}01$, con 2 grados de libertad en el numerador y 217 en el denominador, p-valor $= 4{,}6 \cdot 10^{-6} < 0{,}05$) muestra que los niveles medios de glucosa son claramente diferentes entre los grupos definidos por `peso.evo`. Las hipótesis que chequea este test en este caso son

$$\begin{aligned} &H_0 : \beta_2 = \beta_3 = 0 \\ &H_1 : \text{ al menos uno de los } \beta_i \text{ con } i \text{ entre 2 y 3 es tal que } \beta_i \neq 0 \end{aligned} \tag{72}$$

En este caso se rechaza la hipótesis nula (p-valor $= 4{,}6 \cdot 10^{-6} < 0{,}05$) y se concluye que no todos los β_i con i entre 2 y 3 son simultáneamente iguales a cero. Luego la evolución del peso es útil para predecir el nivel de glucosa. En general este resultado puede leerse en la tabla de ANOVA del ajuste.
Es por este motivo que conviene ingresar en la base de datos a la variable `peso.evo` con sus tres niveles y pedirle al software que compute las dos variables dicotómicas, en vez de ponerlas a mano en el archivo, pues en tal caso no hay cómo decirle al paquete que las dos variables están vinculadas de esta forma.

4.14.5. Comparaciones Múltiples

Una vez que, a través del test F, logramos concluir que la variable categórica es significativa para explicar a la respuesta, aparece el interés de comparar la media de la variable respuesta para los grupos definidos por los distintos niveles de la variable categórica. Es decir, interesa comparar las medias de dos grupos, digamos μ_j y μ_k. Esto suele hacerse estudiando si la diferencia entre ellos $\mu_j - \mu_k$ es distinta de cero. Tal diferencia entre los niveles medios de una variable categórica (o factor) se denomina una comparación de a pares. Cuando una variable categórica toma K niveles, el número de comparaciones de a pares que pueden hacerse es $\binom{K}{2} = \frac{K(K-1)}{2}$.
La salida que da el `summary` del `lm` en `R` que analizamos antes, por ejemplo en la Tabla 32, o la salida estándar que proporciona cualquier paquete estadístico a un ajuste lineal, tiene, en este sentido dos limitaciones importantes.

1. Los p-valores que aparecen en la columna de la derecha son válidos para cada comparación individual.

2. Cuando la variable categórica tiene más de dos niveles, dicha tabla no nos da información de todas las comparaciones de a pares de forma directa. En el ejemplo de `azucar`, vemos en la Tabla 32 que nos falta la comparación entre los niveles 2 y 3 de la variable evolución de peso.

Con la primera limitación veníamos trabajando desde el modelo lineal simple, pero en el caso de regresión con covariables categóricas se hace particularmente seria por la gran cantidad de comparaciones que tienen interés para el experimentador. Cuando se realizan varios tests con los mismos datos, tanto el nivel de significatividad como la potencia de las conclusiones acerca de la familia de tests se ve afectada. Consideremos por ejemplo, la realización de tres tests de t, cada uno a nivel $\alpha = 0{,}05$, para testear las hipótesis

$$\begin{aligned} H_0^{(1)} &: \mu_2 - \mu_1 = 0 \text{ versus } H_1^{(1)} : \mu_2 - \mu_1 \neq 0 \\ H_0^{(2)} &: \mu_3 - \mu_1 = 0 \text{ versus } H_1^{(2)} : \mu_3 - \mu_1 \neq 0 \\ H_0^{(3)} &: \mu_3 - \mu_2 = 0 \text{ versus } H_1^{(3)} : \mu_3 - \mu_2 \neq 0 \end{aligned}$$

La probabilidad de que los tres tests concluyan que las tres hipótesis nulas H_0 son verdaderas cuando en realidad las tres H_0 son verdaderas, asumiendo independencia de los tests, será

$$0{,}95^3 = 0{,}857.$$

Luego, la probabilidad de concluir H_1 para al menos una de las tres comparaciones es $1 - 0{,}857 = 0{,}143$ en vez de 0,05. Vemos que el nivel de significatividad de una familia de tests no es el mismo que para un test individual. Lo mismo pasa para los intervalo de confianza.
El objetivo de hacer estas comparaciones múltiples de manera justa es mantener el error de tipo I acotado, sin inflarlo por sacar muchas conclusiones con el mismo conjunto de datos. Es decir, queremos un test de nivel 0,05 para las hipótesis

$$H_0 : \begin{cases} \mu_2 - \mu_1 = 0 \\ \mu_3 - \mu_1 = 0 \\ \mu_3 - \mu_2 = 0 \end{cases} \quad \text{versus } H_1 : \text{alguna de las 3 igualdades no vale.}$$

Con la notación de regresión lineal múltiple, podemos reescribir a la hipótesis nula del siguiente modo

$$H_0 : \begin{cases} E\left(Y \mid \texttt{peso.evo} = 2\right) - E\left(Y \mid \texttt{peso.evo} = 1\right) = 0 \\ E\left(Y \mid \texttt{peso.evo} = 3\right) - E\left(Y \mid \texttt{peso.evo} = 1\right) = 0 \\ E\left(Y \mid \texttt{peso.evo} = 3\right) - E\left(Y \mid \texttt{peso.evo} = 2\right) = 0 \end{cases}$$

Para eso, primero hay que mirar el resultado del test conjunto F que evalúa la significatividad conjunta de la variable categórica para explicar a la respuesta. Si este test no resulta significativo, suele descartarse la variable categórica de entre las covariables de interés, y se la excluye del modelo. Si este test resulta estadísticamente significativo, entonces suelen mirarse con más detalle cuáles de las comparaciones entre grupos son estadísticamente significativas, para proporcionar un mejor análisis de los datos en consideración. Hay diversas propuestas para llevar estas comparaciones a cabo, de acuerdo esencialmente a cuáles son las comparaciones que resultan más interesantes al experimentador.
¿Qué pasa si alguna de las comparaciones llevadas a cabo resulta no significativa? ¿Conviene redefinir las categorías? La recomendación general es que si esto pasara, de todos modos conviene mantener las categorías originales puesto que de esta forma se estimará mejor a la varianza σ^2 de los errores (resultará menor), y por lo tanto las conclusiones que se obtendrán serán más potentes. Además, cuando el interés esté puesto en la conclusión respecto de una comparación entre dos categorías en particular, el hecho de mantener las grupos originales permitirá clasificar bien a cada observación permitiendo mantener clara la diferencia que se está buscando establecer. No conviene recodificar a posteriori del análisis, es mejor dejar

todos los niveles de la variable categórica en el modelo, aunque algunos resulten no significativos.
Hay distintos métodos disponibles para controlar la tasa de error de tipo I. La comparación múltiple propuesta por Tukey, la *diferencia honestamente significativa* de Tukey (HSD: *honestly significant difference*) es un procedimiento que permite hacer todas las comparaciones con nivel conjunto prefijado. El método de Scheffé también provee un procedimiento que asegura el nivel de todas las comparaciones posibles, incluso asegura el nivel de cualquier combinación lineal de los parámetros (no necesariamente una resta) que pueda interesar. Si sólo interesan comparaciones de a pares, Tukey proporciona intervalos más angostos, y su método es preferible. Los intervalos sólo difieren en el percentil utilizado. En las Tablas 34 y 35 aparecen las salidas de las comparaciones de Tukey para los datos de `azucar` realizado en `R` utilizando dos comandos diferentes para hacerlo: `TukeyHSD` y `glht`, este último del paquete `multcomp`. En ella vemos que el nivel medio de glucosa del grupo "bajó de peso" difiere significativamente tanto del grupo "mantuvo su peso igual" como del grupo "aumentó de peso", así como también vemos que el nivel medio de glucosa de estos dos últimos grupos no difiere (significativamente) entre sí.
En la Sección 4.9.3 presentamos el procedimiento de comparaciones múltiples de Bonferroni. También es aplicable para el contexto de ANOVA, ya sea que interesen las comparaciones de a pares, o combinaciones lineales de los coeficientes mientras estos se hayan fijado **con anterioridad** a hacer el análisis de datos. Otros métodos que pueden aplicarse cuando las comparaciones que interesan tienen características específicas son la *mínima diferencia significativa* de Fisher (LSD: *least significative difference*), el método de Sidak o el procedimiento de Dunnett. Puede verse Seber y Lee [1977] o Kutner et al. [2005] para más detalle sobre las comparaciones múltiples.

4.15. Una predictora cualitativa y una numérica

Ajustemos ahora un modelo de regresión lineal múltiple con una covariable numérica y una categórica. Siguiendo con los datos de la glucosa, proponemos ajustar un modelo donde aparezcan `peso.evo` y `bmi` como variables explicativas, donde la primera es categórica (como ya vimos, la incluimos en el modelo como las 2 dummies definidas por `Ievo`) y la segunda es continua. Proponemos ajustar el siguiente modelo

$$E\left(Y \mid X\right) = \beta_0 + \beta_2 \texttt{Ievo2} + \beta_3 \texttt{Ievo3} + \beta_{BMI} \texttt{bmi} \tag{73}$$

En este caso, $X = (\texttt{Ievo2}, \texttt{Ievo3}, \texttt{bmi})$. Para entender este modelo, nuevamente dejamos que las indicadoras tomen el valor 0 ó 1 de manera de definir los tres

Tabla 34: Intervalos de confianza de nivel simultáneo para la variable `glucosa` para los distintos niveles de evolución del peso, `Ievo` (datos de la base `azucar`), usando el comando `TukeyHSD` del `R`.

```
> tu<-TukeyHSD(aov(glucosa ~ Ievo),"Ievo")
> tu
  Tukey multiple comparisons of means
    95% family-wise confidence level

$Ievo
        diff        lwr       upr     p adj
2-1 6.282828  1.8263489 10.739308 0.0029505
3-1 8.996838  4.8103925 13.183283 0.0000025
3-2 2.714010 -0.4718461  5.899865 0.1121165
```

niveles de `peso.evo`, y obtenemos

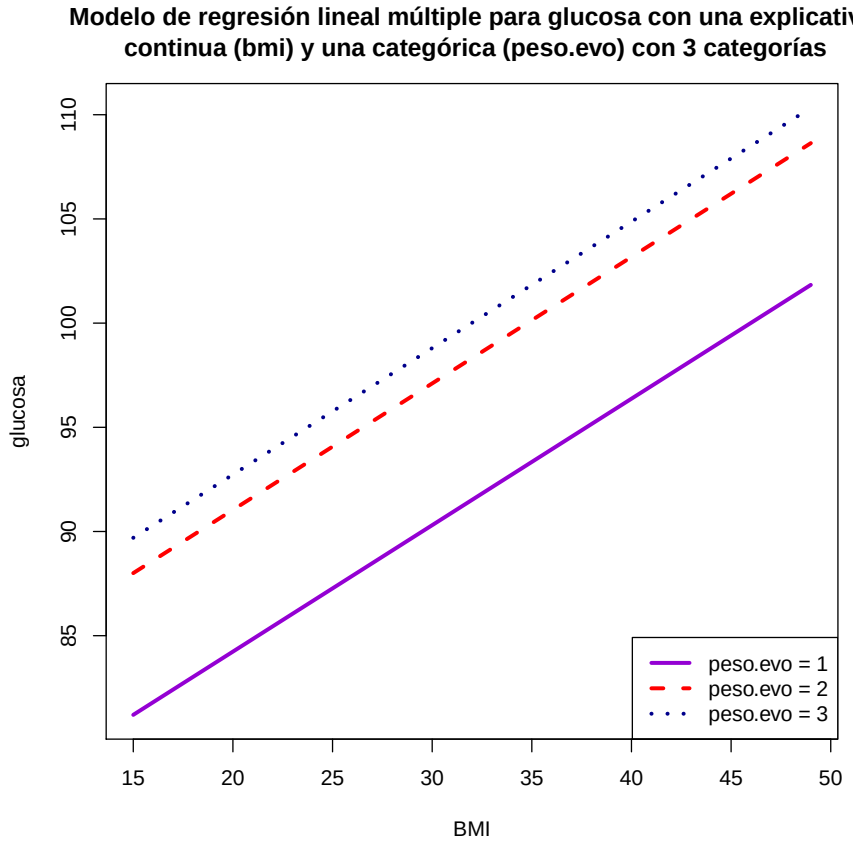
$$E\left(Y \mid X\right) = \begin{cases} \beta_0 + \beta_{BMI}\texttt{bmi} & \text{si } \texttt{peso.evo} = 1 \\ \beta_0 + \beta_2 + \beta_{BMI}\texttt{bmi} & \text{si } \texttt{peso.evo} = 2 \\ \beta_0 + \beta_3 + \beta_{BMI}\texttt{bmi} & \text{si } \texttt{peso.evo} = 3 \end{cases}$$

es decir, que este modelo propone ajustar una recta distinta para la glucosa media de cada grupo, **todas con igual pendiente** que en este caso hemos denominado β_{BMI}, y tres ordenadas al origen diferentes, una por cada grupo. Como vemos, estamos ajustando tres rectas paralelas. Acá β_2 indica cuánto aumenta (o disminuye, dependiendo del signo) el valor medio de glucosa para las personas cuyo nivel de evolución del peso es 2 (las personas "que mantuvieron su peso") respecto de aquellas cuyo nivel de evolución del peso es 1 (las personas que "bajaron de peso"). En la Figura 47 puede verse el gráfico que proponemos para el valor esperado de la glucosa en función de la evolución del peso y del BMI. Como esperamos que a medida que la evolución del peso aumente (o sea, a medida que el paciente aumente de peso) el nivel de glucosa aumente, hemos acomodado las rectas de manera que vayan aumentando al aumentar la variable que codifica esta evolución. Así mismo, es de esperar que a mayor BMI aumente el nivel de glucosa, por eso en el dibujo proponemos una pendiente (común a todos los grupos) positiva, como ya vimos que pasaba en el ajuste anterior.
La Tabla 36 exhibe el modelo ajustado.
En este caso vemos que cuando incorporamos la variable BMI al modelo, todos los coeficientes asociados a la variable `peso.evo` siguen siendo significativos. El test de, por ejemplo, $H_0 : \beta_2 = 0$ da significativo ($t = 3{,}82$, p-valor $= 0{,}000177$)

Tabla 35: Tests e intervalos de confianza de nivel simultáneo para la variable `glucosa` para los distintos niveles de evolución del peso, `Ievo` (datos de la base `azucar`), usando el comando `glht` de la librería `multcomp` de `R`.

```
> ajuste3<-lm(glucosa ~  Ievo)
> library(multcomp)  #para testear combinaciones de los parametros
> tu.otro <- glht(ajuste3, linfct = mcp(Ievo = "Tukey"))
> summary(tu.otro)
         Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: lm(formula = glucosa ~ Ievo)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0    6.283      1.888   3.327  0.00286 **
3 - 1 == 0    8.997      1.774   5.072  < 1e-04 ***
3 - 2 == 0    2.714      1.350   2.010  0.10977
---
> confint(tu.otro)
         Simultaneous Confidence Intervals
Multiple Comparisons of Means: Tukey Contrasts
Quantile = 2.35
95% family-wise confidence level
Linear Hypotheses:
           Estimate lwr     upr
2 - 1 == 0  6.2828   1.8450 10.7206
3 - 1 == 0  8.9968   4.8280 13.1657
3 - 2 == 0  2.7140  -0.4585  5.8865
```

Figura 47: Modelo propuesto para explicar la `glucosa` con una covariable explicativa categórica (`peso.evo`) con tres niveles y otra continua (`bmi`).



indicando que hay diferencia significativa en los niveles medios de glucosa para personas que no bajaron de peso con respecto a las que sí bajaron (grupo basal). Lo mismo sucede al testear la comparación entre la glucosa esperada del grupo que aumentó de peso y el que bajó de peso, cuando en el modelo se ajusta por BMI ($t = 5{,}074$, p-valor $= 8{,}38 \cdot 10^{-7}$). Es decir que los niveles medios de glucosa en los distintos grupos definidos por la evolución del peso difieren del basal. Además, como sus coeficientes estimados crecen al aumentar el peso, vemos que los valores estimados son consistentes con lo que bosquejamos a priori en la Figura 47. Antes de comparar los niveles medios de los distintos grupos entre sí observemos que si queremos evaluar a la variable `peso.evo` en su conjunto, debemos recurrir a un test F que evalue las hipótesis (72), cuando además en el modelo aparece BMI

Tabla 36: Regresión de glucosa en las regresoras: `peso.evo` (categórica) y `bmi` (numérica).

```
> ajuste4<-lm(glucosa ~ Ievo + bmi)
> summary(ajuste4)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  72.0993     3.3764  21.354  < 2e-16 ***
Ievo2         6.8023     1.7823   3.817 0.000177 ***
Ievo3         8.4963     1.6744   5.074 8.38e-07 ***
bmi           0.6068     0.1140   5.324 2.54e-07 ***
---

Residual standard error: 8.466 on 216 degrees of freedom
Multiple R-squared:  0.2107,        Adjusted R-squared:  0.1997
F-statistic: 19.21 on 3 and 216 DF,  p-value: 4.411e-11
```

como explicativa. A presentarlo nos abocamos en la siguiente sección.

4.15.1. Test F para testear si varios parámetros son cero, y tabla de ANOVA para comparar modelos

En forma análoga a la descripta en la Sección 4.8.1, pueden usarse las sumas de cuadrados para comparar el ajuste proporcionado por dos modelos lineales distintos. Esto puede hacerse de manera general, para diversos modelos. Lo describiremos con cierto detalle para la situación que nos interesa ahora. En el caso de los datos de `azucar` queremos testear si la variable categórica que describe la actividad física es significativa para explicar el nivel de glucosa **cuando en el modelo tenemos a BMI como explicativa**. Es decir, para el modelo (64)

$$E\left(Y \mid X\right) = \beta_0 + \beta_2 \texttt{Ievo2} + \beta_3 \texttt{Ievo3} + \beta_{BMI} \texttt{bmi}$$

queremos testear las hipótesis

$$\begin{aligned} &H_0 : \beta_2 = \beta_3 = 0 \\ &H_1 : \text{ al menos uno de los } \beta_i \text{ con } i \text{ entre 2 y 3 es tal que } \beta_i \neq 0 \end{aligned} \tag{74}$$

Para ello, ajustamos dos modelos lineales a los datos y usaremos la suma de cuadrados propuesta en (50) como medida de cuan bueno es cada ajuste, es decir, calcularemos y compararemos las

$$\Delta_{\text{modelo}} = \sum (\text{observados} - \text{modelo})^2$$

para cada uno de dos modelos. En este caso el modelo básico será el que vale si H_0 es verdadera, el modelo lineal simple que tiene a BMI como única explicativa del nivel medio de glucosa:

$$Y_i = \beta_0^{\text{básico}} + \beta_{BMI}^{\text{básico}} \texttt{bmi}_i \ + \varepsilon_i.$$

Para este modelo se calculan las estimaciones de los parámetros $\widehat{\beta}_0^{\text{básico}}$ y $\widehat{\beta}_{BMI}^{\text{básico}}$, y con ellos los predichos

$$\widehat{Y}_i^{\text{básico}} = \widehat{\beta}_0^{\text{básico}} + \widehat{\beta}_{BMI}^{\text{básico}} \texttt{bmi}_i$$

y la suma de cuadrados que mide el desajuste

$$\Delta_{\text{modelo básico}} = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i^{\text{básico}} \right)^2.$$

El modelo más complejo será el que figura en (64), es decir

$$Y_i = \beta_0^{\text{comp}} + \beta_2^{\text{comp}} \texttt{Ievo2}_i + \beta_3^{\text{comp}} \texttt{Ievo3}_i + \beta_{BMI}^{\text{comp}} \texttt{bmi}_i \ + \varepsilon_i.$$

Nuevamente se estiman los parámetros bajo este modelo obteniéndose $\widehat{\beta}_0^{\text{comp}}$, $\widehat{\beta}_2^{\text{comp}}$, $\widehat{\beta}_3^{\text{comp}}$ y $\widehat{\beta}_{BMI}^{\text{comp}}$, con ellos se calculan los predichos para este modelo

$$\widehat{Y}_i^{\text{comp}} = \widehat{\beta}_0^{\text{comp}} + \widehat{\beta}_2^{\text{comp}} \texttt{Ievo2}_i + \widehat{\beta}_3^{\text{comp}} \texttt{Ievo3}_i + \widehat{\beta}_{BMI}^{\text{comp}} \texttt{bmi}_i$$

y la suma de cuadrados que mide el desajuste que tienen los datos a este modelo complejo

$$\Delta_{\text{modelo complejo}} = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i^{\text{comp}} \right)^2.$$

Por supuesto, como el modelo complejo tiene al modelo básico como caso particular, resulta que el ajuste del modelo complejo a los datos será siempre tan satisfactorio como el del modelo básico o más satisfactorio aún, de modo que $\Delta_{\text{modelo complejo}} \leq \Delta_{\text{modelo básico}}$. Es de interés observar que la estimación del coeficiente que acompaña al BMI depende de qué covariables hay en el modelo, excepto cuando todas las covariables presentes en el modelo sean **no correlacionadas** con BMI, lo cual ocurrirá las menos de las veces: en general las variables explicativas

Tabla 37: Tabla de ANOVA para comparar dos modelos de regresión

Modelo	SS	g.l.	Diferencia	g.l.	F
Básico	$\Delta_{\text{mod bás}}$	$n-2$			
Complejo	$\Delta_{\text{mod comp}}$	$n-4$	$\Delta_{\text{mod bás}} - \Delta_{\text{mod comp}}$	2	$\frac{(\Delta_{\text{mod bás}} - \Delta_{\text{mod comp}})/2}{\Delta_{\text{mod comp}}/(n-4)}$

están vinculadas entre sí de manera más o menos estrecha, eso significa que en general estarán (linealmente) correlacionadas.
Nuevamente se puede construir una tabla de ANOVA para resumir la información descripta hasta ahora. En la Tabla 37 describimos la forma en la que se presenta la información.
La resta $\Delta_{\text{modelo básico}} - \Delta_{\text{modelo complejo}}$ mide la mejora en el ajuste debida al modelo más complejo respecto del más sencillo. Los grados de libertad de esta resta será la resta de los grados de libertad de los dos ajustes, en el ejemplo $(n-4)-(n-2) = 2$. Esta cuenta da siempre la diferencia entre el número de coeficientes del modelo más complejo respecto del más básico. El test F se basa en la comparación de la mejora en el ajuste debido al modelo más complejo respecto del simple relativa al ajuste proporcionado por el modelo complejo (el mejor ajuste disponible), ambos divididos por sus grados de libertad. El test F para las hipótesis (74) rechaza H_0 cuando $F > F_{2,n-4,\alpha}$ (el percentil $1-\alpha$ de la distribución F con 2 grados de libertad en el numerador y $n-4$ grados de libertad en el denominador) o, equivalentemente, cuando el $p-$ valor calculado como $P\left(F_{2,n-4} > F_{\text{obs}}\right)$ es menor que α. En general, cuando se comparan

$$\begin{aligned} \text{Modelo complejo: } Y_i &= \beta_0^{\text{c}} + \sum_{k=1}^{p-1} \beta_k^{\text{c}} X_{ik} + \varepsilon_i \\ \text{Modelo simple: } Y_i &= \beta_0^{\text{s}} + \sum_{k=1}^{q-1} \beta_k^{\text{s}} X_{ik} + \varepsilon_i \end{aligned} \tag{75}$$

Es decir, cuando se testea

$$\begin{aligned} &H_0 : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0 \\ &H_1 : \text{ al menos uno de los } \beta_k \text{ con } k \text{ entre } q \text{ y } p-1 \text{ es tal que } \beta_k \neq 0 \end{aligned}$$

en el modelo (75), los grados de libertad del estadístico F serán $p-q$ en el numerador y $n-p$ en el denominador. Para los datos del archivo `azucar`, la tabla de ANOVA para chequear las hipótesis (74) es la que figura en la Tabla 38. Como el p-valor es menor a 0,05 resulta que cuando controlamos a la glucosa por el BMI,

el nivel de evolución del peso de cada paciente resulta significativo. Luego la evolución del peso es útil para predecir el nivel de glucosa, aún cuando controlamos por el BMI.

