

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 tipar 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 \bar{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,

$$\begin{aligned} Y_i &= \text{perímetro cefálico del } i\text{-ésimo bebé de bajo peso} \ (\text{headcirc}) \\ X_i &= \text{edad gestacional o duración de la gestación del } i\text{-ésimo bebé} \\ &\quad \text{de bajo peso} \ (\text{gestage}) \end{aligned}$$

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

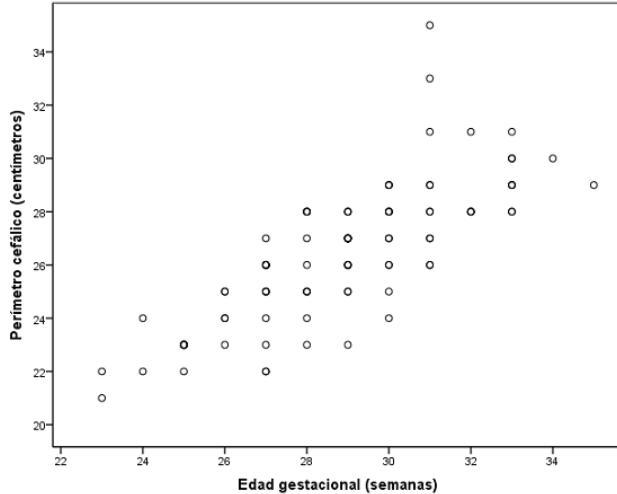
Caso	headcirc	length	gestage	birthwt	momage	toxemia
1	27	41	29	1360	37	0
2	29	40	31	1490	34	0
3	30	38	33	1490	32	0
4	28	38	31	1180	37	0
5	29	38	30	1200	29	1
6	23	32	25	680	19	0
7	22	33	27	620	20	1
8	26	38	29	1060	25	0
9	27	30	28	1320	27	0
10	25	34	29	830	32	1
11	23	32	26	880	26	0
12	26	39	30	1130	29	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 $X = \text{variable predictora o covariante}$ e $Y = \text{variable dependiente o de respuesta}$. El modelo lineal (simple pues sólo vincula una variable predictora con Y) propone que

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad (2)$$

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 β '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

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad (3)$$

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

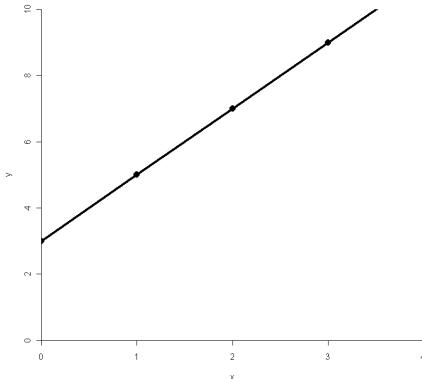
$$y = 2x + 3$$

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 dispersión 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, abundante 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

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad (4)$$

son los siguientes:

1. los ε_i tienen 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

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

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 $E(Y | X)$. 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

$$E(Y | X) = \beta_0 + \beta_1 X \quad (6)$$

o, escrito de otro modo

$$E(Y | X = x_i) = \beta_0 + \beta_1 x_i \quad (7)$$

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.

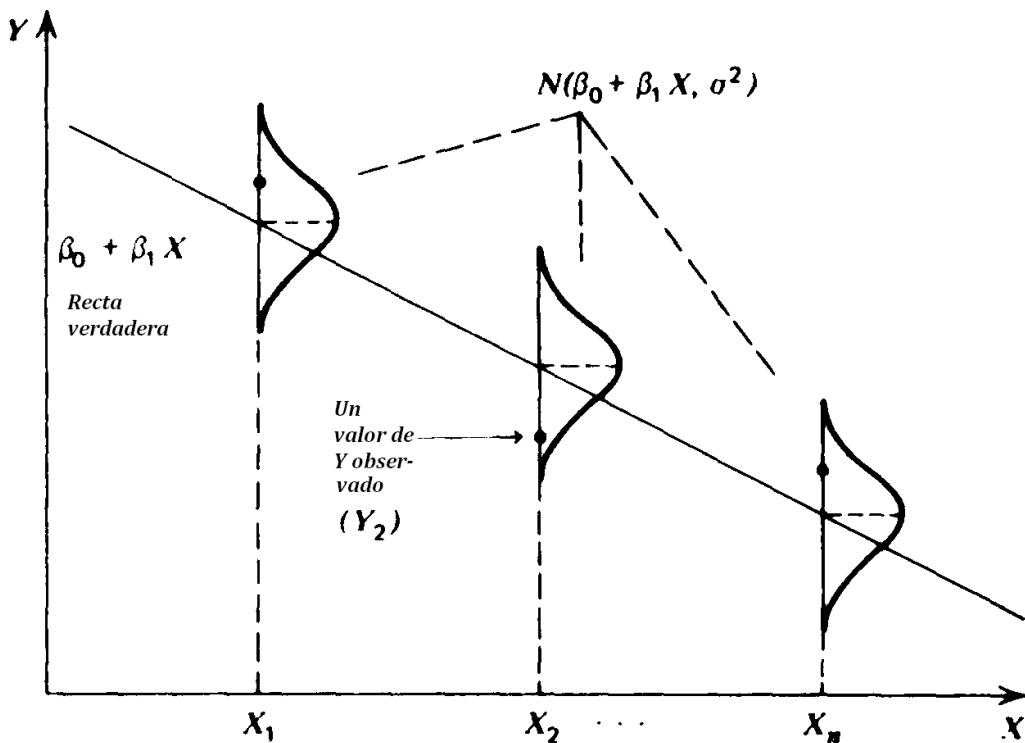
2. La varianza de la variable respuesta Y dado que la predictora está fijada en $X = x$ la denotaremos por $Var(Y | X = x)$. Asumimos que satisface