Tabla 38: Comparación de sumas de cuadrados para evaluar la significatividad de physact (categórica) una vez que se tiene a BMI (numérica) como regresora de glucosa

```
> uno<-lm(glucosa ~ bmi)
> dos<-lm(glucosa ~ Ievo + bmi)
> anova(uno,dos)
Analysis of Variance Table

Model 1: glucosa ~ bmi
Model 2: glucosa ~ Ievo + bmi
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    218 17328
2    216 15480  2    1848.1 12.894 5.128e-06 ***
---
```

4.15.2. Comparaciones múltiples

Cuando usamos un modelo de regresión (73), podemos querer estimar los efectos diferenciales entre los dos niveles de `peso.evo` que no involucren al basal. Cuando la variable categórica tenga más de tres categorías, habrá todavía más comparaciones que considerar, en este caso sólo resta considerar una. Esto puede hacerse estimando diferencias entre coeficientes de regresión. En el ejemplo, $\beta_3 - \beta_2$ indica cuánto más alta (o baja) es la función de respuesta para "subió de peso" (`peso.evo=3`) comparada con "mantuvo su peso" (`peso.evo=1`) para cualquier nivel de BMI pues

$$\begin{aligned} & E\left(Y \mid \texttt{bmi}, \texttt{peso.evo} = 3\right) - E\left(Y \mid \texttt{bmi}, \texttt{peso.evo} = 2\right) \\ & = \beta_0 + \beta_3 + \beta_{BMI} BMI - \beta_0 - \beta_2 - \beta_{BMI} BMI \\ & = \beta_3 - \beta_2. \end{aligned}$$

El estimador puntual de esta cantidad es, por supuesto, $\widehat{\beta}_3 - \widehat{\beta}_2$, y la varianza estimada de este estimador es

$$\widehat{Var}\left(\widehat{\beta}_3 - \widehat{\beta}_2\right) = \widehat{Var}\left(\widehat{\beta}_3\right) + \widehat{Var}\left(\widehat{\beta}_2\right) + 2\widehat{Cov}\left(\widehat{\beta}_3, \widehat{\beta}_2\right).$$

Las varianzas y covarianzas necesarias se pueden obtener a partir de la matriz de covarianza de los coeficientes de regresión.

```
>mm <-  summary(ajuste4)$cov.unscaled * (summary(ajuste4)$sigma)^2
            (Intercept)       Ievo2       Ievo3         bmi
(Intercept)  11.4002401 -2.46807671 -1.88607917 -0.34625635
Ievo2        -2.4680767  3.17653751  2.16249632  0.01112123
Ievo3        -1.8860792  2.16249632  2.80368157 -0.01071534
bmi          -0.3462563  0.01112123 -0.01071534  0.01299155
> sqrt(diag(mm))  #coincide con la columna de Std. Error del lm
(Intercept)       Ievo2       Ievo3         bmi
  3.3764242   1.7822844   1.6744198   0.1139805
```

De todos modos, esta cuenta la realiza, en general, el software estadístico. A continuación vemos las comparaciones de a pares realizadas con el método de la diferencia honestamente significativa de Tukey (HSD: *honestly significant difference*), realizada con nivel conjunto del 95 %. Mostramos los intervalos de confianza calculados con la instrucción `TukeyHSD`. Es muy importante en esta instrucción incorporar primero la variable continua y luego la categórica para obtener los resultados que queremos. También presentamos la salida (intervalos de confianza y tests) del comando `glht`, del paquete `multcomp`. Ahí vemos que exceptuando la diferencia entre los subgrupos dados por los niveles 2 y 3 de actividad física, las restantes diferencias son estadísticamente significativas a nivel conjunto 95 %.

```
#cambiamos el orden, covariable cont primero
> TukeyHSD(aov(glucosa ~ bmi + Ievo),"Ievo")
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = glucosa ~ bmi + Ievo)

$Ievo
        diff       lwr       upr     p adj
2-1 6.827351  2.627495 11.027207 0.0004795
3-1 8.472188  4.526816 12.417560 0.0000026
3-2 1.644837 -1.357564  4.647237 0.4007100

> ajuste4<-lm(glucosa ~  bmi + Ievo)
> library(multcomp)  #para testear combinaciones de los parametros
> posthoc <- glht(ajuste4, linfct = mcp(Ievo = "Tukey"),test = Ftest())
> summary(posthoc)
```

```
         Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: lm(formula = glucosa ~ bmi + Ievo)
Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0    6.802      1.782   3.817 0.000513 ***
3 - 1 == 0    8.496      1.674   5.074  < 1e-04 ***
3 - 2 == 0    1.694      1.287   1.317 0.383040
---
> confint(posthoc)
         Simultaneous Confidence Intervals
Multiple Comparisons of Means: Tukey Contrasts
Fit: lm(formula = glucosa ~ bmi + Ievo)
Quantile = 2.3509
95% family-wise confidence level

Linear Hypotheses:
           Estimate lwr     upr
2 - 1 == 0  6.8023   2.6123 10.9923
3 - 1 == 0  8.4963   4.5599 12.4327
3 - 2 == 0  1.6940  -1.3305  4.7186
```

En la Figura 48 se ve un gráfico de estos intervalos de confianza de nivel simultáneo 95 %. Sólo el intervalo para la diferencia de medias entre los grupos 2 y 3 contiene al cero. Los restantes quedan ubicados a la izquierda del cero. En el gráfico esto se ve más fácilmente que leyendo la tabla. Para el modelo con las covariables `Ievo` pero sin `bmi` también podríamos haber exhibido un gráfico como éste (de hecho, no lo hicimos puesto que ambos dan muy parecidos).

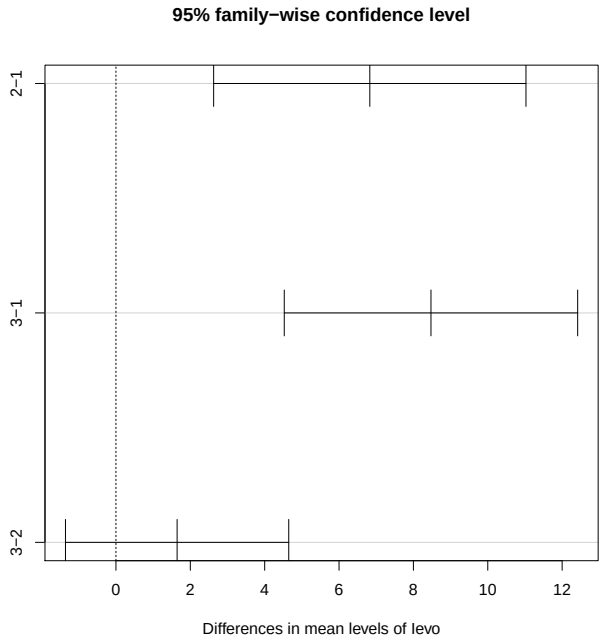
4.16. Modelos con interacción entre variables cuantitativas y cualitativas

Como ya dijimos, cuando proponemos un modelo de regresión lineal múltiple del estilo de

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \tag{76}$$

estamos asumiendo que los efectos de las variables X_1 y X_2 sobre la respuesta Y no interactúan entre sí: es decir, que el efecto de X_1 en Y no depende del valor que tome X_2 (y al revés, cambiando X_1 por X_2, el efecto de X_2 en Y no depende del valor que tome X_1). Cuando esto no sucede, es inadecuado proponer el modelo (76), y es necesario agregarle a dicho modelo un témino que intente dar cuenta

Figura 48: Intervalos de confianza de nivel simultáneo para las diferencias de los niveles medios de glucosa de cada grupo, controlados por el BMI.



de la *interacción* entre X_1 y X_2 en su relación con Y, es decir, del hecho de que el efecto de un predictor sobre la respuesta difiere de acuerdo al nivel de otro predictor. La manera estándar de hacerlo es agregarle al modelo (76) un término de interacción, es decir

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} \cdot X_{i2} + \varepsilon_i. \tag{77}$$

El modelo (77) es un caso particular del modelo de regresión lineal múltiple. Sea $X_{i3} = X_{i1} \cdot X_{i2}$ el producto entre las variables X_1 y X_2 medidas en el iésimo individuo, entonces el modelo (77) puede escribirse de la forma

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i,$$

que es un caso particular del modelo de regresión lineal múltiple presentado en (44). Algunas veces al coeficiente de la interacción se lo nota con los subíndices $1:2$, es decir $\beta_{1:2} = \beta_3$ para explicitar que es el coeficiente asociado a la interacción. Veamos un ejemplo.

Ejemplo 4.1 *Consideremos datos sobre la frecuencia cardíaca o pulso medido a 40 personas antes y después de ejercitar. Estos datos aparecen publicados en el manual del paquete BMDP, sin citar las fuentes. Figuran en la carpeta de datos de este apunte en el archivo* `pulso.txt`*. Se les pidió que registraran su pulso, luego que corrieran una milla, y luego volvieran a registrar su pulso. Además se registró su sexo, edad y si eran o no fumadores. De este modo, para cada individuo, se midieron las siguientes variables*

$$
\begin{aligned}
Y &= \textit{pulso luego de correr una milla (}\texttt{Pulso2}\textit{)} \\
X_1 &= \textit{pulso en reposo (}\texttt{Pulso1}\textit{)} \\
X_2 &= \begin{cases} 1 & \textit{si la persona es mujer} \\ 0 & \textit{en caso contrario} \end{cases} \\
X_3 &= \begin{cases} 1 & \textit{si la persona fuma} \\ 0 & \textit{en caso contrario} \end{cases} \\
X_4 &= \textit{edad}
\end{aligned}
$$

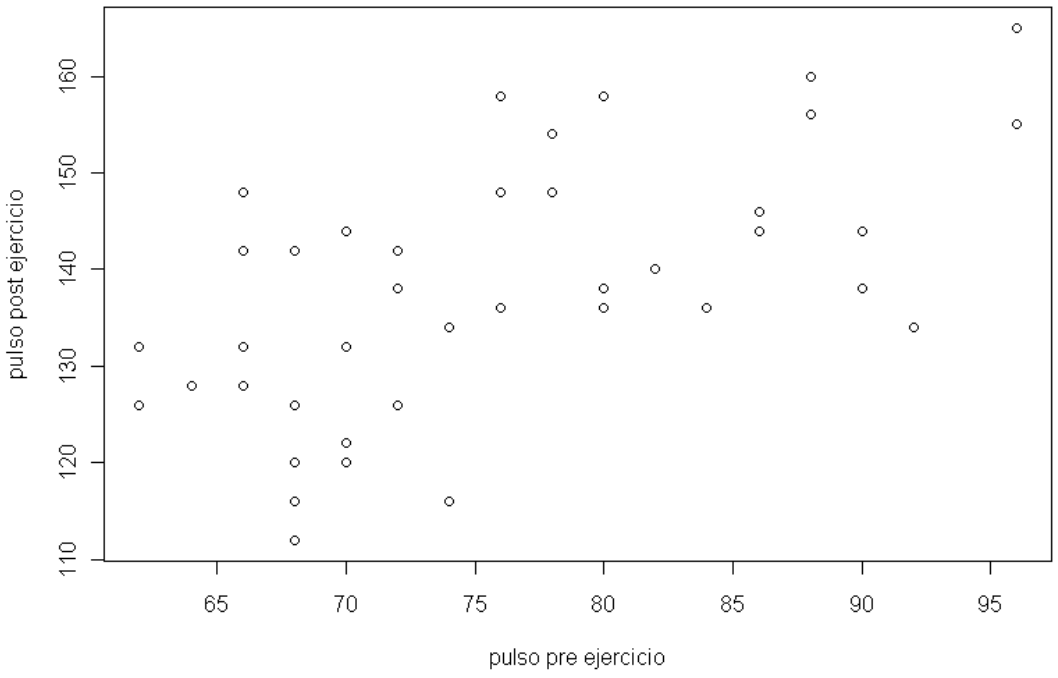
Interesa explicar el pulso post-ejercicio, en función de algunas de las demás covariables. Es de interés saber si la edad, o el hábito de fumar inciden en él. La frecuencia cardíaca es el número de contracciones del corazón o pulsaciones por unidad de tiempo. Su medida se realiza en unas condiciones determinadas (reposo o actividad) y se expresa en latidos por minuto.
Tanto el sexo como la condición de fumador son variables dummies o binarias. En la base de datos se las denomina $X_2 =$ `mujer` *y* $X_3 =$ `fuma`*. Las restantes son variables continuas. En la Figura 49 hacemos un scatter plot de* Y *versus* X_1*. En él se puede ver que a medida que* X_1 *crece también lo hace* Y*, y que una relación lineal es una buena descripción (inicial) de la relación entre ellas.*

Si identificamos en ese gráfico a las observaciones según su sexo, obtenemos el gráfico de dispersión que aparece en la Figura 50. En él observamos que el género de la persona parece influir en la relación entre ambas variables.
Querríamos cuantificar el efecto del género en el pulso medio post ejercicio. Para ello vamos a ajustar un modelo de regresión lineal múltiple con el pulso post ejercicio como variable dependiente. Proponemos un modelo lineal múltiple para estos datos. El modelo múltiple sería en este caso

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \tag{78}$$

Como ya vimos en la Sección 4.13.2, este modelo sin interacción propone que el pulso medio post-ejercicio es una función lineal del pulso pre-ejercicio, con dos

Figura 49: Gráfico de dispersión del pulso post-ejercicio versus el pulso pre-ejercicio, para 40 adultos. Archivo: `pulso.txt`



rectas diferentes para las mujeres y los hombres, pero estas rectas tienen la misma pendiente. O sea, la ecuación (78) propone que para las mujeres, (o sea, cuando $X_2 = 1$)

$$\begin{aligned} E\left(Y \mid X_1, X_2 = 1\right) &= \beta_0 + \beta_1 X_1 + \beta_2 \\ &= (\beta_0 + \beta_2) + \beta_1 X_1 \end{aligned}$$

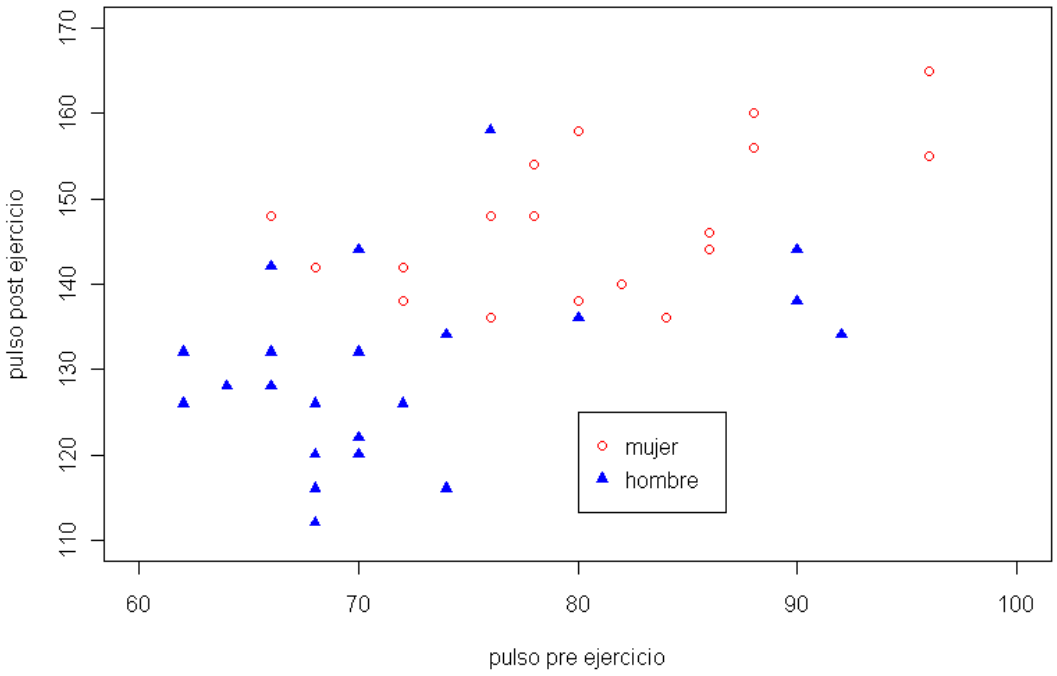
mientras que para los hombres (cuando $X_2 = 0$) se tiene

$$E\left(Y \mid X_1, X_2 = 0\right) = \beta_0 + \beta_1 X_1.$$

La salida del ajuste del modelo está en la Tabla 39. De acuerdo a ella, la recta ajustada es

$$\widehat{Y} = 93{,}0970 + 0{,}5157 \cdot X_1 + 12{,}7494 \cdot X_2$$

Figura 50: Gráfico de dispersión del pulso post-ejercicio versus el pulso pre-ejercicio, identificando el sexo de cada observación.



El coeficiente estimado de `mujer` es positivo, indicando que cuando la variable X_2 aumenta de 0 a 1 (`mujer` $= 0$ quiere decir que se trata de un hombre), el pulso medio post ejercicio crece, es decir, el pulso medio de las mujeres es mayor que el de los hombres si uno controla por pulso en reposo. ¿Será estadísticamente significativa esta observación? Para evaluarlo, hacemos un test de

$$H_0 : \beta_2 = 0 \text{ versus } H_0 : \beta_2 \neq 0$$

asumiendo que el modelo contiene al pulso en reposo. El estadístico observado resulta ser $t_{obs} = 3{,}927$ y $p-$valor $= 0{,}000361$. Entonces, rechazamos la hipótesis nula y concluimos que $\beta_2 \neq 0$. Si construyéramos un intervalo de confianza para β_2, éste resultaría contenido enteramente en $(-\infty, 0)$. Por eso concluimos que el verdadero valor poblacional de β_2 es menor a cero. Es decir, para las dos poblaciones de personas (hombres y mujeres) con el mismo pulso en reposo, en promedio los pulsos medios luego de ejercitar serán mayores en las mujeres que en los hombres.

Tabla 39: Ajuste del modelo lineal múltiple $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$, donde X_1 = pulso pre ejercicio (`Pulso1`), X_2 = indicador de mujer (`mujer`), Y = pulso post ejercicio (`Pulso2`).

```
> ajuste1<-lm(Pulso2~ Pulso1+mujer)
> summary(ajuste1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  93.0970    12.5157   7.438 7.44e-09
Pulso1        0.5157     0.1715   3.007 0.004725
mujer        12.7494     3.2468   3.927 0.000361

Residual standard error: 9.107 on 37 degrees of freedom
Multiple R-squared: 0.5445, Adjusted R-squared: 0.5199
F-statistic: 22.12 on 2 and 37 DF,  p-value: 4.803e-07
```

Para entender mejor este modelo escribimos las dos rectas ajustadas en cada caso. El modelo ajustado para las mujeres, $(X_2 = 1)$ es

$$\begin{aligned}\widehat{Y} &= (93{,}0970 + 12{,}7494) + 0{,}5157 \cdot X_1 \\ &= 105{,}85 + 0{,}5157 \cdot X_1\end{aligned}$$

mientras que para los hombres $(X_2 = 0)$

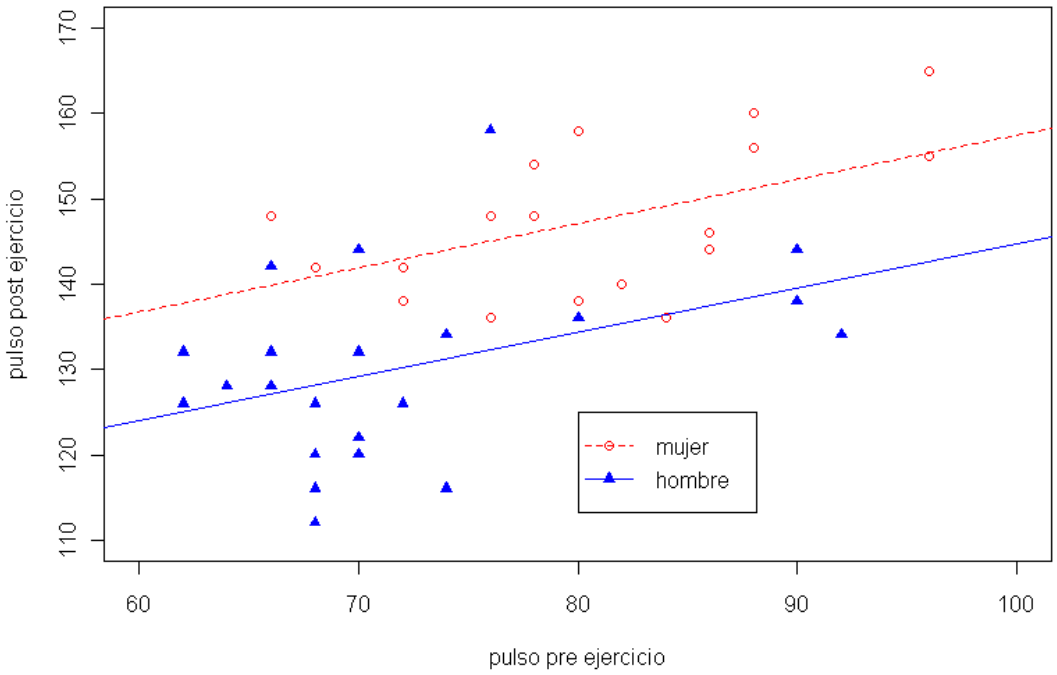
$$\widehat{Y} = 93{,}0970 + 0{,}5157 \cdot X_1.$$

Las dos rectas están graficadas en la Figura 51, junto con las observaciones identificadas por sexo. Observemos que ambas rectas son paralelas: en ambos grupos una unidad (un latido por minuto) de aumento en el pulso en reposo está asociado con un incremento en 0,5157 latidos por minuto de la frecuencia cardíaca post ejercicio, en promedio. Esto es consecuencia del modelo propuesto.
Ahora queremos proponer un modelo con interacción para estos datos. Es decir proponemos el modelo

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_{1:2} X_{i1} \cdot X_{i2} + \varepsilon_i \tag{79}$$

Como la variable X_2 asume solamente valores 0 y 1, el término de la interacción $X_{i1} \cdot X_{i2}$ valdrá 0 siempre que $X_2 = 0$ (o sea para los hombres), y será igual a

Figura 51: Rectas ajustadas para los dos g éneros (modelo sin interacción).



X_1 siempre que $X_2 = 1$ (o sea para las mujeres). En la población de personas ejercitando, esta nueva variable tendrá coeficiente $\beta_{1:2}$. Llamemos $X = (X_1, X_2)$. Si escribimos el modelo propuesto para los dos grupos de observaciones, tendremos que cuando `mujer` $= 1$,

$$\begin{aligned} E(Y \mid X) &= \beta_0 + \beta_1 X_1 + \beta_2 1 + \beta_{1:2} X_1 \cdot 1 \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_{1:2}) X_1 \quad \text{mujeres} \end{aligned}$$

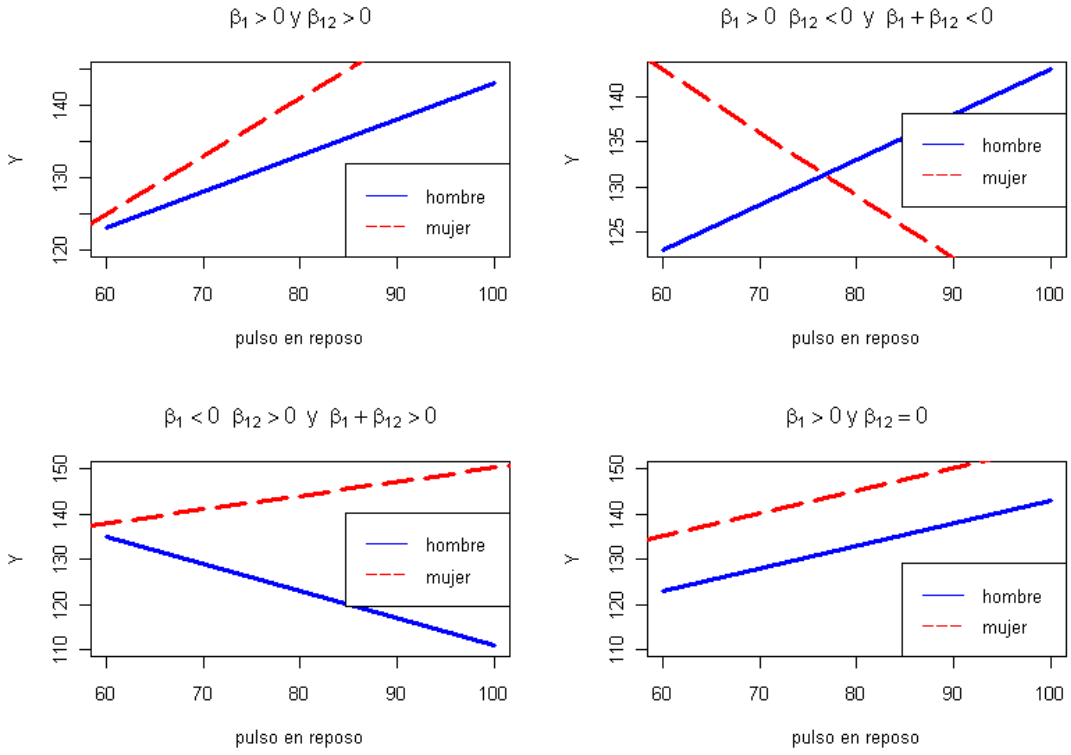
Mientras que cuando `mujer` $= 0$, proponemos

$$\begin{aligned} E(Y \mid X) &= \beta_0 + \beta_1 X_1 + \beta_2 0 + \beta_{1:2} X_1 \cdot 0 \\ &= \beta_0 + \beta_1 X_1 \quad \text{hombres} \end{aligned}$$

Es decir que para cada grupo estamos proponiendo ajustar dos rectas distintas. Observemos que estas rectas no están atadas (como sí lo estaban en el modelo aditivo con una explicativa binaria y una continua, en el que ajustábamos **dos**

rectas paralelas). Por otro lado, la interpretación de los coeficientes del modelo cambia. Analicemos cada uno. El coeficiente de X_1 (β_1) es la pendiente del `pulso1` en el grupo de hombres. Indica que por cada aumento en una unidad en el pulso en reposo entre los hombres, el pulso medio post ejercicio aumenta (o disminuye, según el signo) β_1 unidades. El coeficiente de la interacción ($\beta_{1:2}$) representa el aumento (o la disminución) de la pendiente en el grupo de las mujeres con respecto al de los hombres. Si $\beta_{1:2} = 0$ esto significaría que ambas rectas son paralelas. Los distintos valores que pueden tomar β_1 y $\beta_{1:2}$ dan lugar a distintos posibles tipos de interacción entre las variables, según se ve en la Figura 52.

Figura 52: Gráfico de posibles combinaciones de valores de β_1 y $\beta_{1:2}$ para el modelo (79).



El ajuste del modelo con interacción a los datos se muestra en la Tabla 40. El test de

$$H_0 : \beta_{1:2} = 0 \text{ versus } H_0 : \beta_{1:2} \neq 0$$

asumiendo que el modelo contiene al pulso en reposo y a la indicadora de mujer, tiene por estadístico $t_{obs} = 0{,}211$ y $p-$valor $= 0{,}834$. Esto nos dice que esta muestra

Tabla 40: Ajuste del modelo lineal con interacción entre X_1 = pulso pre ejercicio (`Pulso1`), X_2 = indicador de mujer (`mujer`), Y = pulso post ejercicio (`Pulso2`).

```
> ajuste2<-lm(Pulso2~ Pulso1 * mujer)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  95.42838   16.80929   5.677 1.88e-06
Pulso1        0.48334    0.23157   2.087    0.044
mujer         7.05575   27.14749   0.260    0.796
Pulso1:mujer  0.07402    0.35033   0.211    0.834

Residual standard error: 9.227 on 36 degrees of freedom
Multiple R-squared: 0.5451, Adjusted R-squared: 0.5072
F-statistic: 14.38 on 3 and 36 DF,  p-value: 2.565e-06
```

no provee evidencia suficiente de que el pulso en reposo tenga un efecto diferente en el pulso post ejercicio dependiendo del sexo de la persona.
Como la interacción no es estadísticamente significativa, no la retendremos en el modelo de regresión. Sin embargo, veamos cuanto dan las dos rectas ajustadas en este caso. Cuando `mujer` $= 1$,

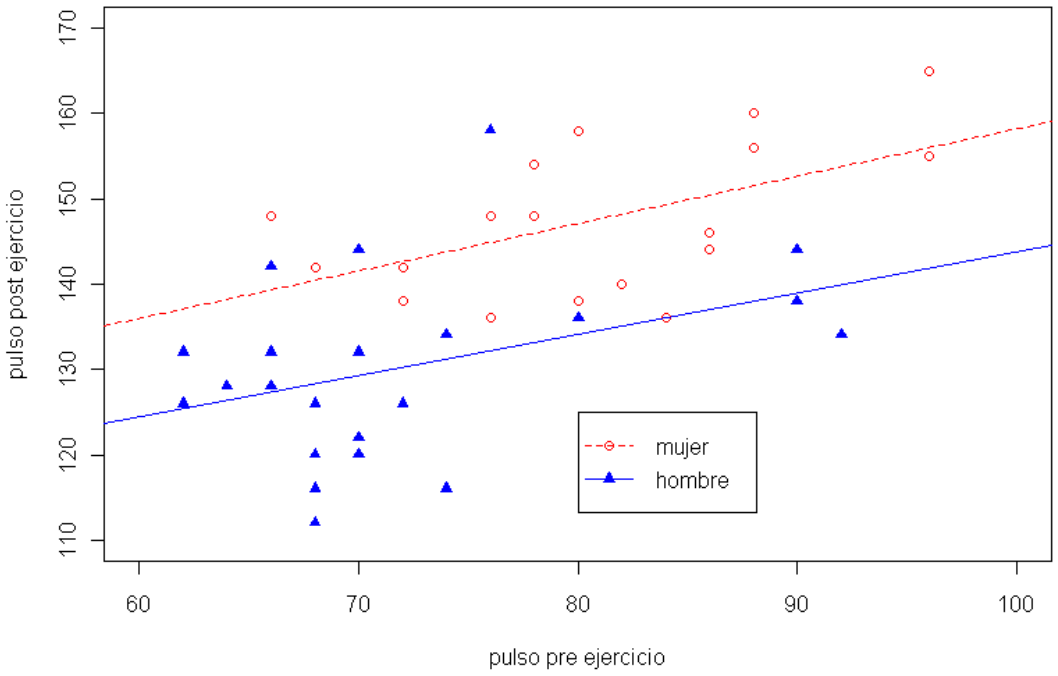
$$\begin{aligned}\widehat{Y} &= 95{,}429 + 0{,}483 \cdot X_1 + 7{,}056 \cdot 1 + 0{,}074 X_1 \cdot 1 \\ &= 102{,}485 + 0{,}557 \cdot X_1\end{aligned}$$

Mientras que cuando `mujer` $= 0$, resulta

$$\begin{aligned}\widehat{Y} &= 95{,}429 + 0{,}483 \cdot X_1 + 7{,}056 \cdot 0 + 0{,}074 X_1 \cdot 0 \\ &= 95{,}429 + 0{,}483 \cdot X_1\end{aligned}$$

El gráfico de ambas rectas puede verse en la Figura 53. Estas dos rectas no tienen la misma pendiente, ni la misma ordenada al origen. En el rango de interés, sin embargo, la recta que describe el pulso medio post-ejercicio para las mujeres está completamente sobre la de los hombres. Esto implica que a lo largo de todos los valores relevantes del pulso en reposo, predeciremos valores de pulso post-ejercicio mayores para las mujeres que para los hombres. Si comparamos los ajustes obtenidos para los modelos que explican a Y con las variables `Pulso1` y `mujer` sin interacción (78) y con interacción (79), que aparecen en las Tablas 39 y 40, respectivamente, vemos que son muy diferentes.

Figura 53: Rectas ajustadas por mínimos cuadrados para distintos niveles de sexo, con el término de interacción incluido.



En la Tabla 41 resumimos lo observado. Cuando el término de la interacción se incluye en el modelo, el coeficiente de `mujer` se reduce en magnitud, casi a la mitad. Además, su error estándar aumenta multiplicándose por un factor de 8. En el modelo sin término de interacción, el coeficiente de `mujer` es significativamente distinto de cero, a nivel 0,05; esto no ocurre cuando incluimos el término de interacción en el modelo, en ese caso la variable `mujer` deja de ser significativa. El coeficiente de determinación (R^2) no cambia al incluir la interacción, sigue valiendo 0,545. Más aún, el coeficiente de determinación ajustado decrece ligeramente con la incorporación de una covariable más al modelo. Al tomar en cuenta simultáneamente todos estos hechos, concluimos que la inclusión del término de interacción de `Pulso1` · `mujer` en el modelo no explica ninguna variabilidad adicional en los valores observados del pulso post-ejercicio, más allá de lo que es explicado por las variables `mujer` y `Pulso1` en forma aditiva. La información proporcionada por este término es redundante.

¿Por qué sucede esto? Muchas vece sucede que al incorporar una nueva variable al modelo ajustado, se pierde la significatividad de alguna o varias variables ya incluidas previamente. Si además de suceder esto aparece una inestabilidad de los coeficientes estimados, difiriendo sustancialmente los valores estimados de algunos coeficientes en los dos modelos, y en particular, se observa un aumento grosero de los errores estándares: esto suele ser un síntoma de *colinealidad o multicolinealidad* entre los predictores. La colinealidad ocurre cuando dos o más variables explicativas están altamente correlacionadas, a tal punto que, esencialmente, guardan la misma información acerca de la variabilidad observada de Y. En la Sección 5.3.1 presentaremos algunas maneras de detectar y resolver la multicolinealidad.
En este caso, la variable artificial `Pulso1` · `mujer` está fuertemente correlacionada con `mujer` ya que el coeficiente de correlación de Pearson es $r_{\texttt{mujer,Pulso1·mujer}} = 0{,}99$, como aparece en la Tabla 42. Como la correlación entre las variables es tan grande, la capacidad explicativa de `Pulso1` · `mujer` cuando `mujer` está en el modelo es pequeña.

Tabla 41: Tabla comparativa de los ajustes con y sin interacción para las covariables `Pulso1` y `mujer`.

	Sin interacción	Con interacción
Coeficiente $\widehat{\beta}_2$	12,749	7,056
Error estándar de $\widehat{\beta}_2$	3,247	27,147
Valor del estadístico t	3,927	0,26
p−valor	0,000361	0,796
R^2	0,5445	0,5451
R^2 ajustado	0,5199	0,5072

Tabla 42: Correlaciones de Pearson entre X_1 = pulso pre ejercicio (`Pulso1`), X_2 = indicador de mujer (`mujer`) e Y = pulso post ejercicio (`Pulso2`).

	Pulso1	mujer	Pulso1·mujer
Pulso1	1	0,453	0,53
mujer	0,453	1	0,99
Pulso1·mujer	0,53	0,99	1

Un modo de resolver el problema de la multicolinealidad es trabajar con los datos centrados para la o las variables predictoras que aparecen en más de un término del modelo. Esto es, usar no la variable X tal como fue medida, sino la diferencia entre el valor observado y el valor medio en la muestra.

4.17. Interacción entre dos variables cuantitativas

En la sección anterior presentamos la interacción entre dos variables cuando una es cualitativa y la otra cuantitativa. Ahora nos ocuparemos de estudiar la situación en la que las dos variables que interesan son cuantitativas. Vimos que el modelo aditivo propone que cuando la covariable X_j aumenta una unidad, la media de Y aumenta en β_j unidades independientemente de cuáles sean los valores de las otras variables. Esto implica paralelismo de las rectas que relacionan a Y y X_j, cualesquiera sean los valores que toman las demás variables.
En nuestro ejemplo de los bebés de bajo peso, analizado en la Sección 4.7, propusimos un modelo de regresión lineal múltiple con dos variables predictoras. Recordemos que habíamos definido

$$\begin{aligned} Y_i &= \text{ perímetro cefálico del iésimo niño, en centímetros (}\texttt{headcirc}\text{)} \\ X_{i1} &= \text{ edad gestacional del iésimo niño, en semanas (}\texttt{gestage}\text{)} \\ X_{i2} &= \text{ peso al nacer del iésimo niño, en gramos (}\texttt{birthwt}\text{)} \end{aligned}$$

Propusimos el siguiente modelo,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon. \tag{80}$$

El modelo ajustado figura en la Tabla 19, página 125. La superficie ajustada resultó ser

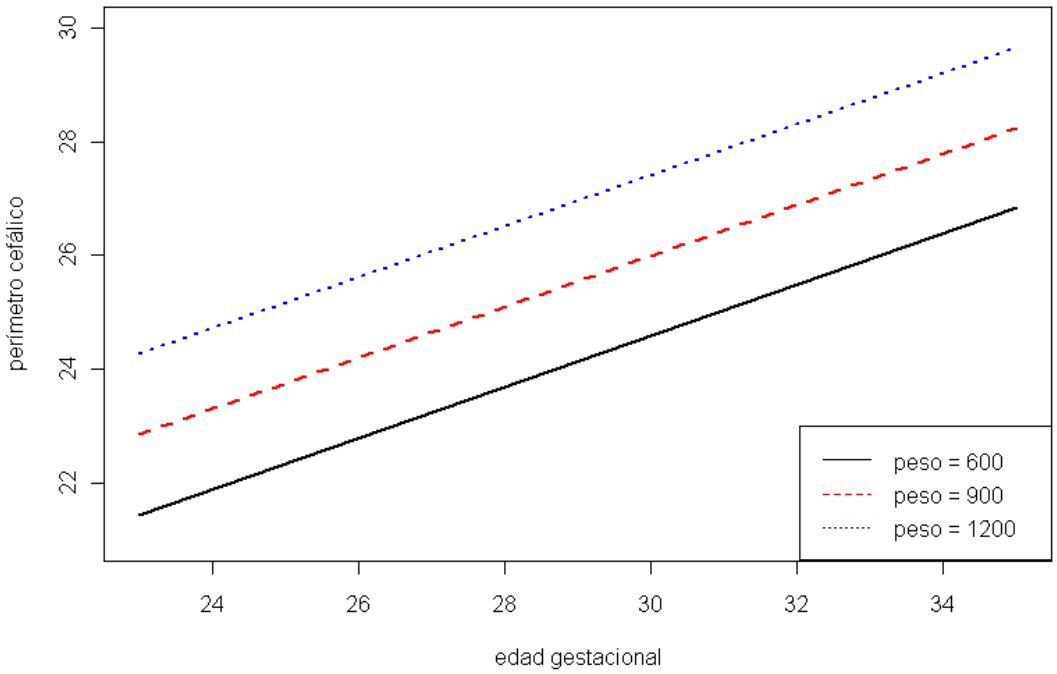
$$\widehat{Y} = 8{,}3080 + 0{,}4487 X_1 + 0{,}0047 X_2.$$

Cuando controlamos por X_2 (peso al nacer), la ecuación (parcial) ajustada que relaciona el perímetro cefálico y la edad gestacional es

$$\begin{aligned} X_2 &= 600, \quad \widehat{Y} = 8{,}3080 + 0{,}4487 X_1 + 0{,}0047 \cdot 600 = 11{,}128 + 0{,}4487 X_1 \\ X_2 &= 900, \quad \widehat{Y} = 8{,}3080 + 0{,}4487 X_1 + 0{,}0047 \cdot 900 = 12{,}538 + 0{,}4487 X_1 \\ X_2 &= 1200, \quad \widehat{Y} = 8{,}3080 + 0{,}4487 X_1 + 0{,}0047 \cdot 1200 = 13{,}948 + 0{,}4487 X_1 \end{aligned}$$

Para cada nivel posible de peso al nacer, por cada unidad de aumento en la edad gestacional se espera un aumento de 0,448 unidades (cm.) en el perímetro cefálico al nacer. Gráficamente, esto se ve representado en la Figura 54. Lo mismo sucedería si controláramos por X_1 en vez de X_2 : tendríamos rectas paralelas, de pendiente 0,0047.
Este modelo asume que no existe interacción entre las variables. El modelo (80) fuerza a que los efectos de las covariables en la variable dependiente sean aditivos, es decir, el efecto de la edad gestacional será el mismo para todos los valores del peso al nacer, y viceversa, porque el modelo no le permitirá ser de ninguna otra forma. A menudo este modelo es demasiado simple para ser adecuado, aunque en

Figura 54: Perímetro cefálico esperado en función de la edad gestacional, controlando por peso al nacer, para tres posibles valores de peso al nacer (600, 900 y 1200g.) en el modelo sin interacción.



muchos conjuntos de datos proporciona una descripción satisfactoria del vínculo entre las variables.
Cuando esto no suceda, es decir, cuando pensemos que tal vez la forma en que el perímetro cefálico varíe con la edad gestacional dependa del peso al nacer del bebé, será necesario descartar (o validar) esta conjetura. Una manera de investigar esta posibilidad es incluir un término de interacción en el modelo. Para ello, creamos la variable artificial que resulta de hacer el producto de las otras dos: $X_3 = X_1 \cdot X_2 =$ `gestage` $\cdot$ `birthwt`, y proponemos el modelo

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \\ Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{1:2} X_1 \cdot X_2 + \varepsilon \end{aligned} \tag{81}$$

Este es un caso especial de un modelo de regresión con tres variables regresoras. ¿Cómo se interpreta este modelo para dos variables cuantitativas? En este caso

decimos que existe interacción estadística cuando **la pendiente** de la relación entre la variable respuesta y una variable explicativa **cambia** para distintos niveles de las otras variables. Para entenderlo, escribamos el modelo propuesto cuando controlamos el valor de X_2.

$$\begin{aligned} E\left(Y \mid X_1, X_2 = 600\right) &= \beta_0 + \beta_1 X_1 + \beta_2 600 + \beta_{1:2} X_1 \cdot 600 \\ &= \underbrace{\beta_0 + \beta_2 600}_{\text{ordenada al origen}} + \underbrace{(\beta_1 + \beta_{1:2} 600)}_{\text{pendiente}} X_1 \end{aligned}$$

$$\begin{aligned} E\left(Y \mid X_1, X_2 = 900\right) &= \beta_0 + \beta_1 X_1 + \beta_2 900 + \beta_{1:2} X_1 900 \\ &= \underbrace{\beta_0 + \beta_2 900}_{\text{ordenada al origen}} + \underbrace{(\beta_1 + \beta_{1:2} 900)}_{\text{pendiente}} X_1 \end{aligned}$$

$$\begin{aligned} E\left(Y \mid X_1, X_2 = 1200\right) &= \beta_0 + \beta_1 X_1 + \beta_2 1200 + \beta_{1:2} X_1 1200 \\ &= \underbrace{\beta_0 + \beta_2 1200}_{\text{ordenada al origen}} + \underbrace{(\beta_1 + \beta_{1:2} 1200)}_{\text{pendiente}} X_1 \end{aligned}$$

En general

$$\begin{aligned} E\left(Y \mid X_1, X_2\right) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{1:2} X_1 X_2 \\ &= \underbrace{\beta_0 + \beta_2 X_2}_{\text{ordenada al origen}} + \underbrace{(\beta_1 + \beta_{1:2} X_2)}_{\text{pendiente}} X_1 \end{aligned} \tag{82}$$

Luego, en el modelo (81), la pendiente de la relación entre X_1 e Y depende de X_2, decimos entonces que existe interacción entre las variables.
Entonces, cuando X_2 aumenta en una unidad, la pendiente de la recta que relaciona Y con X_1 aumenta en $\beta_{1:2}$. En este modelo, al fijar X_2, $E\left(Y \mid X_1, X_2\right)$ es una función lineal de X_1, pero la pendiente de la recta depende del valor de X_2. Del mismo modo, $E\left(Y \mid X_1, X_2\right)$ es una función lineal de X_2, pero la pendiente de la relación varía de acuerdo al valor de X_1.
Si $\beta_{1:2}$ no fuera estadísticamente significativa, entonces los datos no avalarían la hipótesis de que el cambio en la respuesta con un predictor dependa del valor del otro predictor, y podríamos ajustar directamente un modelo aditivo, que es mucho más fácil de interpretar.

Ejemplo 4.2 *Consideremos un ejemplo de datos generados. Para* $n = 40$ *pacientes se miden tres variables:*
$X_1 =$ *cantidad de droga A consumida*
$X_2 =$ *cantidad de droga B consumida*

Y = *variable respuesta*
Proponemos un modelo con interacción para los datos, que figuran en el archivo ***ejemploint.txt****. Antes de ajustar un modelo, veamos los estadísticos descriptivos de las dos variables, en la Tabla 43. Ajustamos el modelo (81). En la Tabla 44 aparece la salida. Vemos que el coeficiente asociado al término de interacción*

Tabla 43: Estadísticos descriptivos de las variables $X_1 =$ `drogaA` y $X_2 =$ `drogaB`.

```
> summary(drogaA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.207   4.449   7.744   8.107  11.100  13.590

> summary(drogaB)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.18   38.44   63.02   59.58   82.61   93.76
```

es 2,771 *y el test t rechaza la hipótesis* $H_0 : \beta_{1:2} = 0$ *(*$p-valor < 2 \cdot 10^{-16}$*). Concluímos que la interacción resulta estadísticamente significativa, así como lo son los restantes coeficientes asociados a las dos drogas. Luego, hay variación en la pendiente de la relación entre la respuesta y la cantidad de droga A ingerida, al variar la cantidad de droga B consumida. Esto puede verse más fácilmente en el gráfico de la Figura 55. En este caso vemos que las dos drogas potencian su efecto en la variable respuesta, ya que a medida que la cantidad de droga A crece (en el gráfico pasa de 4 a 7 y luego a 11) la variable respuesta crece al crecer la droga B, con pendientes cada vez mayores. Tienen interacción positiva. Las rectas graficadas en dicha figura son*

$$\begin{aligned} \texttt{drogaA} &= 4 \qquad \widehat{Y} = -53{,}92 + 16{,}59 \cdot 4 + 6{,}22X_2 + 2{,}77 \cdot 4 \cdot X_2 \\ \widehat{Y} &= 12{,}44 + 17{,}3X_2 \end{aligned}$$

$$\begin{aligned} \texttt{drogaA} &= 7 \qquad \widehat{Y} = -53{,}92 + 16{,}59 \cdot 7 + 6{,}22X_2 + 2{,}77 \cdot 7 \cdot X_2 \\ \widehat{Y} &= 62{,}21 + 25{,}61X_2 \end{aligned}$$

$$\begin{aligned} \texttt{drogaA} &= 11 \qquad \widehat{Y} = -53{,}92 + 16{,}59 \cdot 11 + 6{,}22X_2 + 2{,}77 \cdot 11 \cdot X_2 \\ \widehat{Y} &= 128{,}57 + 36{,}69X_2 \end{aligned}$$

Debería resultar claro en este caso, que necesitamos conocer el valor de la droga A para poder decir cuánto aumenta la respuesta media al aumentar en una unidad la

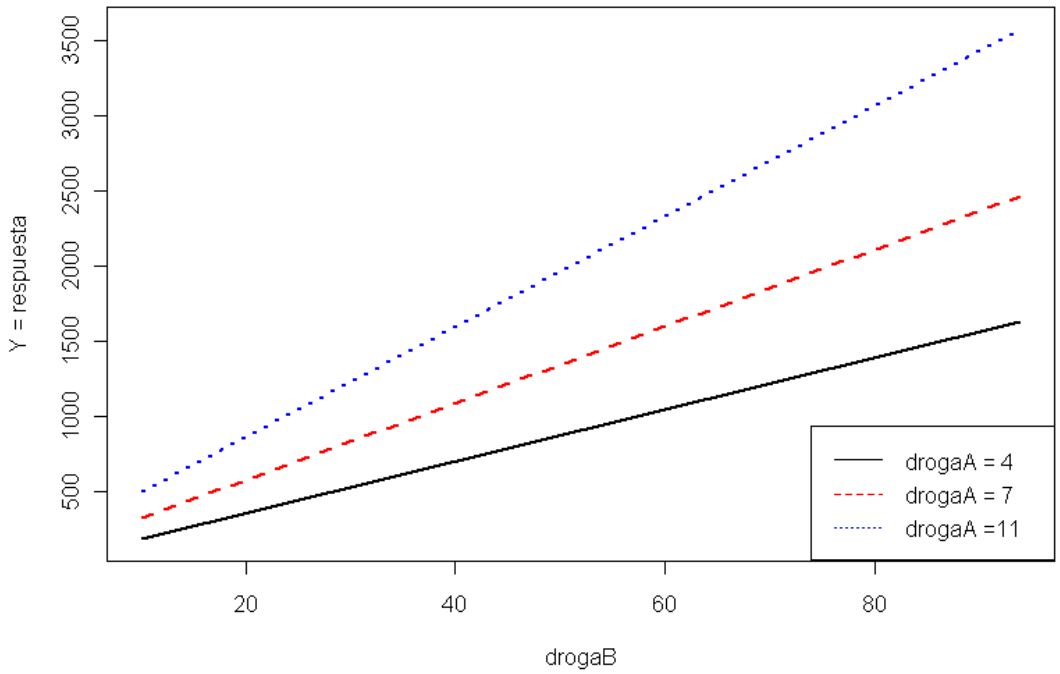
Tabla 44: Ajuste del modelo lineal múltiple (81) $E(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{1:2} X_1 X_2$, donde $X_1 =$ `drogaA`, $X_2 =$ `drogaB`, $Y =$ `respuesta`.

```
> summary(  ajuste5)

Call:
lm(formula = YY ~ drogaA * drogaB)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -53.92176   42.27242  -1.276  0.21027
drogaA         16.59288    4.92500   3.369  0.00181
drogaB          6.22153    0.63436   9.808 1.04e-11
drogaA:drogaB   2.77152    0.07774  35.651  < 2e-16
---

Residual standard error: 44.04 on 36 degrees of freedom
Multiple R-squared: 0.9979,     Adjusted R-squared: 0.9977
F-statistic:  5650 on 3 and 36 DF,  p-value: < 2.2e-16
```

cantidad de droga B consumida. Para los tres valores graficados, tendríamos tres respuestas distintas: 17,3, 25,61 *y* 36,69 *(para* `drogaA` $= 4$, *7 y 11, respectivamente).*

Hay que tener mucho cuidado en la interpretación de los coeficientes de cada covariable cuando el modelo contiene interacciones. Este modelo es mucho más complicado que el aditivo. Por esta razón, cuando se ajusta un modelo con interacción y no se rechaza la hipótesis de que la interacción sea cero, es mejor eliminar el término de interacción del modelo antes de interpretar los efectos parciales de cada variable. Sin embargo, cuando existe clara evidencia de interacción (se rechaza $H_0 : \beta_{1:2} = 0$), hay que conservar los términos asociados a las variables originales en el modelo lineal, aún cuando no resulten ser significativos, ya que el efecto de cada variable cambia según el nivel de las otras variables, ver (82). Es decir, si en el ajuste presentado en la Tabla 44 la interacción hubiera resultado significativa y el efecto de la droga A no hubiera resultado significativo, de todos modos, debería conservarse la droga A como covariable en el modelo, puesto que se conservará la interacción.
Veamos un ejemplo donde el efecto de la interacción es más fuerte aún.

Figura 55: Variable respuesta `Y` ajustada en función de la `drogaB`, controlando por `drogaA`, para tres posibles valores de `drogaA` (4, 7 y 11) en el modelo con interacción.



Ejemplo 4.3 *El conjunto de datos está en el archivo* `ejemploint3.txt`. *Nuevamente se trata de datos generados para los que se midieron las tres variables descriptas en el principio de esta sección, es decir, niveles de droga A* (X_1), *droga B* (X_2) *y la respuesta* (Y). *El modelo ajustado figura en la Tabla 45.*
Ahí vemos que tanto el coeficiente de la interacción, como los otros dos coeficientes que acompañan a las covariables son significativamente distintos de cero. En la Figura 56 vemos las rectas ajustadas para tres valores fijos de `drogaB`. *En ella vemos que el efecto de la interacción cambia de sentido al vínculo entre la respuesta* `Y` *y la* `drogaA` *al aumentar la cantidad de* `drogaB`, *ya que pasa de ser un potenciador de la variable respuesta, auméntadola considerablemente al aumentar la cantidad de* `drogaA`, *cuando la cantidad de* `drogaB` *es 10, a tener un vínculo inverso con* `Y` *cuando la cantidad de* `drogaB` *es 90, en el sentido que a mayor cantidad de* `drogaA` *la variable respuesta disminuye en este caso. En el caso de* `drogaB` $= 50$,

Tabla 45: Modelo ajustado para los datos del archivo `ejemploint3.txt`, con las variables explicativas X_1 = `drogaA` y X_2 = `drogaB` y la interacción entre ellas, para explicar a Y.

```
> summary(ajuste7)
Call:
lm(formula = Y7 ~ drogaA * drogaB)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   2488.19403   31.27861   79.55  < 2e-16
drogaA         151.87124    3.64415   41.67  < 2e-16
drogaB           4.92268    0.46938   10.49 1.71e-12
drogaA:drogaB   -3.00872    0.05752  -52.30  < 2e-16
---
Residual standard error: 32.59 on 36 degrees of freedom
Multiple R-squared: 0.9965,        Adjusted R-squared: 0.9962
F-statistic:  3427 on 3 and 36 DF,  p-value: < 2.2e-16
```

vemos que el vínculo entre ***drogaA*** *y la respuesta desaparece, ya que la recta parece horizontal (la pendiente estimada es exactamente cero cuando* ***drogaB*** $= 50{,}47703$*). Las tres rectas graficadas son*

$$\begin{aligned} \texttt{drogaB} &= 10 \qquad \widehat{Y} = 2488{,}194 + 151{,}871X_1 + 4{,}923 \cdot 10 - 3{,}0087 \cdot X_1 \cdot 10 \\ \widehat{Y} &= 2537{,}4 + 121{,}78X_1 \end{aligned}$$

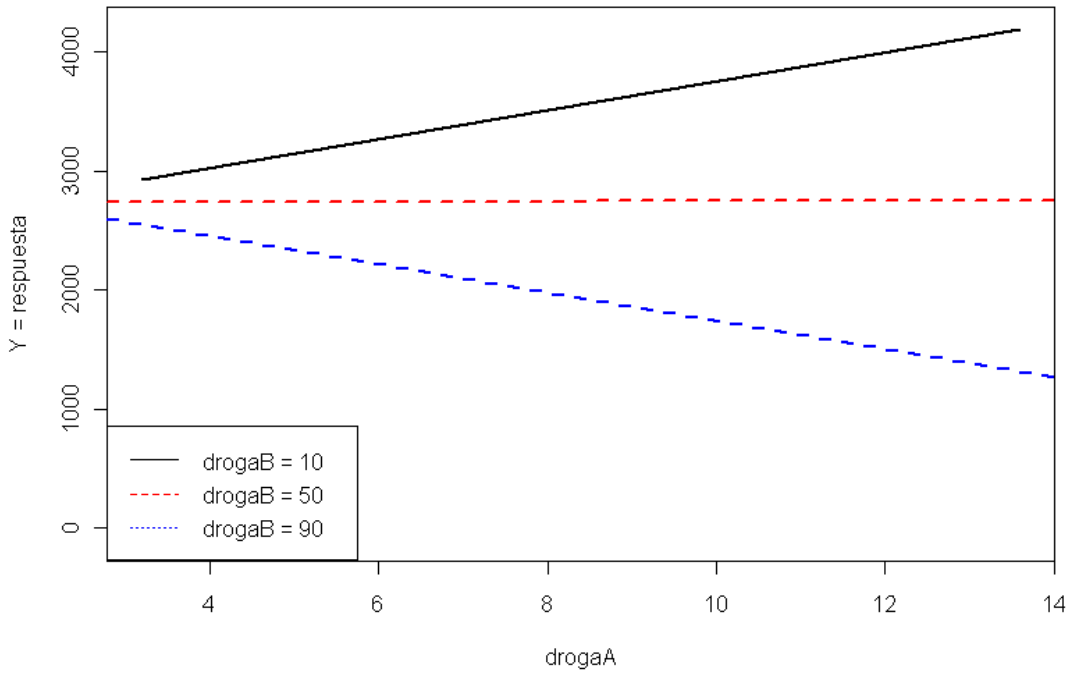
$$\begin{aligned} \texttt{drogaB} &= 50 \qquad \widehat{Y} = 2488{,}194 + 151{,}871X_1 + 4{,}923 \cdot 50 - 3{,}0087 \cdot X_1 \cdot 50 \\ \widehat{Y} &= 2734{,}2 + 1{,}436X_1 \end{aligned}$$

$$\begin{aligned} \texttt{drogaB} &= 90 \qquad \widehat{Y} = 2488{,}194 + 151{,}871X_1 + 4{,}923 \cdot 90 - 3{,}0087 \cdot X_1 \cdot 90 \\ \widehat{Y} &= 2931{,}3 - 118{,}91X_1 \end{aligned}$$

En este caso observamos que para hablar del efecto que tiene en la media de Y el aumento de una unidad de la ***drogaA debemos saber cuál es el valor de la drogaB*** *(ya que Y podría crecer, quedar constante o incluso disminuir) con un aumento de una unidad de la* ***drogaA****. En el modelo aditivo (sin interacción) uno podía siempre cuantificar la variación de la respuesta ante un aumento de*

una unidad de una covariable sin necesidad de conocer siquiera el valor de la otra covariable, mientras se mantuviera constante. Decíamos, en el ejemplo de los bebés de bajo peso, que manteniendo el peso constante, el aumento de una semana en la edad gestacional de un bebé repercutía en un aumento de 0,45 *cm. del perímetro cefálico esperado del bebé al nacer. Esto vale tanto para bebés que pesan 600 g., 900 g. o 1200 g. al nacer. Cuando hay interacción, esta interpretación se dificulta.*

Figura 56: Variable respuesta `Y` ajustada en función de la `drogaA`, controlando por `drogaB`, para tres posibles valores de `drogaB` (10, 50 y 90) en el modelo con interacción, para los datos de `ejemplointer3.txt`.



Ejercicio 4.2 *Hacer el ejercicio 3 del Taller 3.*

Ejercicio 4.3 *Hacer el ejercicio 4 del Taller 3.*

4.18. Interacción entre dos variables cualitativas

Finalmente restaría presentar un modelo de regresión lineal con interacción entre dos variables cualitativas. Retomemos el ejemplo del pulso post ejercicio.

Ejemplo 4.4 *A cuarenta personas se les miden el pulso antes y después de ejercitar, junto con otras covariables. Estos datos fueron presentados en el Ejemplo 4.1. Para cada individuo, se midieron las siguientes variables*

$$
\begin{aligned}
Y &= \textit{pulso luego de correr una milla } (\texttt{Pulso2}) \\
X_2 &= \begin{cases} 1 & \textit{si la persona es mujer} \\ 0 & \textit{en caso contrario} \end{cases} \\
X_3 &= \begin{cases} 1 & \textit{si la persona fuma} \\ 0 & \textit{en caso contrario} \end{cases}
\end{aligned}
$$

Antes de presentar el modelo con interacción, proponemos un modelo aditivo para explicar el pulso post-ejercicio, en función de las covariables X_2 y X_3. Tanto el sexo como la condición de fumador son variables dummies o binarias. En la base de datos se las denomina $X_2 = \texttt{mujer}$ y $X_3 = \texttt{fuma}$.

El modelo (aditivo) es

$$E\left(Y \mid X_2, X_3\right) = \beta_0 + \beta_M \texttt{mujer} + \beta_F \texttt{fuma}. \tag{83}$$

Hemos puesto el subíndice de los beta de acuerdo a la variable explicativa que acompañan. En la Tabla 46 escribimos el significado del modelo para las cuatro combinaciones posibles de los valores de X_2 y X_3.

Tabla 46: Modelo de regresión lineal múltiple aditivo para el pulso post-ejercicio con covariables $X_2 = \texttt{mujer}$ y $X_3 = \texttt{fuma}$.

Grupo	$X_2 =$ mujer	$X_3 =$ fuma	$E\left(Y \mid X_2, X_3\right)$
1	0	0	β_0
2	0	1	$\beta_0 + \beta_F$
3	1	0	$\beta_0 + \beta_M$
4	1	1	$\beta_0 + \beta_F + \beta_M$

En la Tabla 46 vemos que β_F representa el aumento (o disminución, según el signo) en el pulso medio post ejercicio al comparar el grupo de hombres fumadores con

el grupo de hombres no fumadores (grupo 2 menos grupo 1), pues

$$\begin{aligned} & E\left(Y \mid \texttt{mujer} = 0, \texttt{fuma} = 1\right) - E\left(Y \mid \texttt{mujer} = 0, \texttt{fuma} = 0\right) \\ = \; & (\beta_0 + \beta_F) - \beta_0 \\ = \; & \beta_F \end{aligned}$$

y también representa el cambio en el pulso medio post ejercicio al comparar el grupo de mujeres fumadoras con el de las mujeres no fumadoras (grupo 4 menos grupo 3), pues

$$\begin{aligned} & E\left(Y \mid \texttt{mujer} = 1, \texttt{fuma} = 1\right) - E\left(Y \mid \texttt{mujer} = 1, \texttt{fuma} = 0\right) \\ = \; & (\beta_0 + \beta_F + \beta_M) - (\beta_0 + \beta_M) \\ = \; & \beta_F. \end{aligned}$$

Como ambas diferencias dan el mismo número, decimos que β_F representa el cambio en el valor esperado del pulso post-ejercicio por efecto de fumar, cuando se controla (o estratifica) por la variable sexo, o sea, cuando mantenemos la otra variable fija sin importar su valor. Observemos que esta es la misma interpretación que hemos hecho de los coeficientes en los modelos de regresión lineal aditivos. Del mismo modo, β_M representa la diferencia en el pulso medio post-ejercicio entre mujeres y varones, al controlar por la variable `fuma`.
Observemos que este modelo dispone de tres coeficientes β_0, β_M y β_F para reflejar las medias de cuatro grupos distintos.
¿Cómo es el modelo que explica a Y con X_2, X_3 y la interacción entre ambas? El modelo es el siguiente

$$E\left(Y \mid X_2, X_3\right) = \beta_0 + \beta_M \texttt{mujer} + \beta_F \texttt{fuma} + \beta_{M:F} \texttt{mujer} \cdot \texttt{fuma}. \tag{84}$$

Como tanto $X_2 = \texttt{mujer}$ y $X_3 = \texttt{fuma}$ son variables dicotómicas, el término producto $X_2 \cdot X_3 = \texttt{mujer} \cdot \texttt{fuma}$ también resulta ser una variable indicadora o dicotómica, en este caso

$$X_2 \cdot X_3 = \begin{cases} 1 & \text{si la persona es mujer y fuma} \\ \\ 0 & \text{en caso contrario.} \end{cases}$$

Nuevamente, en la Tabla 47, escribimos el significado del modelo para las cuatro combinaciones posibles de los valores de $X_2 = \texttt{mujer}$ y $X_3 = \texttt{fuma}$.
Hagamos las mismas comparaciones que hicimos en el modelo aditivo. Comparamos el valor medio de la variable respuesta del grupo 2 con el del grupo 1:

$$\begin{aligned} & E\left(Y \mid \texttt{mujer} = 0, \texttt{fuma} = 1\right) - E\left(Y \mid \texttt{mujer} = 0, \texttt{fuma} = 0\right) \\ = \; & (\beta_0 + \beta_F) - \beta_0 \\ = \; & \beta_F \end{aligned}$$

Tabla 47: Modelo de regresión lineal múltiple con interacción, para el pulso post-ejercicio con covariables $X_2 = \texttt{mujer}$ y $X_3 = \texttt{fuma}$.

Grupo	$X_2 =$ mujer	$X_3 =$ fuma	$X_2 \cdot X_3$	$E(Y \mid X_2, X_3)$
1	0	0	0	β_0
2	0	1	0	$\beta_0 + \beta_F$
3	1	0	0	$\beta_0 + \beta_M$
4	1	1	1	$\beta_0 + \beta_F + \beta_M + \beta_{M:F}$

Ahora comparemos los valores medios de la respuesta en los grupos 4 y 3:

$$\begin{aligned} & E(Y \mid \texttt{mujer} = 1, \texttt{fuma} = 1) - E(Y \mid \texttt{mujer} = 1, \texttt{fuma} = 0) \\ = \; & (\beta_0 + \beta_M + \beta_F + \beta_{M:F}) - (\beta_0 + \beta_M) \\ = \; & \beta_F + \beta_{M:F}. \end{aligned}$$

Por lo tanto, β_F mide el efecto de fumar en los hombres, y $\beta_F + \beta_{M:F}$ mide el efecto de fumar en las mujeres. De modo que el término de la interacción $\beta_{M:F}$ da la diferencia del pulso medio post-ejercicio por efecto de fumar de las mujeres respecto de los hombres. Si $\beta_{M:F} > 0$, el hecho de fumar en las mujeres redunda en un aumento de la respuesta media respecto de la de los hombres. Un test de $H_0 : \beta_{M:F} = 0$ versus $H_1 : \beta_{M:F} \neq 0$ para el modelo (84) es una prueba para la igualdad del efecto de fumar en el pulso medio post-ejercicio de hombres y mujeres. Observemos que si no se rechaza H_0, tenemos un modelo aditivo: el efecto de fumar en el pulso post-ejercicio resulta ser el mismo para hombres y mujeres.
También se podrían tomar diferencias análogas entre los grupos 1 y 3 (no fumadores) y entre los grupos 4 y 2 (fumadores) y llegar a la misma interpretación de la interacción. En este ejemplo, esta aproximación parece menos intuitiva, ya que interesa evaluar el efecto de fumar (controlando por el sexo) en la respuesta.
Antes de pasar a los ajustes de los modelos, propongamos un modelo de comparación de las medias de cuatro muestras aleatorias normales, todas con la misma varianza (o sea, una generalización del test de t para de dos muestras). Tal modelo, propondría que se tienen 4 muestras de pulso post-ejercicio tomadas en 4 grupos diferentes (en este caso, los definidos en la primer columna de la Tabla 47) y para cada uno de ellos proponemos

$$\begin{aligned} Y_{i1} &\sim N(\mu_1, \sigma^2) \quad (1 \leq i \leq n_1) \quad \text{grupo 1 (hombres no fumadores)} \\ Y_{i2} &\sim N(\mu_2, \sigma^2) \quad (1 \leq i \leq n_2) \quad \text{grupo 2 (hombres fumadores)} \\ Y_{i3} &\sim N(\mu_3, \sigma^2) \quad (1 \leq i \leq n_3) \quad \text{grupo 3 (mujeres fumadoras)} \\ Y_{i4} &\sim N(\mu_4, \sigma^2) \quad (1 \leq i \leq n_4) \quad \text{grupo 4 (mujeres no fumadoras).} \end{aligned} \tag{85}$$

Todas las observaciones son independientes entre sí. Este modelo propone ajustar 4 parámetros que dan cuenta de la media (uno para cada grupo, que hemos denominado μ_k que se estimarán con las observaciones del respectivo grupo $k-$ ésimo) y un parámetro que da cuenta de la varianza de cada observación en el modelo homoscedástico (σ^2 que se estimará de forma conjunta con todas las $n_1+n_2+n_3+n_4$ observaciones). Si comparamos este modelo con el propuesto en (84), vemos que ambos tienen 4 parámetros para las medias. Más aún, resultará que se vinculan de la siguiente forma, por lo desarrollado en la Tabla 47.

$$\begin{aligned} \mu_1 &= \beta_0 \\ \mu_2 &= \beta_0 + \beta_F \\ \mu_3 &= \beta_0 + \beta_M \\ \mu_4 &= \beta_0 + \beta_F + \beta_M + \beta_{F:M}. \end{aligned} \tag{86}$$

Otra forma de escribir el modelo (85) es la siguiente

$$\begin{aligned} Y_{i1} &= \mu_1 + \epsilon_{i1} \quad (1 \leq i \leq n_1) \qquad \text{grupo 1 (hombres no fumadores)} \\ Y_{i2} &= \mu_2 + \epsilon_{i2} \quad (1 \leq i \leq n_2) \qquad \text{grupo 2 (hombres fumadores)} \\ Y_{i3} &= \mu_3 + \epsilon_{i3} \quad (1 \leq i \leq n_3) \qquad \text{grupo 3 (mujeres fumadoras)} \\ Y_{i4} &= \mu_4 + \epsilon_{i4} \quad (1 \leq i \leq n_4) \qquad \text{grupo 4 (mujeres no fumadoras),} \end{aligned} \tag{87}$$

donde los $\epsilon_{ik} \sim N\left(0, \sigma^2\right)$ y son todos independientes

Vemos pues que ambos modelos (84) y (85) son equivalentes, ya que conociendo los parámetros de uno de ellos (los μ_k por ejemplo) podemos despejar los valores del otro (los β_h por ejemplo) por medio de las ecuaciones (86). O al revés, obtener los μ_k a partir de los β_h. La varianza del error se estimará en forma conjunta en ambos modelos. La diferencia está en el significado de los parámetros. En el modelo (85), μ_k representa el valor esperado de la variable respuesta en el grupo $k-$ésimo, mientras que en el modelo (84) los β_h representan (algunas de) las diferencias entre los valores de las respuestas medias entre los distintos grupos.
En las Tablas 48 y 49 se muestran los valores ajustados de los modelos aditivos (83) y con interacción (84).
Analicemos primero el modelo con interacción. En la salida vemos que el coeficiente de la interacción no resulta significativo (el p$-$valor es 0,245 que no es menor a 0,05), por lo tanto concluimos que el efecto de fumar en el pulso medio post-ejercicio de mujeres y varones es el mismo. Luego, para los datos del pulso el modelo apropiado es el aditivo (83). En dicho ajuste vemos que todos los coeficientes son significativos, y que el hecho de fumar aumenta el pulso post-ejercicio en 7,36 pulsaciones por minuto, cuando uno controla por sexo. Es interesante graficar

Tabla 48: Ajuste del modelo lineal múltiple aditivo $Y_i = \beta_0 + \beta_M X_{i2} + \beta_F X_{i3} + \varepsilon_i$, donde X_2 = indicador de mujer (`mujer`), X_3 = indicador de fumar (`fuma`), e Y = pulso post ejercicio (`Pulso2`).

```
> ajusteA<-lm(Pulso2 ~ mujer + fuma)
> summary(ajusteA)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  126.926      2.452  51.754  < 2e-16
mujer         18.064      3.027   5.967 6.96e-07
fuma           7.362      3.074   2.395   0.0218
---

Residual standard error: 9.453 on 37 degrees of freedom
Multiple R-squared: 0.5093,     Adjusted R-squared: 0.4828
F-statistic:  19.2 on 2 and 37 DF,  p-value: 1.906e-06
```

las cuatro medias muestrales y los cuatro valores esperados bajo el modelo. Esos valores figuran en la Tabla 50.
Mirando la Tabla 50 podemos corroborar que los estimadores obtenidos con el modelo con interacción son los mismos que obtendríamos si estimáramos las medias de cada grupo por separado. En este caso además, vemos que el ajuste obtenido por el modelo sin interacción no difiere demasiado del con interacción, en sus valores ajustados, es por eso que la interacción no resulta significativa en este modelo. El Gráfico 57 permite visualizar más claramente la situación. En él vemos que al pasar del grupo de no fumadores al grupo de fumadores, aumenta el pulso medio post-ejercicio, tanto en hombres como en mujeres, siempre en una cantidad parecida (tan parecida, que la diferencia entre ambos no es estadísticamente significativa). Este gráfico suele llamarse gráfico de interacción. Sirve para evaluar si tiene sentido ajustar un modelo con interacción a nuestros datos. Si dicho gráfico resultara como se muestra en alguno de los dos de la Figura 58, entonces se justificaría agregar el término de interacción al modelo con dos covariables categóricas. En el gráfico A vemos un ejemplo donde al pasar del grupo no fumador al grupo fumador, para las mujeres se produce un aumento de la respuesta media, y para los hombres una disminución de la respuesta media. Para este ejemplo, tiene sentido incluir el término de la interacción, ya que la respuesta cambia de sentido para distintas combinaciones de las dos explicativas. En el gráfico B sucede algo parecido: cuando controlamos por el sexo de la persona, el efecto de fumar es diferente en los dos

Tabla 49: Ajuste del modelo lineal múltiple con interacción $Y_i = \beta_0 + \beta_M X_{i2} + \beta_F X_{i3} + \beta_{M:F} X_{i2} \cdot X_{i3} + \varepsilon_i$, donde $X_2 =$ indicador de mujer (`mujer`), $X_3 =$ indicador de fumar (`fuma`), $Y =$ pulso post ejercicio (`Pulso2`).

```
> ajusteB <-lm(Pulso2 ~ mujer * fuma)
> summary(ajusteB)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  128.333      2.714  47.280  < 2e-16
mujer         15.250      3.839   3.973 0.000326
fuma           4.267      4.026   1.060 0.296306
mujer:fuma     7.317      6.190   1.182 0.244922
---

Residual standard error: 9.403 on 36 degrees of freedom
Multiple R-squared: 0.5276,     Adjusted R-squared: 0.4883
F-statistic:  13.4 on 3 and 36 DF,  p-value: 4.978e-06
```

grupos, para las mujeres aumenta la media de la respuesta, para los hombres la deja igual.

4.19. Generalización a más de dos variables.

Cuando el número de variables regresoras es mayor que dos, se pueden incluir términos de interacción para cada par de covariables. Por ejemplo, en un modelo con tres variables regresoras X_1, X_2 y X_3, podemos tener:

$$\begin{aligned} E(Y \mid X_1, X_2, X_3) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\ &\quad + \beta_{1:2} X_1 \cdot X_2 + \beta_{1:3} X_1 \cdot X_3 + \beta_{2:3} X_2 \cdot X_3 \end{aligned}$$

En este modelo hemos considerado las interacciones de las variables tomadas de a pares, a las que se denomina interacciones de segundo orden. Pero podríamos haber considerado además la interacción de tercer orden incorporando un término $\beta_{1:2:3} X_1 \cdot X_2 \cdot X_3$.
A partir de los tests de significación podremos evaluar si alguna(s) de estas interacciones son necesarias en el modelo.

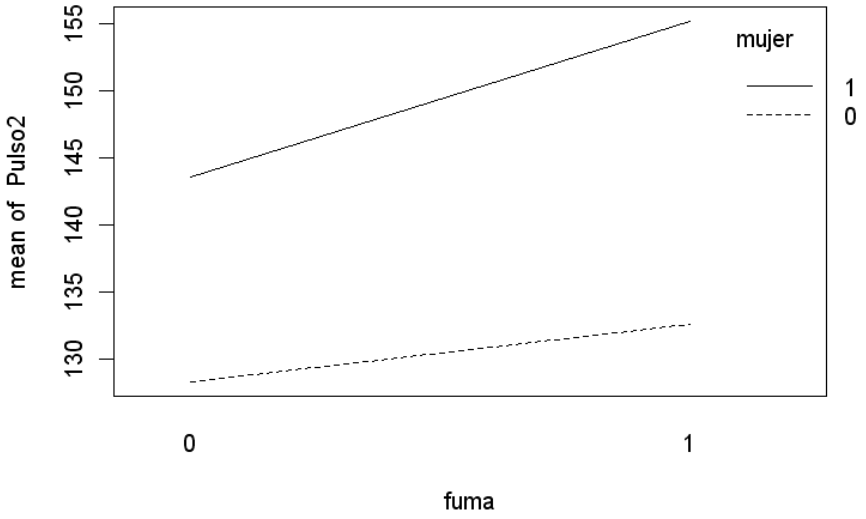
Tabla 50: Medias muestrales calculadas por grupos, comparadas con el ajuste de los modelos sin y con interacción, para el pulso post-ejercicio con covariables $X_2 =$ `mujer` y $X_3 =$ `fuma`.

Grupo	X_2	X_3	Media muestral	$E(Y \mid X_2, X_3)$ sin interacción
1	0	0	128,3333	$\widehat{\beta}_0 = 126{,}926$
2	0	1	132,6	$\widehat{\beta}_0 + \widehat{\beta}_F = 126{,}926 + 7{,}362 = 134{,}29$
3	1	0	143,5833	$\widehat{\beta}_0 + \widehat{\beta}_M = 126{,}926 + 18{,}064 = 144{,}99$
4	1	1	155,1667	$\widehat{\beta}_0 + \widehat{\beta}_F + \widehat{\beta}_M = 126{,}926 + 7{,}362$ $+18{,}064 = 152{,}35$

Grupo	X_2	X_3	Media muestral	$E(Y \mid X_2, X_3)$ con interacción
1	0	0	128,3333	$\widehat{\beta}_0 = 128{,}333$
2	0	1	132,6	$\widehat{\beta}_0 + \widehat{\beta}_F = 128{,}333 + 4{,}267 = 132{,}6$
3	1	0	143,5833	$\widehat{\beta}_0 + \widehat{\beta}_M = 128{,}333 + 15{,}25 = 143{,}58$
4	1	1	155,1667	$\widehat{\beta}_0 + \widehat{\beta}_F + \widehat{\beta}_M + \widehat{\beta}_{F:M} = 128{,}333 + 4{,}267$ $+15{,}25 + 7{,}317 = 155{,}1667$

Cuando alguna interacción es significativa y el modelo debe incluir estos términos, es más compleja la presentación de los resultados. Una aproximación posible es graficar una colección de rectas como en las figuras anteriores, para describir gráficamente cómo cambia la relación con los valores de las demás variables.

Figura 57: Gráfico de las medias muestrales de los cuatro grupos, de los datos de pulso-post ejercicio.

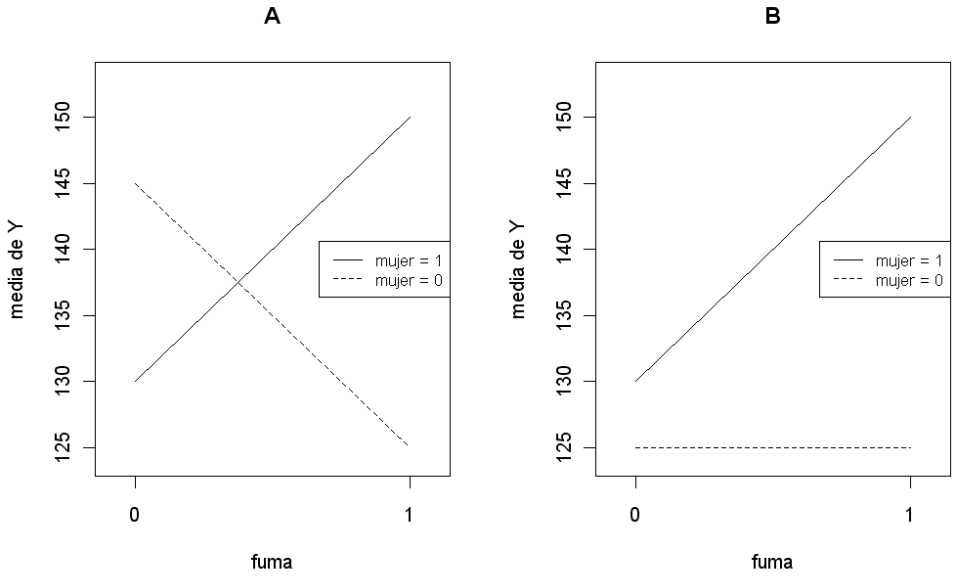


5. Diagnóstico del modelo

5.1. Diagnóstico del modelo: definiciones y gráficos

Los métodos de inferencia anteriormentes descriptos (cálculo de p-valores, intervalos de confianza y de predicción, por citar algunos) requieren que se satisfagan los cuatro supuestos (45) que subyacen a nuestro modelo. El diagnóstico del modelo consiste en la validación de estos supuestos para los datos en cuestión. Esta validación puede hacerse a través de una serie de gráficos, muchos de los cuales ya describimos en la regresión lineal simple, o bien a través de diversos cálculos. El diagnóstico desempeña un papel importante en el desarrollo y la evaluación de los modelos regresión múltiple. La mayoría de los procedimientos de diagnóstico para la regresión lineal simple que hemos descrito anteriormente se trasladan directamente a la regresión múltiple. A continuación revisaremos dichos procedimientos de diagnóstico.
Por otro lado, también se han desarrollado herramientas de diagnóstico y procedimientos especializados para la regresión múltiple. Algunas de las más importantes se discuten en la Sección 5.2.

Figura 58: Gráficos de las medias de una variable respuesta Y para dos ejemplos ficticios, en las figuras A y B.

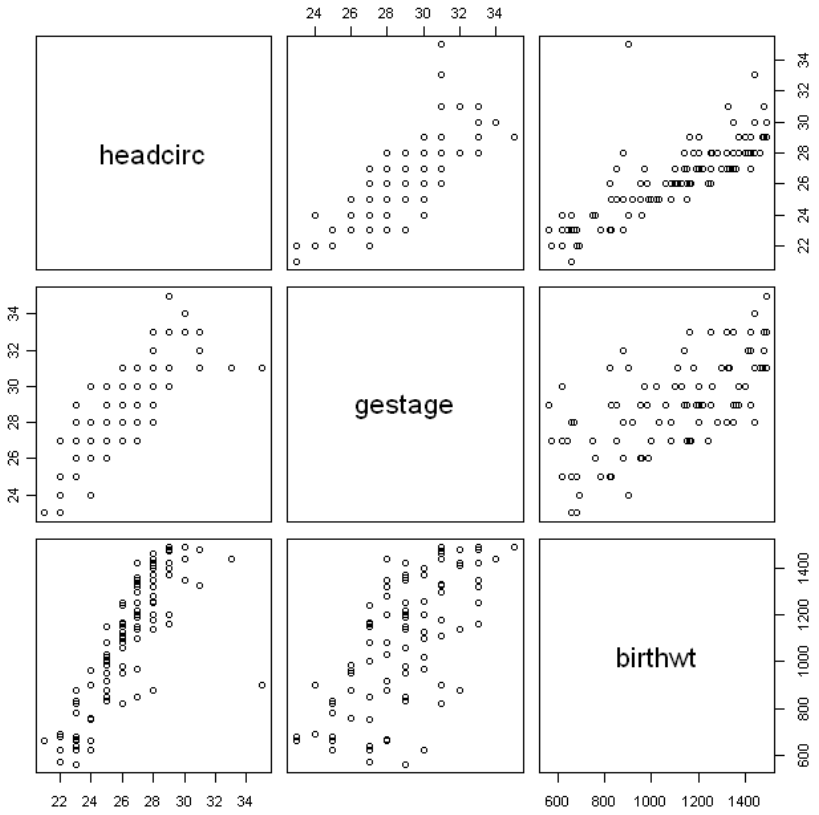


5.1.1. Matriz de scatter plots o gráficos de dispersión

Los boxplots, histogramas, diagramas de tallo y hojas, y gráficos de puntos para cada una de las variables predictoras y para la variable de respuesta pueden proporcionar información univariada preliminar y útil sobre estas variables. Los diagramas de dispersión (scatter plots) de la variable de respuesta versus cada variable predictora pueden ayudar a determinar la naturaleza y la fuerza de las relaciones bivariadas entre cada una de las variables de predicción y la variable de respuesta así como pueden permitir la identificación de lagunas en las regiones de datos. También pueden permitir identificar outliers u observaciones atípicas o alejadas del patrón del resto de los datos. Los diagramas de dispersión de cada variable predictora versus cada una de las otras variables de predicción son útiles para el estudio de las relaciones bivariadas entre las distintas variables predictoras y también para buscar espacios con ausencia de datos y detectar valores atípicos. El análisis resulta más fácil si los gráficos de dispersión se ensamblan en una matriz diagrama de dispersión (*scatter plot matrix*), como vemos en la Figura 59. En esta figura, la variable graficada en el eje vertical para cualquier gráfico de dispersión es aquella cuyo nombre se encuentra en su fila, y la variable graficada en el eje

horizontal es aquella cuyo nombre se encuentra en su columna. Por lo tanto, la matriz de gráfico de dispersión en la Figura 59 muestra en la primera fila los gráficos de Y (perímetro cefálico: `headcirc`) versus X_1, (edad gestacional: `gestage`) y de Y versus X_2 (peso: `birthwt`). En la segunda fila tenemos los gráficos de X_1 versus Y y de X_1 versus X_2. Finalmente, en la tercer fila tenemos los gráficos de X_2 versus Y y de X_2 versus X_1. Una matriz de diagramas de dispersión facilita el estudio de las relaciones entre las variables mediante la comparación de los diagramas de dispersión dentro de una fila o una columna. Esta matriz muestra, por supuesto, información repetida de los datos. Bastaría con dar los scatter plots que quedan por encima (o bien, por debajo) de la diagonal.
Si el dataframe que contiene a los datos se denomina `low`, la matriz de scatterplots se realiza con `pairs(low)`, en `R`.

Figura 59: Matriz de scatter plots para los datos de bebés con bajo peso, con las covariables edad gestacional y peso



Un complemento a la matriz de diagramas de dispersión que puede ser útil a veces es la matriz de correlaciones. Esta matriz contiene los coeficientes de correlación simple $r_{YX_1}, r_{YX_2}, ..., r_{YX_{p-1}}$ entre Y y cada una de las variables predictoras, así como todos los coeficientes de correlación simple entre las distintas variables predictoras entre sí. El formato de la matriz de correlación sigue el de la matriz de scatter plots

$$\begin{bmatrix} 1 & r_{YX_1} & r_{YX_2} & \cdots & r_{YX_{p-1}} \\ r_{YX_1} & 1 & r_{X_1X_2} & \cdots & r_{X_1X_{p-1}} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{YX_{p-1}} & r_{X_1X_{p-1}} & r_{X_2X_{p-1}} & \cdots & 1 \end{bmatrix}$$

y en el caso de los datos de bebés de bajo peso es

```
> cor(low)
          headcirc   gestage   birthwt
headcirc 1.0000000 0.7806919 0.7988372
gestage  0.7806919 1.0000000 0.6599376
birthwt  0.7988372 0.6599376 1.0000000
```

Observemos que la matriz de correlación es simétrica y en la diagonal contiene unos pues el coeficiente de correlación de una variable consigo misma es 1.

5.1.2. Gráficos de dispersión en tres dimensiones

Algunos paquetes estadísticos proporcionan gráficos de dispersión en tres dimensiones y permiten girar estos gráficos para permitir al usuario ver la nube de puntos desde diferentes perspectivas. Esto puede ser muy útil para identificar los patrones que sólo se desprenden de la observación desde ciertas perspectivas. En `R` la instrucción `plot3d` de la librería `rgl` permite hacerlo. Incluso se puede rotar el gráfico y guardar la película de la rotación con la instrucción `movie3d`.

5.1.3. Gráficos de residuos

Es importante recalcar que aunque las observaciones $(X_{i1}, X_{i2}, \ldots, X_{i(p-1)}, Y)$ no puedan graficarse en el caso de tener más de dos covariables, siempre tanto el valor predicho o ajustado $\widehat{Y}_i$ como el residuo e_i están en $\mathbb{R}$ y pueden ser graficados. De modo que un gráfico de los residuos contra los valores ajustados es útil para evaluar la idoneidad de la regresión múltiple para modelar los datos observados, y la homoscedasticidad (constancia de la varianza) de los términos de error. También

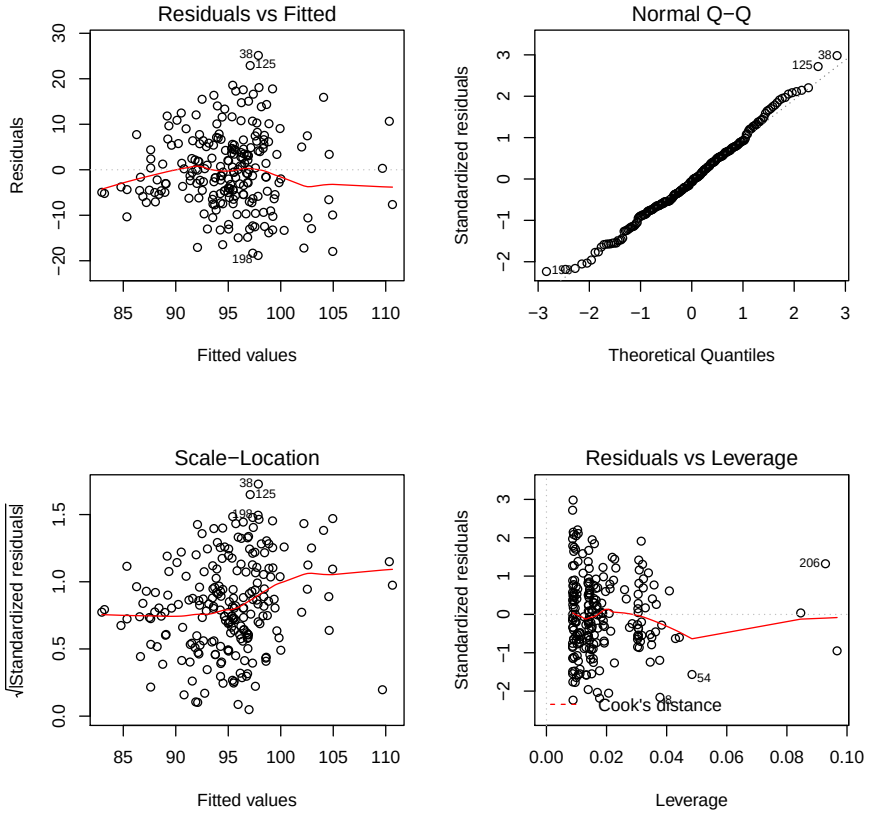
permite proporcionar información acerca de los valores extremos como lo vimos en la regresión lineal simple. Del mismo modo, un gráfico de los residuos contra el tiempo (u orden en el que fueron recopilados los datos, si este fuera relevante) o en contra de otra secuencia puede proporcionar información acerca de las posibles correlaciones entre los términos de error en la regresión múltiple. Boxplots y gráficos de probabilidad normal de los residuos son útiles para examinar si el supuesto de distribución normal sobre los términos de error se satisface razonablemente para los valores observados.
Además, los residuos deben ser graficados versus cada una de las variables predictivas. Cada uno de estos gráficos pueden proporcionar más información sobre la idoneidad de la función de regresión con respecto a la variable de predicción (por ejemplo, si un efecto que dé cuenta de la curvatura es necesario para dicha variable) y también puede proporcionar información sobre la posible variación de la varianza del error en relación con dicha variable predictora.
Los residuos también deben ser graficados versus cada una de las variables de predicción importantes que se omitieron del modelo, para ver si las variables omitidas tienen importantes efectos adicionales sobre la variable de respuesta que aún no han sido reconocidos en el modelo de regresión. Además, los residuos deben graficarse versus términos de interacción para los posibles efectos no incluidos en el modelo de regresión (trabajamos con interacción en la Sección 4.16 y subsiguientes), para ver si es necesario incluir algún término de interacción en el modelo.
Un gráfico de los residuos o los residuos al cuadrado contra los valores ajustados es útil para examinar si la varianza de los términos de error es constante. Si se detecta que la varianza no es constante, suele ser apropiado realizar gráficos del valor absoluto de los residuos, o de sus cuadrados, versus cada una de las variables predictoras. Estos gráficos pueden permitir la identificación de una o más variables predictoras con las que se relaciona la magnitud de la variabilidad del error. Se puede incluir una transformación de esta variable en el modelo para tratar de eliminar la estructura observada en los residuos.
Por supuesto, cualquier paquete estadístico realizará estos gráficos de manera más o menos automática. R produce automáticamente cuatro gráficos de diagnóstico cuando uno aplica la función `plot()` directamente a una salida del `lm()`. En general, este comando produce un gráfico por vez, apretando enter en el teclado producirá un gráfico por vez. Para ver los cuatro juntos, una opción es dividir la ventana gráfica en cuatro con la instrucción `mfrow`:

```
par(mfrow=c(2,2))
plot(ajuste4)
```

Alternativamente, uno puede calcular los residuos a partir de un ajuste lineal con las instrucciones `residuals()`, `rstudent()` para los residuos estudentizados,

`rstandard()` para los estandarizados, `fitted.values()` para los valores predichos. Y luego graficarlos usando la función `plot`. En la Figura 60 vemos el `plot` de los residuos para el `ajuste4` propuesto en el modelo (73), página 173.

Figura 60: Gráficos de diagnóstico producidos por la instrucción plot aplicada a la salida lm. Se grafican, los residuos versus los valores ajustados, el Q-Q plot normal de los residuos estandarizados, la raíz cuadrada de los residuos versus los valores ajustados, y finalmente los residuos estandarizados versus los leverage.



5.2. Identificación de outliers y puntos de alto leverage

Como en el caso de regresión lineal simple, aquí también tiene sentido identificar las observaciones que no siguen el patrón de los demás datos. Medidas para identificar estas observaciones son, nuevamente, el leverage, los residuos estudentizados, y algunas que no estudiaremos aquí como los DFFITS y los DFBETAS.

5.2.1. Ajuste robusto: permite ignorar a los outliers automáticamente

Como vimos para regresión lineal simple, el problema de las observaciones atípicas en un conjunto de datos es que su presencia puede influir de manera dramática sobre el ajuste del modelo propuesto, incluso llegando a tergiversar por completo las conclusiones que se pueden extraer de él cuando el ajuste se lleva a cabo usando estimadores de mínimos cuadrados. Esta distorsión de los valores ajustados además tiene la potencia de enmascarar las observaciones atípicas y muchas veces disimularlas entre los datos, dificultando el buen funcionamiento de las herramientas de detección de atipicidad más difundidas en el área: leverage, distancias de Cook, dfits, etc. Como ya dijimos en la Sección 3.2.4, una forma automática de evitar estos problemas consiste en cambiar el método de estimación de mínimos cuadrados por un ajuste robusto de los coeficientes. Los MM-estimadores de regresión son una buena alternativa. El ajuste que presentamos en la Sección 3.2.4 a traves de la rutina `lmrob` de la librería `robustbase` se extiende trivialmente para el caso de regresión lineal múltiple. A modo de ejemplo, en la Tabla 51 vemos el ajuste de dicha rutina a los datos del archivo `azucar` considerados previamente, el mismo ajuste de mínimos cuadrados figura en la Tabla 36.
Si el ajuste robusto y el clásico (o sea el de mínimos cuadrados) no difieren, esta es una señal de que no hay observaciones atípicas en el conjunto de datos con el que se trabaja. Como el ajuste por mínimos cuadrados es el más difundido en estadística, cuando el interés del análisis incluya la comunicación de los resultados a otros especialistas, es recomendable reportar la salida obtenida con el ajuste clásico. Esto es lo que sucede para el ajuste de los datos de `azucar`.
En cambio, cuando el ajuste robusto y el clásico difieren entre sí, esto se deberá a la presencia de datos atípicos. Con el ajuste robusto estos se podrán detectar claramente: corresponderán a aquellas observaciones cuyos pesos (robustos) asignados por el ajuste del `lmrob` sean muy chicos (pesos cero o muy cercanos a él). Estos se calculan como `ajusterob$rweights` para el ajuste presentado en la Tabla 51. Como los pesos, que van entre 0 y 1, se calculan en función de los residuos, un criterio equivalente será investigar aquellas observaciones con residuos grandes del ajuste robusto. En el caso de los datos de `azucar`, vemos que el menor peso corresponde a 0,24. La instrucción `plot` aplicada al ajuste robusto dado por `lmrob` proporciona 5 gráficos que permite visualizar las observaciones extremas, basadas en el cómputo del leverage robusto (*robust distances*) y que pueden verse en la Figura 61 para el ajuste robusto de los datos de `azucar`. Una descripción más detallada de los estimadores robustos disponibles puede consultarse en los Capítulos 4 y 5 de Maronna et al. [2006].

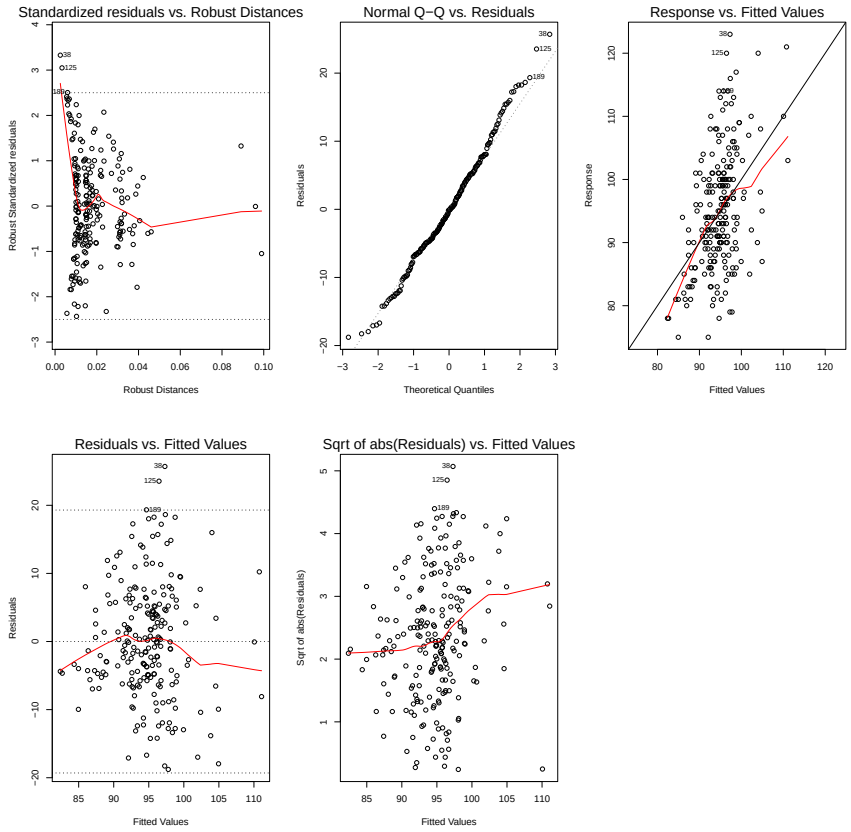
Tabla 51: Ajuste de un MM-estimador de regresión para el modelo con `glucosa` como variables respuesta y `peso.evo` (categórica) y `bmi` (numérica), archivo de datos `azucar`. El ajuste clásico puede verse en la Tabla 36, página 177.

```
> library(robustbase)
> Ievo<-factor(peso.evo)
> ajusterob <- lmrob(glucosa ~ bmi + Ievo, data = azucar)
> summary(ajusterob)
 \--> method = "MM"
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  70.6589     3.7731  18.727  < 2e-16 ***
bmi           0.6552     0.1356   4.832 2.56e-06 ***
Ievo2         7.2255     1.3691   5.278 3.18e-07 ***
Ievo3         7.9919     1.4531   5.500 1.07e-07 ***
---
Robust residual standard error: 7.721
Multiple R-squared:  0.2222,        Adjusted R-squared:  0.2114
Convergence in 12 IRWLS iterations

Robustness weights:
 18 weights are ~= 1. The remaining 202 ones are summarized as
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2452  0.8557  0.9502  0.8921  0.9815  0.9990
```

5.2.2. Leverage

Vimos en la Observación 4.5, en la Sección 4.6 que los residuos no son homoscedásticos. Y además vimos que la varianza dependía del leverage de una observación, que también definimos en esa sección a partir de la matriz de proyección o "*hat matrix*" H. El leverage de la i-ésima observación será el elemento h_{ii} de la matriz de proyección, y en general será calculado por el software. En el caso de regresión múltiple, sin embargo, es mucho más importante asegurarse que no haya observaciones potencialmente influyentes, o si uno sospecha de algunas, estudiar cómo cambia el ajuste cuando esa observación es eliminada de la base de datos. Para la detección de observaciones potencialmente influyentes en regresión lineal simple, muchas veces basta mirar con cuidado el scatter plot de los datos. El problema que aparece aquí es que no podemos, en general, dibujar el scatter plot de los datos, por lo que tendremos que calcular el leverage de cada observación. El criterio para

Figura 61: Salida de `plot(ajusterob)` para los datos de `azucar`, cuyo ajuste figura en la Tabla 51.



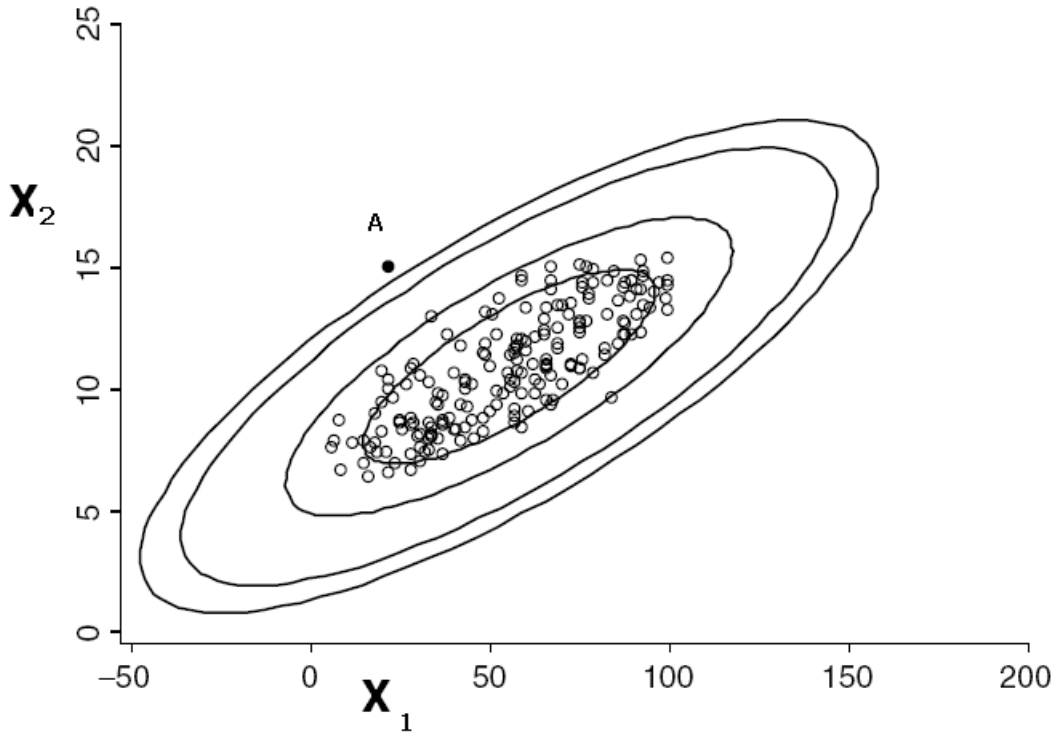
encontrar observaciones potencialmente influyentes será la extensión del visto anteriormente. El leverage alto indica que una observación no sigue el patrón de las demás covariables X. Nuevamente se tiene

$$0 \leq h_{ii} \leq 1 \qquad \sum_{i=1}^{n} h_{ii} = p$$

donde p es el número de parámetros de regresión (betas) que hay en la función de regresión, incluyendo el término de intercept. Puede mostrarse que h_{ii} es una medida de la distancia entre los valores de las covariables X de la i-ésima observación respecto del valor del promedio de todas las X observadas en los n casos. Es lo que se conoce como **distancia de Mahalanobis** de la i-ésima observación $\left(X_{i1}, X_{i2}, \ldots, X_{i(p-1)}\right)$ cuando se tiene una muestra de ellas. Este concepto se es-

tudia en detalle en los cursos de análisis multivariado. La idea subyacente es que las distancia usual no expresa bien las distancias entre observaciones cuando hay dependencia entre las covariables, entonces esta correlación o dependencia se toma en cuenta para definir una nueva noción de distancia entre puntos. En la Figura 62 se ve un gráfico de dispersión para un conjunto de observaciones, con curvas superpuestas. Estas curvas representan los puntos que tienen el mismo leverage. Vemos que son elipses. En el gráfico hay una observación alejada, indicada con A.

Figura 62: Contornos de leverage constante en dos dimensiones. Las elipses más pequeñas representan un menor leverage. Vemos una observación identificada con el nombre A que tiene alto leverage y no sigue el patrón de las restantes. Fuente: Weisberg [2005], pág. 170



Los dos criterios para evaluar si una observación tiene alta palanca presentados en el caso de regresión lineal simple se extienden sin grandes modificaciones al caso múltiple. Ellos son

1. *(Ajustado por la cantidad de covariables)* Declarar a la observación i-ésima con alto leverage si $h_{ii} > 2\overline{h} = \frac{2}{n}\sum_{j=1}^{n} h_{jj} = \frac{2p}{n}$.

2. *(Sin ajustar por la cantidad de covariables)* Declarar a la observación i-ésima con muy alto leverage si $h_{ii} > 0{,}5$ y con leverage moderado si $0{,}2 < h_{ii} \leq 0{,}5$.

Una evidencia adicional para declarar que una cierta observación tiene un leverage notoriamente alto, consiste en graficar un histograma de los h_{ii} y ver si existe una brecha destacable que separa al mayor leverage o a un pequeño conjunto de mayores leverages del resto de las observaciones.

5.2.3. Uso de la matriz de proyección para identificar extrapolaciones

La matriz H de proyección también es útil para determinar si una inferencia respecto de la respuesta media o de la predicción para una nueva observación X_{nueva} de valores de las predictoras involucra una extrapolación sustancial respecto del rango de los valores observados. Cuando sólo tenemos dos predictoras X_1 y X_2 esto puede resolverse con un scatter plot como muestra la Figura 62. Este sencillo análisis gráfico no se encuentra disponible si $p \geq 3$, donde las extrapolaciones pueden ocultarse.
Para detectarlas podemos utilizar los cálculos de leverage presentados anteriormente. Para una nueva combinación de variables

$$X_{\text{nue}} = (X_{1\text{nue}}, \ldots, X_{p-1\ \text{nue}})$$

para la que interesa hacer predicción se puede calcular

$$h_{\text{nue}} = X_{\text{nue}}^t \left(\mathbf{X}^t\mathbf{X}\right)^{-1} X_{\text{nue}}$$

donde la matriz $\mathbf{X}$ tiene dimensión $n \times p$ y se armó en base a la muestra con la que se calculó el modelo ajustado, ver (46), en la página 118. Si h_{nue} está bien incluida dentro del rango de leverages observados en el conjunto de datos disponibles, estamos seguros de que no hay extrapolación involucrada. Si, por el contrario, h_{nue} es mucho mayor que los leverages observados, entonces no debería llevarse a cabo la estimación o predicción para esta combinación X_{nue} de covariables.

5.2.4. Residuos estudentizados y distancias de Cook

Ambos estadísticos se definen y calculan del mismo modo que en regresión lineal simple. Las distancia de Cook para la i-ésima observación se define por

$$D_i = \frac{\sum_{j=1}^{n} \left(\widehat{Y}_j - \widehat{Y}_{j(i)}\right)^2}{pMS\text{Res}}$$

donde $\widehat{Y}_j$ es el valor ajustado para la j-ésima observación, cuando se usaron las n observaciones en el ajuste del modelo, y $\widehat{Y}_{j(i)}$ es el valor ajustado para la j-ésima observación, cuando se usaron $n-1$ observaciones en el ajuste del modelo,

todas menos la i-ésima. Esto se repite para cada observación, para poder calcular todas las Distancias de Cook. Afortunadamente, las D_i pueden ser calculadas sin necesidad de ajustar una nueva función de regresión cada vez, en la que se deja una observación distinta afuera del conjunto de datos. Esto es porque puede probarse la siguiente igualdad que permite calcular las distancias de Cook

$$D_i = \frac{e_i^2}{pMS\text{Res}} \left[\frac{h_{ii}}{(1-h_{ii})^2} \right].$$

Observemos que las Distancias de Cook dependen de dos factores:

1. el tamaño del residuo i-ésimo, e_i
2. el leverage i-ésimo, h_{ii}.

Cuanto más grande sean e_i o h_{ii}, mayor será D_i. Luego el i-ésimo caso puede ser influyente por

1. tener un alto residuo e_i y sólo un moderado valor de leverage h_{ii},
2. o bien por tener un alto valor de leverage h_{ii} con sólo un moderado valor de residuo e_i,
3. o bien por tener tanto un alto valor de leverage h_{ii} como un alto valor de residuo e_i.

Los puntos de corte sugeridos para detectar una observación influyente con la Distancia de Cook suelen ser percentiles de la distribución F de Fisher con p grados de libertad en el numerador y $n-p$ en el denominador. Si la $D_i \geq F\,(p, n-p, 0{,}50)$ la observación i-ésima es considerada influyente.
El residuo estudentizado (o estudentizado eliminado) se define por

$$restud_i = \frac{Y_i - \widehat{Y}_{i(i)}}{\frac{MS\text{Res}_{(i)}}{1-h_{ii}}},$$

donde $\widehat{Y}_{i(i)}$ es el valor ajustado para la i-ésima observación, cuando se usaron $n-1$ observaciones en el ajuste del modelo, todas menos la i-ésima y $MS\text{Res}_{(i)}$ es el cuadrado medio de los residuos cuando el caso i-ésimo es omitido en el ajuste de la regresión lineal. Nuevamente, no necesitamos ajustar las regresiones excluyendo los casos de a uno por vez, pues una expresión alternativa para el residuo estudentizado es

$$restud_i = e_i \left[\frac{n-p-1}{SS\text{Res}\,(1-h_{ii}) - e_i^2} \right]^{1/2}$$

Los puntos de corte sugeridos para detectar una observación influyente con el residuo estudentizado están dados por el criterio de Bonferroni y consiste en declarar influyente a una observación si

$$|restud_i| > t_{n-p-1,\ 1-\frac{\alpha}{2n}}.$$

5.3. Colinealidad de los predictores

5.3.1. Diagnóstico de multicolinealidad

Cuando las variables predictoras incluidas en el modelo están correlacionadas entre ellas, decimos que existe intercorrelación o multicolinealidad. Algunos de los problemas típicos que aparecen cuando las variables regresoras están fuertemente correlacionadas son:

1. Los coeficientes de regresión estimados se modifican sustancialmente cuando se agregan o se quitan variables del modelo.

2. Los errores estándares de los estimadores de los coeficientes aumentan espúreamente cuando se incluyen covariables muy correlacionadas en el modelo. Esto se denomina *inflar la varianza estimada de los estimadores.*

3. Los coeficientes pueden ser no significativos aún cuando exista una asociación verdadera entre la variable de respuesta y el conjunto de variables regresoras.

5.3.2. Diagnóstico informal

Las siguientes situaciones son indicativas de multicolinealidad severa:

1. Cambios importantes en los coeficientes estimados al agregar o quitar variables o al modificar levemente las observaciones.

2. Tests no significativos para los coeficientes asociados a variables que teóricamente son importantes predictores, aún cuando observamos que existe una relación estadística entre las predictoras y la respuesta. El modelo puede tener R^2 cercano a 1 y el test F para el modelo ser fuertemente significativo y los tests para los coeficientes pueden ser no significativos. Recordemos que el efecto de la multicolinealidad es inflar la varianza estimada y en consecuencia el estadístico t asociado a cada coeficiente beta será pequeño. Por lo tanto, cuando existe multicolinealidad es difícil evaluar los efectos parciales.

3. Coeficientes estimados con signo contrario al que se espera según consideraciones teóricas.

4. Coeficientes de correlación grandes para las predictoras tomadas de a pares.

Aunque este último diagnóstico parece ser el modo más simple de detectar multicolinealidad, adolece de un problema: al calcular los coeficientes de correlación de Pearson de todas las variables regresoras tomadas de a pares sólo estamos mirando los vínculos lineales entre dos covariables. El problema es que podría haber un vínculo lineal muy estrecho entre una colección de variables y otra variable en particular. Un enfoque más apropiado es hacer una regresión de cada variable regresora sobre las demás variables regresoras. Cuando el R^2 de alguna de estas regresiones sea cercano a 1, deberíamos preocuparnos por el efecto de la multicolinealidad.
Finalmente diremos que la interpretación de los coeficientes se vuelve dudosa cuando existe multicolinealidad. Recordemos que en regresión múltiple (aditiva) cada coeficiente representa el efecto de la variable regresora cuando todas las demás variables se mantienen constantes. Pero si dos variables regresoras, por ejemplo X_1 y X_2, están fuertemente correlacionadas tiene poco sentido pensar en el efecto de X_1 sobre Y cuando X_2 se mantiene constante.

5.3.3. Diagnóstico formal

Un método formal para detectar la presencia de multicolinealidad que está ampliamente difundido es el uso de los factores de inflación de la varianza, más conocidos como **Variance Inflation Factor (VIF).** Es un número que se calcula para cada covariable. El VIF de la $k-$ésima covariable se calcula del siguiente modo

$$VIF_k = \frac{1}{1 - R_k^2}, \qquad 1 \leq k \leq p-1,$$

donde R_k^2 es el coeficiente de determinación múltiple cuando X_k es regresado en las $p-2$ restantes covariables X en el modelo.
El VIF_k es igual a uno si $R_k^2 = 0$, es decir si la $k-$ésima covariable no está correlacionada con las restantes covariables. Cuando $R_k^2 \neq 0$, el VIF_k es mayor a uno. Cuando R_k^2 está muy cerca de uno, el VIF_k se vuelve un número enorme.
Para un conjunto de datos, el mayor VIF observado se usa como medida de diagnóstico. Si el máximo VIF es mayor a 10, eso es señal de multicolinealidad. Otro criterio es que cuando el promedio de los VIF es considerablemente mayor a uno se está frente a problemas de multicolinealidad.

5.3.4. ¿Cómo tratar el problema de multicolinealidad?

El recurso más simple es elegir un subconjunto de las variables regresoras poco correlacionadas. Si detectamos dos variables muy correlacionadas ¿cómo decidir cuál omitir? En general, conviene omitir aquella que tenga:

- mayor número de datos faltantes,
- mayor error de medición o
- que sea menos satisfactoria en algún sentido.

Otra posibilidad es eliminar variables a través de procedimientos de selección automáticos (se presentarán más adelante).
Cuando varios predictores están altamente correlacionados y son indicadores de una característica común, uno puede construir un índice combinando estas covariables. Los índices de bienestar, como el IDH (índice de desarrollo humano), o el índice de inflación, construidos como promedios ponderados de variables que miden el bienestar en una cierta región o bien los precios asociados a una determinada canasta, son ejemplos clásicos de esta construcción. En aplicaciones en las que se miden varias covariables muy correlacionadas esta puede resultar una buena solución.
En modelos polinómicos o que contienen interacciones, una solución al problema de multicolinealidad es trabajar con los datos centrados para la o las variables predictoras que aparecen en más de un término del modelo. Esto es, no usar la variable X tal como fue medida, sino la diferencia entre el valor observado y el valor medio de X en la muestra.
Existen otros procedimientos para tratar multicolinealidad, tanto clásicos como modernos, que complementan el ajuste de regresión lineal múltiple que hemos descrito, potenciando su alcance. Podemos agruparlos en tres clases importantes de métodos.

- Selección de modelos (o selección de variables). Este enfoque implica la identificación de un subconjunto de predictores que creemos están relacionados con la respuesta. Una vez seleccionado el subconjunto de variables relevantes, ajustamos con mínimos cuadrados un modelo que explica la respuesta con el conjunto reducido de variables. Describimos con detalle este procedimiento en la Sección 5.4.

- Regularización o penalización. Este enfoque involucra ajustar un modelo que contiene a todos los predictores. Sin embargo, el método de ajuste fuerza a encojer a los estimadores, achicándolos en valor absoluto, en relación con los estimadores que se obtienen usando mínimos cuadrados. Este encojimiento (también conocido en la literatura matemática como regularización) tiene el efecto de reducir la varianza estimada. Dependiendo de qué regularización se realice, la estimación de algunos coeficientes terminará siendo exactamente cero. Por eso, los métodos de regularización o penalización también pueden llevar a cabo la selección de variables. Los estimadores de regularización

más utilizados son los estimadores ridge o los lasso, y una combinación de ambos que se denominan elastic net. El libro de James, Witten, Hastie, y Tibshirani [2013] constituye una fuente muy accesible y actualizada, que además comenta los comandos de `R` para apropiados para implementarlos. Un enfoque más técnico por los mismos autores es Friedman, Hastie, y Tibshirani [2008]. No nos ocuparemos de estos temas en el curso.

- Reducción de la dimensión. Este enfoque propone proyectar las $p-1$ variables predictoras en un espacio de dimensión M, con $M < p$. Esto se realiza calculando M combinaciones lineales, o proyecciones, diferentes de los predictores. Luego, estas M proyecciones se utilizan como nuevas covariables para ajustar un modelo de regresión lineal por mínimos cuadrados. Puede verse el libro de James et al. [2013] para una introducción al tema, que tampoco trataremos en estas notas. Los métodos comprendidos en esta categoría son *Principal Components Regression*, *Partial Least Squares*, *Fitted Principal Components*, *Projection Pursuit Regression*, entre otros.

5.4. Selección de modelos

Ya hemos observado que cuando tenemos K covariables disponibles y una variable a explicar Y, pueden, en principio, ajustarse 2^K modelos distintos. Decimos en principio, pues este total de modelos no incluye aquellos que tienen interacciones. En esta sección estamos pensando que si uno quiere evaluar ciertas interacciones, las debe incluir en esas K covariables iniciales. Lo mismo si uno quisiera evaluar algunas potencias de las covariables originales, o algunas transformaciones más complicadas de ellas. De este modo, cuando K es un número grande, la cantidad de modelos posibles crece exponencialmente, y evaluarlos uno por uno puede ser inmanejable. Por ejemplo, para $K = 8$, hay $2^8 = 256$ modelos posibles: hay un modelo sin covariables, 8 modelos de regresión lineal simple, cada uno con una sola covariable, $\binom{8}{2} = 28$ modelos con dos covariables $\{X_1, X_2\}$, $\{X_1, X_3\}$, $\{X_1, X_4\}$, $\{X_2, X_3\}$, etc. , $\binom{8}{3} = 56$ modelos con tres covariables, etcétera.
Lo que se denomina selección de modelos corresponde a la tarea de elegir el mejor modelo para nuestros datos.

5.4.1. Criterios para comparar modelos

Una vez que se tienen todas las variables, es de interés contar con un criterio numérico para resumir la bondad del ajuste que un modelo lineal con un cierto conjunto de covariables da a la variable dependiente observada. A partir de este criterio se podrán ranquear los modelos y elegir un conjunto de unos pocos buenos candidatos para estudiar luego en detalle.

A continuación presentamos algunos de los criterios más frecuentemente utilizados en regresión lineal para la selección de modelos. No son los únicos, pero sí los más difundidos. Cuando ajustamos un modelo con $p-1$ covariables, es decir, con p coeficientes $\beta's$ podemos tomar como criterio para evaluar el ajuste a:

- R_p^2 o $SS\text{Res}_p$: Un primer criterio para comparar modelos es mirar el R^2 obtenido con cada uno de ellos y elegir aquél con mayor R^2. Usamos el subíndice p para indicar la cantidad de parámetros $\beta's$ hay en el modelo (es decir, $p-1$ covariables). Como tenemos que

$$R_p^2 = 1 - \frac{SS\text{Res}_p}{SSTotal},$$

resulta que comparar modelos usando el criterio de elegir aquél cuyo R_p^2 sea lo más grande posible equivale a elegir aquel que tenga la menor suma de cuadrados de residuos $SS\text{Res}_p$ (ya que la suma de cuadrados total $SSTotal = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$ no depende de las covariables del modelo ajustado y por eso permanece constante). Pero como ya observamos, el R^2 aumenta al aumentar $p-1$, el número de covariables, sean estas apropiadas para ajustar los datos o no. Es por eso que el criterio no es identificar el modelo con mayor R^2 (ese será siempre el modelo con todas las covariables disponibles) sino encontrar el punto a partir del cual no tiene sentido agregar más variables ya que estas no inciden en un aumento importante del R^2. Muchas veces esto sucede cuando se han incorporado unas pocas variables al modelo de regresión. Por supuesto, encontrar el punto donde este aumento se empieza a estancar es un asunto de criterio individual. Suele ser bastante informativo graficar el mejor R_p^2 en función de p y evaluar gráficamente cuándo el crecimiento en el R^2 es tan poco que no justifica la inclusión de la covariable adicional.

- $R_{a,p}^2$ o MSE_p : Como el R_p^2 no toma en cuenta el número de parámetros en el modelo de regresión, un criterio de decisión mucho más objetivo y automatizable es calcular y comparar modelos por medio del R_a^2. Lo subindicaremos como $R_{a,p}^2$ para indicar la cantidad de coeficientes $\beta's$ presentes en el modelo. Recordemos que

$$R_{a,p}^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SS\text{Res}_p}{SSTotal} = 1 - \frac{MS\text{Res}_p}{\frac{SSTotal}{n-1}}.$$

Como $\frac{SSTotal}{n-1}$ está fijo en un conjunto de datos dado (sólo depende de las Y observadas), el $R_{a,p}^2$ aumenta si y sólo si el $MS\text{Res}_p$ disminuye. Luego, el coeficiente de determinación múltiple ajustado $R_{a,p}^2$ y el cuadrado medio del error $MS\text{Res}_p$, proveen información equivalente acerca del ajuste obtenido.

Al usar este criterio buscamos el subconjunto de $p-1$ covariables que maximicen el $R^2_{a,p}$, o un subconjunto de muchas menos covariables para las cuales $R^2_{a,p}$ esté muy cerca del máx $R^2_{a,p}$, en el sentido que el aumento en el R^2_a sea tan pequeño que no justifique la inclusión de la o las covariables extra.

- *C_p de Mallows*: Para utilizar esta medida hay que asumir que en el modelo con el total de las K covariables (el más grande posible) están todas las covariables importantes de modo que en ese modelo completo, la estimación de la varianza del error, σ^2, es insesgada. El valor del C_p se define por

$$C_p = \frac{SS\text{Res}_p}{MS\text{Res}\,(X_1, \ldots, X_K)} - (n - 2p)$$

donde $SS\text{Res}_p$ es la suma de los cuadrados de los errores del modelo con p parámetros (es decir, con $p-1$ covariables) y $MS\text{Res}\,(X_1, \ldots, X_K)$ es el estimador de la varianza del error σ^2, calculado bajo el modelo con todas las posibles covariables $X_1, \ldots, X_K$. Cuando se usa el C_p como criterio, se busca aquel subconjunto de p covariables X que tengan un C_p pequeño, lo más cercano a p posible. Es fácil ver que para el modelo completo, $C_K = K$.

- *AIC_p*, o el *Criterio de Akaike* y *SBC_p* o el *Criterio Bayesiano de Schwartz*, son otros dos criterios que, al igual que el C_p de Mallows, penalizan a los modelos con muchas covariables. Se buscan los modelos que tienen valores pequeños de AIC_p o SBC_p, donde estas cantidades están dadas por

$$\begin{aligned} AIC_p &= n \ln (SS\text{Res}_p) - n \ln (n) + 2p \\ SBC_p &= n \ln (SS\text{Res}_p) - n \ln (n) + p \ln (n) \end{aligned}$$

Observemos que para ambas medidas, el primer sumando decrece al aumentar p. El segundo sumando está fijo (puesto que n lo está, para un conjunto de datos) y el tercer sumando crece al crecer p, es decir, el número de covariables. Ambas medidas representan una buena ponderación entre ajuste apropiado (es decir, $SS\text{Res}_p$ pequeña) y parsimonia del modelo (es decir, pocos parámetros a ajustar, o sea, p pequeño). El Criterio SBC_p también se llama *Criterio Bayesiano de Información* (*BIC*, por sus siglas en inglés).

5.4.2. ¿Cuál de estos criterios utilizar?

Todos estos criterios miden cualidades deseables en un modelo de regresión. Ocasionalmente, una única ecuación de regresión produce valores óptimos de los cuatro criterios simultáneamente, con lo que uno puede confiar que éste es el mejor modelo en términos de estos criterios.

Desafortunadamente esto raramente ocurre y diferentes instrumentos identifican diferentes modelos. Sin embargo, tomados en conjunto estos criterios permiten identificar un conjunto pequeño de modelos de regresión que pueden ser construidos a partir de las variables independientes relevadas. Conviene entonces estudiar estos pocos modelos más detalladamente, teniendo en cuenta los objetivos del estudio, nuestro conocimiento del área del que provienen los datos y la evaluación de los supuestos del análisis de regresión para realizar una selección criteriosa de cual es el "mejor" modelo.

5.4.3. Selección automática de modelos

Al inicio de la Sección 5.4 hemos visto que en el proceso de selección de modelos es necesario comparar un número muy grande de modelos entre sí. Para simplificar esta tarea, existen una variedad de procedimientos automáticos de selección de modelos, programados en los paquetes estadísticos.
Un gran problema de estas búsquedas automáticas es que en general, están programadas para trabajar con la base completa de $n \times K$ observaciones. Si hubiera una observación faltante (*missing data*) (es decir, un caso para el cual no se registró **una** de las variables) estos algoritmos remueven el caso **completo** y hacen la selección de modelos basados en $n - 1$ observaciones. Esto puede volverse un problema si n es pequeño y hay varias variables con observaciones faltantes.
Los métodos más populares de selección de variables son:

1. Todos los subconjuntos posibles (*Best subset*).
2. Eliminación *backward* (hacia atrás).
3. Selección *forward* (incorporando variables).
4. *Stepwise regression* (regresión de a pasos).

A continuación los describimos. Asumimos que $n > K$ (o sea, que tenemos más observaciones que covariables).

5.4.4. Todos los subconjuntos posibles (*Best subset*)

Estos algoritmos ajustan **todos** los submodelos posibles (los 2^K) y luego los ranquean de acuerdo a algún criterio de bondad de ajuste. Por supuesto, esto involucra hacer 2^K regresiones. Siempre que sea posible es aconsejable usar este procedimiento ya que es el único método que garantiza que se obtendrá el modelo final que realmente optimice la búsqueda con el criterio elegido: por ejemplo mayor R^2_a, o mejor C_p, etc. Es decir, garantiza que el modelo final es el "mejor" para el presente conjunto de datos y para los criterios utilizados.

Una vez que todos los modelos han sido ajustados, en general el paquete exhibe los 10 (o una cantidad prefijable) mejores modelos de acuerdo al criterio elegido, entre todos los que tienen el mismo número de variables.
Cuando la cantidad original de potenciales covariables es muy grande, K mayor a 40, por ejemplo, no es posible ajustar todos los modelos posibles ya que $2^{40} = 1\,099\,511\,627\,776$. Se vuelve necesario usar otro tipo de procedimientos, computacionalmente más realizables, que buscan elegir un modelo luego de una búsqueda que explora una sucesión de modelos de regresión que en cada paso agrega o quita una covariable X. El criterio para agregar o quitar una covariable, en el caso secuencial, puede escribirse equivalentemente en términos de la suma de los cuadrados de los residuos, los estadísticos F parciales, el estadístico t asociado a un coeficiente, o el R_a^2. Son los tres procedimientos que describimos a continuación.

5.4.5. Eliminación *backward* (hacia atrás).

El procedimiento comienza construyendo el modelo con todas las predictoras y en cada paso se elimina una variable. La secuencia del procedimiento es la siguiente. Se define un nivel de significación fijo α.

1. El modelo inicial contiene todos los potenciales predictores (que hemos denominado K).

2. Si todas las variables producen una contribución parcial significativa (es decir, un estadístico t con p−valor $< \alpha$) entonces el modelo completo es el modelo final.

3. De otro modo, se elimina la variable que tenga la menor contribución parcial (es decir, el mayor p−valor de su estadístico t) cuando todas las demás están en el modelo.

4. Se ajusta el nuevo modelo con $(K-1)$ predictores y se repiten los pasos 2 y 3 hasta que todas las variables en el modelo tengan un coeficiente estimado cuyo p−valor asociado al estadístico t sea menor a α.

Si hay una alta multicolinealidad en el conjunto de los K predictores, este procedimiento no es muy recomendable.

5.4.6. Selección *forward* (incorporando variables)

En este caso, comenzamos con el modelo sin variables y vamos agregando las variables de a una por vez. Ingresa la variable que más contribuye a explicar a Y cuando las otras ya están en el modelo. Se elige un nivel de significación fijo α. La secuencia de pasos es la siguiente:

1. Primero se ajustan todos los modelos de regresión lineal simple con Y como respuesta y una sola covariable explicativa. Se elige la que tiene el mayor valor del estadístico F o, equivalentemente, el menor p−valor del estadístico t asociado al coeficiente, siempre que dicho p−valor sea inferior a α, sino el procedimiento termina y se elige el modelo sin covariables

2. En el segundo paso, se busca elegir entre todos los modelos de dos covariables que tienen a la que fue seleccionada en el primer paso aquél para el cuál el test F parcial dé mas significativo. El test F parcial es el que compara el ajuste del modelo con dos variables con el ajuste del modelo con una variable elegido en el primer paso. Es decir, es el test que mide la significatividad de la segunda variable a ser incorporada en el modelo cuando la primera ya está en él. Para aquel modelo que tenga el F parcial más significativo o, equivalentemente, el test t asociado al coeficiente de la variable a ser incorporada más significativo, o sea, el menor p−valor, se compara a dicho p−valor con el valor crítico α. Si el p−valor es menor que α se elige dicho modelo, si el p−valor supera el valor crítico, el procedimiento se detiene, y el output del proceso es el modelo que tiene una única covariable significativa, que fue seleccionada en el paso 1.

3. Ahora se calculan los estadísticos F parciales de todos los modelos con tres covariables, que tienen a las dos covariables ya elegidas e incorporan una tercera. Se continua de esta manera (como en el paso 2) hasta que ninguna variable produce un F parcial (o t) significativo.

Si se usa un punto de corte muy exigente (digamos $\alpha < 0{,}01$) serán incluídas menos variables y existe la posibilidad de perder covariables importantes. Si se usa un punto de corte menos exigente ($\alpha < 0{,}20$) es menos probable que se pierdan covariables explicativas importantes pero el modelo contendrá más variables.
Una vez que el procedimiento finaliza, no todas las variables en el modelo necesariamente tendrán coeficientes parciales significativos.

5.4.7. Selección stepwise

Es una modificación del procedimiento forward que elimina una variable en el modelo si ésta pierde significación cuando se agregan otras variables. La aproximación es la misma que la selección forward excepto que a cada paso, después de incorporar una variable, el procedimiento elimina del modelo las variables que ya no tienen contribución parcial significativa. Una variable que entró en el modelo en una etapa, puede eventualmente, ser eliminada en un paso posterior.
En este caso será necesario definir un punto de corte para que ingrese una variable α_I y otro para eliminarla del modelo α_E. Uno puede desear ser menos exigente

(mayor p−valor) en el punto de corte para que una variable salga del modelo una vez que ingresó, o usar el mismo valor para ambos.
Este procedimiento, en general produce modelos con menos variables que la selección forward.

5.4.8. Limitaciones y abusos de los procedimientos automáticos de selección de variables

Cualquier método automático de selección de variables debe ser usado con precaución y no debería ser sustituto de un investigador que piensa, ya que no hay garantías que el modelo final elegido sea "óptimo". Conviene tener en cuenta las siguientes observaciones.

- Cuando se proponen términos de interacción entre las variables regresoras, el modelo debe contener las interacciones significativas y los efectos principales de las variables involucradas en estas interacciones, sean éstas significativas o no. De otro modo el modelo carece de interpretación. La mayoría de los procedimientos automáticos no tienen este cuidado. Lo mismo sucede cuando uno incorpora una variable categórica codificada con dummies: o entran todas las dummies, o ninguna, pero no es correcto poner algunas de ellas (las significativas) y otras no, porque sino el modelo carece de interpretación. Con las categóricas, otra posibilidad consiste en recategorizarlas, agrupando algunas categorías, y luego ajustar el modelo nuevamente, esperando que se obtenga un mejor ajuste (no siempre ocurre).

- El hecho de que un modelo sea el mejor en términos de algún criterio (C_p o R^2_a, por ejemplo) no significa que sea el mejor desde el punto de vista práctico. Ni tampoco que para este modelo valgan los supuestos.

- El procedimiento de selección automática puede excluir del modelo variables que realmente deberían estar en el modelo de acuerdo a otros criterios teóricos.Una posibilidad es forzar a que ciertas variables aparezcan en el modelo, independientemente del hecho de que tengan coeficientes significativos. Por ejemplo, podemos hacer una regresión backward sujeta a la restricción de que el modelo incluya ciertos términos especificados de antemano. Esto asegura que el modelo final contiene las variables de interés primario y toda otra variable o interacción que sea útil a los efectos de predicción. Algunos paquetes permiten esta alternativa.

- Una vez que hemos seleccionado un modelo final usando cualquier procedimiento de selección, la inferencia realizada sobre ese modelo es **sólo aproximada**. En particular, los p−valores serán menores y los intervalos de confianza más angostos que lo que deberían ser, puesto que el modelo seleccionado

es aquél que más fuertemente refleja los datos. (Hemos hecho uso y abuso de nuestros datos para obtener un modelo, es de esperar que otra muestra aleatoria de observaciones del mismo tipo a la que se le ajuste este modelo tenga menor capacidad predictiva).

- Existe una diferencia sustancial entre selección de modelos explicativos y exploratorios. En investigación explicativa, uno tiene un modelo teórico y pretende testearlo a través de un análisis de regresión. Uno podría querer testear si una relación que se propone como espúrea desaparece al incorporar una nueva variable en el modelo. En este enfoque, los procedimientos de selección automática en general no son apropiados, ya que es la teoría la que determina cuáles son las variables que deben estar en el modelo.

- En investigación exploratoria el objetivo es encontrar un buen conjunto de predictores. Uno intenta maximizar R^2 independientemente de explicaciones teóricas.

- ¿Por qué podría dejarse en el modelo final una variable que no resulta estadísticamente significativa? Muchas veces pueden aparecer variables en el modelo seleccionado para las cuáles el p−valor del test t no es menor que 0,05. Esto puede deberse a que haya motivos teóricos que indican que la respuesta depende de dicha covariable y que tal vez el tamaño de muestra no haya sido lo suficientemente grande como para comprobarse la significatividad estadística. Se deja para que el modelo no resulte sesgado. Los estimadores de los coeficientes son insesgados si el modelo es correcto (es decir, contiene todas las covariables apropiadas en la forma correcta, dejar covariables con sustento teórico para que estén permite que los estimadores de los efectos de otras covariables sean insesgados). Otro motivo para dejarla puede ser porque su presencia ayuda a reducir la varianza estimada del error, permitiendo que otros coeficientes resulten significativos. Y también pueden dejarse covariables aunque no sean significativas pero que permitan comparar el modelo presentado con otros modelos publicados con antelación.

En resumen, los procedimientos de selección automática de modelos no son sustitutos de una cuidadosa construcción teórica que guíe la formulación de los modelos.

5.4.9. Validación de modelos

El paso final en el proceso de construcción o selección de modelos lo constituye el proceso de validación de los modelos. Esta etapa de validación involucra, usualmente, chequear el modelo candidato con datos independientes a los utilizados para proponer el modelo. Hay cuatro formas básicas de validar un modelo de regresión:

1. Recolectar un nuevo conjunto de datos que permita chequear el modelo y su habilidad predictiva.

2. Comparar los resultados con las expectativas teóricas, resultados empíricos previos y resultados de simulaciones.

3. Cuando fuera posible, usar otras técnicas experimentales para confirmar el modelo. Esto, por supuesto, dependerá de las herramientas propias de cada disciplina.

4. Cuando el tamaño de muestra lo permitiera, otra posibilidad es dividir al conjunto de observaciones disponible en dos grupos disjuntos. Con uno de ellos se selecciona el modelo más apropiado. Este grupo se denomina *muestra de entrenamiento* (*training sample*). Con el segundo grupo, que se llama *muestra de validación* (*validation set*) se evalúa la razonabilidad y la capacidad predictiva del modelo seleccionado. A este proceso de validación se lo denomina a veces, *cross-validation*, es decir, validación cruzada.

A. Talleres

A.1. Taller 1: Coeficiente de Correlación y Regresión Lineal Simple

T1.Ej1. En una ciudad con graves problemas de obesidad en la población, se solicitó a un grupo de 100 adolescentes que registrara durante un mes la cantidad de horas que dedicaban cada día a actividades sedentarias (mirar televisión, estudiar o utilizar la computadora) y las promediaran. El archivo "*Adol horas.xls*" presenta la edad en años (*edad*), el género (Varón, Mujer), el promedio de horas por día dedicadas a actividades sedentarias (*horas*) como así también un número (*ID*) para identificar a cada participante. Importe los datos que se encuentran en el archivo "*Adol horas.xls*". Estos datos fueron artificialmente generados.

a) Obtenga el diagrama de dispersión de *horas* (en el eje vertical) en función de la *edad* (en el eje horizontal), ya sea global y coloreando las observaciones por género, y un diagrama de dispersión para cada género por separado.

b) Calcule el coeficiente de correlación entre la variable *edad* y la variable *horas* para todos los datos juntos y para cada género.

c) Describa el tipo de asociación que muestran los diagramas de dispersión de las variables *edad* y *horas* del ítem a). Compare con los correspondientes coeficientes de correlación.

d) Testee la asociación lineal de las variables en cada uno de los casos considerados en el ítem b). ¿Cuáles p-valores obtiene? ¿Qué se está testeando?

T1.Ej2. Abra y examine las variables del archivo "ingresos.txt". éste corresponde a una base de datos de 40 individuos, para los que se registraron las variables: *Id* (identificador, un número entre 1 y 40 que identifica al número de observación), *nivelEduc* (nivel educativo), *edad* y *salario*. La variable *nivelEduc* está codificada de 1 a 10, donde 1 corresponde al menor nivel educativo alcanzado y 10 al mayor. La variable *salario* corresponde al salario bruto mensual (es decir, antes de impuestos), en dólares. La variable *edad* está medida en años. Suponga, siempre que lo necesite, que los datos tienen distribución normal.

a) Obtenga el coeficiente de correlación y el p-valor correspondiente al test asociado entre *nivelEduc* y *salario*. Interprete el resultado. Realice el diagrama de dispersión (con *nivelEduc* en el eje horizontal). Describa

el tipo de asociación que muestran las variables. ¿Le parece que es un resultado lógico? Justifique brevemente.

b) Obtenga el coeficiente de correlación y el p-valor correspondiente al test asociado entre *nivelEduc* y *salario* para cada edad, interpretando el resultado. Compare con el resultado obtenido en a).

c) Realice el diagrama de dispersión entre *nivelEduc* y *salario* para cada *edad*. ¿Qué observa? ¿Puede explicar ahora las diferencias encontradas entre a) y b)?

d) Haremos un cambio de unidades en las que está expresada la variable *salario* para facilitar la interpretación. Para ello defina una nueva variable: *sal100* (salario en cientos) que es igual a la variable *salario* dividida por 100. Ajuste una recta de cuadrados mínimos para la variable respuesta *sal100* y la variable explicativa *nivelEduc* sin tener en cuenta la variable *edad*. Describa e interprete cada uno de los resultados. ¿Qué significa el coeficiente de la variable explicativa *nivelEduc*?

e) Para cada edad, ajuste una recta de cuadrados mínimos con *sal100* como variable respuesta y *nivelEduc* como variable explicativa. ¿Qué significa el coeficiente de la variable explicativa en cada una de las regresiones ajustadas?

A.2. Ejercicio domiciliario

El valor energético (en kcal. por cada 100g.) de galletitas de agua de marca A (Y) se relaciona con la cantidad de grasas totales (en g.) (X) involucradas en su producción. Un experimentador toma una muestra de tamaño 22 (es decir, compra 22 paquetes de galletitas y elige una de cada uno) para verificar la adecuación de un modelo de regresión lineal a esta relación. Utilizando el archivo de datos *galletitas.xls* responda a las siguientes preguntas: (no hace falta que copie en su respuesta las salidas del paquete ni los comandos con los que las hizo, simplemente responda brevemente a las preguntas, en general bastará con una o dos oraciones).

1. Exprese el modelo de regresión lineal indicando claramente los parámetros y variables involucradas. Escriba los supuestos necesarios para que sean válidas las conclusiones respecto de los tests y los intervalos de confianza.
2. Ajuste el modelo. Dé la ecuación de la recta estimada.
3. ¿Es la pendiente del modelo significativa? Es decir, ¿hay un vínculo lineal entre las valor energético (en kcal. por cada 100g.) de galletitas de agua de marca A (Y) y la cantidad de grasas totales (en g.) (X) involucradas en su producción? Conteste a nivel 0.05. Al escribir su respuesta, escriba claramente las hipótesis que testea, el pvalor obtenido y su conclusión.
4. ¿Es la ordenada al origen significativa al nivel 0.05?
5. Estime la varianza del error (σ^2).
6. Interprete los parámetros estimados (en su respuesta a esta pregunta debería aparecer una frase que comience más o menos así: "Por cada aumento de 1g. en la cantidad de grasas totales....")
7. Al investigador le interesa calcular la cantidad de calorías esperadas para 100g de galletitas de agua de marca A producidas con $X = 30g.$ de grasas totales. Diga cuál es el valor esperado, en base a los datos dados.
8. Dé un intervalo de confianza de 0.95 del valor calculado en el ítem anterior.
9. Halle un intervalo de confianza para la pendiente correspondiente al ajuste de la marca A de nivel 0.95.
10. ¿Cuánto vale el coeficiente de determinación R^2? ¿Cómo interpreta este valor para estos datos?
11. El fabricante de las galletitas de marca A le regala al investigador un paquete de galletitas producidas con 40g. de grasas totales, que no usa para hacer su análisis. Antes de comer una, el investigador se pregunta cuántas calorías

estará ingiriendo. Responda a esta pregunta calculando dicho valor. Además, dé un intervalo de predicción del 99 % para dicho valor.

12. Ídem la pregunta anterior pero para un paquete de galletitas producidas con 90g. de grasas totales. Con el ajuste obtenido, ¿se puede realizar este cálculo?

13. ¿Para qué valores de grasas totales involucradas en la producción de galletitas puede contestar la pregunta 11 con los datos dados? (es decir, diga para qué valores de X no está extrapolando al calcular valores predichos).

14. ¿Para cuál de los valores posibles para X la pregunta anterior el intervalo a calcular resultará más corto? ¿Para cuál (o cuáles) más largo?

15. Responda a la pregunta 3 con otro test.

16. Diga verdadero o falso:

    a) Los residuos son observables.
    b) Los errores son observables.
    c) Los residuos son iguales que los errores.
    d) Los residuos son aleatorios.
    e) Los errores son aleatorios.

17. Analice la adecuación del modelo. Indique en que salidas/gráficos se basan sus conclusiones.

A.3. Taller 2: Regresión Lineal: medidas de diagnóstico y transformaciones

T2.Ej1. Visualice los datos que brinda el archivo "*lowbwi.txt*" correspondientes a 100 bebés de bajo peso al nacer, que fueron extraídos del libro de Pagano y Gauvreau[5]. Usaremos sólo las variables *headcirc*, que es el perímetro cefálico en centímetros, y *birthwt*, que es el peso al nacer en gramos. Para una mejor interpretación, consideraremos la variable *pesokg*, el peso en kilogramos, que será igual a *birthwt* dividida por 1000.

a) Escriba, en papel, el modelo lineal considerando a *pesokg* como variable explicativa y a *headcirc* como variable dependiente, e indique qué significan β_0 y β_1.

b) Ajuste el modelo lineal propuesto en el ítem a) y repita a) para el modelo estimado.

[5]Pagano, M., Gauvreau, K. (2000) *Principles of Biostatistics*, Second Edition, Duxbury Thomson Learning.

c) ¿Cuánto estima que aumentará el perímetro cefálico esperado (o medio) de un bebé si aumenta en 10 g.? ¿Y si aumenta 100 g.? ¿Tiene sentido responder esta pregunta si se trata de 1 kg. de aumento?

d) Prediga el perímetro cefálico medio para la población de bebés que pesaron 820 g. al nacer. Lo mismo para los bebés de 1200 g. Calcule los intervalos de confianza y los intervalos de predicción de nivel 0,95 en cada caso.

e) Obtenga el diagrama de dispersión de los datos junto con la recta de regresión ajustada. Superponga luego las bandas de los intervalos de confianza y los intervalos de predicción de nivel 0,95.

f) Hacer un gráfico de residuos (los que quiera, no difieren mucho entre sí en este caso) versus la covariable. Y también un gráfico de residuos versus predichos. ¿Son muy diferentes? ¿Tenemos evidencia de que no se cumplan los supuestos del modelo lineal? ¿Podemos identificar en este gráfico alguna observación con residuo alto?

g) Efectúe un ajuste robusto de los datos. Compare con lo obtenido ajustando mínimos cuadrados.

h) ¿Hay alguna observación influyente en este conjunto de datos? Verifique esto utilizando los estadísticos pertinentes.

i) Ahora contaminemos los datos. Agreguemos dos datos a la base: 101 y 102 que figuran a continuación. Rehaga los ítems b), e) y h) con el

Dato	*pesokg*	*headcirc*
101	0,40	50
102	0,35	20

nuevo conjunto de datos.

j) Complete la siguiente tabla:

Datos	$\widehat{\beta}_0$ (p-valor)	$\widehat{\beta}_1$ (p-valor)	R^2
1 a 100			
1 a 102			
1 a 101			
1 a 100 y 102			

T2.Ej2. Abra el archivo "*gross national product.txt*", que corresponde a los datos tratados en el libro, ya mencionado, de Pagano y Gauvreau, capítulo 18. Las

variables son tasa de natalidad por 1000 habitantes (*birthrt*) y producto nacional bruto expresado en dólares estadounidenses (*gnp*). Las observaciones conciernen a 143 países distintos.

a) Obtenga el diagrama de dispersión de los datos tomando a *gnp* como variable explicativa. ¿Le parece razonable ajustar un modelo lineal a estos datos?

b) Ajuste un modelo lineal de todos modos. Mire la salida y el gráfico de residuos. ¿Qué ve en este gráfico?

c) Transforme la variable explicativa en $lgnp = \ln(gnp)$ y repita los ítems a) y b) para el modelo que explica la tasa de natalidad con esta nueva covariable. Escriba el modelo propuesto y el modelo ajustado. Interprete: ¿cómo impacta en la variable respuesta un 1 % de incremento en la covariable *gnp*?

d) Prediga la tasa de natalidad (por cada 1000 habitantes) para un país con producto nacional bruto de 1816 dólares estadounidenses y brinde los intervalos de confianza y de predicción para dicha tasa.

A.4. Taller 3: Regresión Lineal Múltiple

T3.Ej1. Para los datos de niños de bajo peso, se encontró una relación lineal significativa entre la presión sistólica y la edad gestacional. Los datos están en el archivo "lowbwt.txt" y fueron generados artificialmente. Las mediciones de presión sistólica están guardadas bajo el nombre *presSist*, y las correspondientes edades gestacionales bajo *edadG*. También en ese archivo figuran los datos de *apgar5*, el score Apgar a los 5 minutos para cada niño recién nacido. El score Apgar es un indicador del estado general de salud del niño a los 5 minutos de haber nacido; aunque en realidad es una medida ordinal que se la suele tomar como si fuera continua.

a) Realice un diagrama de dispersión de la presión sistólica versus el score Apgar. ¿Parece haber una relación lineal entre estas dos variables?

b) Usando la presión sistólica como variable respuesta y la edad gestacional y el score Apgar como variables explicativas, ajuste el modelo lineal $\mathbb{E}(Y|\mathbf{X}) = \beta_0 + \beta_1 \cdot edadG + \beta_2 \cdot apgar5$ donde $\mathbf{X} = (edadG,\ apgar5)$. Interprete los coeficientes estimados.

c) ¿Cuál es la presión media estimada para la población de niños de bajo peso cuya edad gestacional es 31 semanas y cuyo score Apgar es 7?

d) Construya un intervalo de confianza de nivel 0,95 para la verdadera media para el caso anterior.

e) Testee la hipótesis de $H_0 : \beta_2 = 0$ a nivel 0,05.

f) Comente la magnitud de R^2. La inclusión del score Apgar en el modelo que ya contiene a la edad gestacional, ¿mejora su habilidad para predecir a la presión sistólica? Observe que el modelo lineal múltiple considerado en b) mejora la habilidad para predecir a la presión sistólica respecto del modelo lineal simple que tiene a *apgar5* de covariable.

g) Construya un gráfico de los residuos versus los valores ajustados. ¿Qué le dice este gráfico del ajuste obtenido?

T3.Ej2. La idea de este ejercicio es discutir qué significan distintos modelos de regresión múltiple. Probaremos distintos modelos en un solo conjunto de datos. Retomamos el Ejercicio T1.Ej2 del Taller 1. Eran datos guardados en el archivo "ingresos.txt". Consistían en 40 datos de salarios (ingresos), niveles de educación y edad. Para modelar esos datos, propusimos ajustar dos modelos, que recordamos ahora:

- Un modelo lineal simple con *salario* como variable respuesta y *nivelEduc* como variable explicativa: vimos que había una asociación negativa entre ellas, lo cual era ilógico.
- 4 modelos lineales simples basados en 10 datos cada uno, con *salario* como variable respuesta y *nivelEduc* como variable explicativa, pero separados por tramos de edad (*edad* = 20, 30, 40 y 50, respectivamente).

Llamemos Modelo A al ajustado en el Taller 1, o sea a aquel tal que

$$\mathbb{E}(salario|nivelEduc) = \beta_0 + \beta_1 \cdot nivelEduc\,.$$

Consideremos ahora el modelo lineal múltiple, que llamaremos Modelo B, cumpliendo

$$\mathbb{E}(salario|nivelEduc, edad) = \beta_0 + \beta_1 \cdot nivelEduc + \beta_2 \cdot edad\,.$$

a) ¿Cuáles son los supuestos necesarios en el Modelo B para que sean válidas las conclusiones respecto de los tests y los intervalos de confianza? Interprete los parámetros del modelo.

b) Ajuste el Modelo B. Brinde los parámetros estimados. Mejor aún, escriba el modelo ajustado.

c) Evalúe la bondad del ajuste con el test F. Indique si los coeficientes son significativos. Evalúe la adecuación del modelo con el R^2. ¿Qué porcentaje de variabilidad del salario queda explicada por el modelo que tiene a *nivelEduc* y a *edad* como explicativas?

T3.Ej3. Usando los datos del Ejercicio T3.Ej1.

a) Considere el modelo lineal que tiene a *presSist* como variable respuesta y que sólo contiene *edadG* como covariable. Agregue la variable explicativa *varon* al modelo (vale 1 si el bebé es varón y 0 si es mujer). Ajuste el modelo. Comente la significatividad de los parámetros. Dados dos niños con igual edad gestacional, uno varón y otro mujer, ¿cuál tendrá presión sistólica más alta? ¿Por qué?

b) Haga un diagrama de dispersión de presión sistólica versus edad gestacional separando nenes de nenas. Superponga las rectas ajustadas. ¿Es la presión sistólica media de los nenes con una edad gestacional fija significativamente distinta de la presión sistólica media de las nenas con la misma edad gestacional?

c) Agregue la interacción varón – edad gestacional. Ajuste el nuevo modelo.

d) Al modelo que tiene edad gestacional, ¿incluiría a varón como variable explicativa? ¿Incluiría a la interacción como variable explicativa del modelo? ¿Por qué?

T3.Ej4. Usando los datos del Ejercicio T3.Ej2.

a) Considere, ahora, un modelo lineal con interacción, que denominaremos Modelo C

$$\mathbb{E}(salario|\mathbf{X}) = \beta_0 + \beta_1 \cdot nivelEduc + \beta_2 \cdot edad + \beta_{1:2} \cdot nivelEduc \cdot edad$$

donde $\mathbf{X}$=(*nivelEduc*, *edad*, *nivelEduc*$\cdot$*edad*). Interprete los parámetros del modelo.

b) Ajuste el Modelo C y brinde los parámetros estimados.

c) Evalúe la bondad del ajuste con el test F. Indique si los coeficientes son significativos. Evalúe la adecuación del modelo con el R^2. ¿Qué porcentaje de variabilidad del *salario* queda explicada por el modelo que tiene *nivelEduc*, *edad* y la interacción entre ambas como explicativas? ¿Con cuál de los dos modelos (B o C) se quedaría?

d) Considere el modelo, que llamaremos Modelo D, que incluye a *edad* como variable categórica. ¿Cuántas dummies o variables binarias hay que considerar en el nuevo ajuste? Escriba el nuevo modelo, ajústelo y brinde los parámetros estimados.

e) Evalúe la bondad del ajuste con el test F. Testee si es significativa la inclusión de las variables dummies de edad (o sea la variable edad como

cualitativa) cuando en el modelo aparece la variable *nivelEduc*[6]. Indique si los coeficientes son significativos. Evalúe la adecuación del modelo con el R^2. ¿Qué porcentaje de variabilidad del salario queda explicada por el modelo que tiene a *nivelEduc* y a *edad* como explicativas? ¿Con cuál de los modelos se quedaría?

f) Por último, considere el Modelo E que tiene de covariables a *nivelEduc*, las 3 variables binarias que representan las distintas edades y las interacciones entre ellas. ¿Cuál es la diferencia entre este modelo y los 4 modelos lineales simples considerados en el Taller 1? Ajuste el modelo, brinde los parámetros estimados y realice el ítem e) adaptado a este modelo.

[6]Recuerde que esto se responde con otro test F.

Referencias

Draper, N. R., y Smith, H. (1998). Wiley series in probability and statistics. *Applied Regression Analysis, Third Edition*, 707–713.

Field, A. (2005). *Discovering statistics with spss*. SAGE publications Ltd London, UK.

Friedman, J., Hastie, T., y Tibshirani, R. (2008). *The elements of statistical learning: Data mining, inference and prediction*.

Heinz, G., Peterson, L. J., Johnson, R. W., y Kerk, C. J. (2003). Exploring relationships in body dimensions. *Journal of Statistics Education*, *11*(2).

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Kendall, M., y Gibbons, J. (1990). Rank correlation methods, trans. *JD Gibbons, 5th edn edition (London, Edward Arnold)*.

Kutner, M. H., Nachtsheim, C., Neter, J., y Li, W. (2005). *Applied linear statistical models*. McGraw-Hill Irwin.

Leviton, A., Fenton, T., Kuban, K. C., y Pagano, M. (1991). Labor and deliver characteristics and the risk of germinal matrix hemorrhage in low birth weight infants. *Journal of child neurology*, *6*(1), 35–40.

Maronna, R., Martin, R. D., y Yohai, V. (2006). *Robust statistics*. John Wiley & Sons, Chichester. ISBN.

McCullagh, P., y Nelder, J. (1989). *Generalized linear models* (2nd. edn. ed.). Chapman-Hall, London.

Pagano, M., Gauvreau, K., y Pagano, M. (2000). *Principles of biostatistics* (Vol. 2). Duxbury Pacific Grove, CA.

Pinheiro, J. C., y Bates, D. M. (2000). *Mixed-effects models in s and s-plus*. Springer-Verlag New York, Inc.

R Core Team. (2015). R: A language and environment for statistical computing [Manual de software informático]. Vienna, Austria. Descargado de `https://www.R-project.org/`

Rosner, B. (2006). *Fundamentals of biostatistics* (6th. ed. ed.). Thomson Brooks Cole.

Seber, G. A., y Lee, A. J. (1977). *Linear regression analysis.* Wiley, New York.

Wasserman, L. (2010). *All of statistics: A concise course in statistical inference.* Springer Publishing Company, Incorporated.

Weisberg, S. (2005). *Applied linear regression* (3rd. ed. ed.). John Wiley & Sons.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 642–656.