$$Var(Y | X = x_i) = \sigma^2,$$

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 $Y | X = x_i$) tienen distribución normal, es decir, $Y | X = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.
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.

¹En realidad, se pueden hacer supuestos más débiles aún: asumir que $E(\varepsilon_i | X_i) = 0$, y $Var(\varepsilon_i | X_i) = \sigma^2$. Para los test se asume que $\varepsilon_i | X_i \sim N(0, \sigma^2)$, $1 \leq i \leq n$, 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 pármetros (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 $\hat{\beta}_0$ y $\hat{\beta}_1$ a los estimadores de β_0 y β_1 . Los valores $\hat{\beta}_0$ y $\hat{\beta}_1$ corresponderán a la recta de ordenada al origen $\hat{\beta}_0$ y pendiente $\hat{\beta}_1$ que “mejor ajuste” a los datos $(X_1, Y_1), \dots, (X_n, Y_n)$ 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

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

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 $(X_1, Y_1), \dots, (X_n, Y_n)$ 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 $\hat{\beta}_0$ y $\hat{\beta}_1$ que anulan sus derivadas. Sus derivadas son

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

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

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

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

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad (11)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \quad (12)$$

La pendiente estimada también se puede escribir de la siguiente forma

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

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 $\hat{\beta}_0$ y $\hat{\beta}_1$ no sea sólo un punto crítico, sino también un mínimo. Afortunadamente, en la práctica, los cálculos para hallar a $\hat{\beta}_0$ y $\hat{\beta}_1$ 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*

$$g(a, b) = \text{mediana} \{ [Y_1 - (a + bX_1)]^2, \dots, [Y_n - (a + bX_n)]^2 \}$$

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

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

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

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

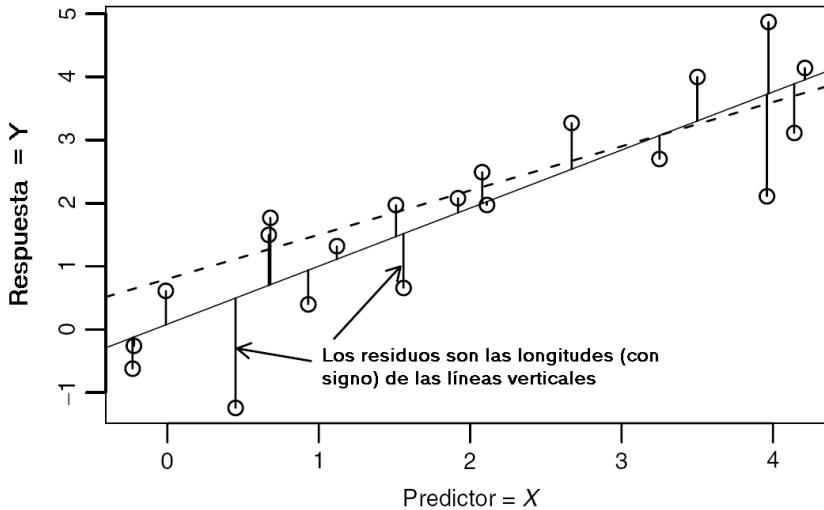
Definición 2.1 El valor $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ calculado para el valor X_i observado se denomina (*valor*) **predicho o ajustado i-ésimo**.

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

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

El residuo i-ésimo representa la distancia vertical entre el punto observado (X_i, Y_i) y el punto predicho por el modelo ajustado, (X_i, \hat{Y}_i) , 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

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

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

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

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 covariante 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

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

y el residuo sería

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

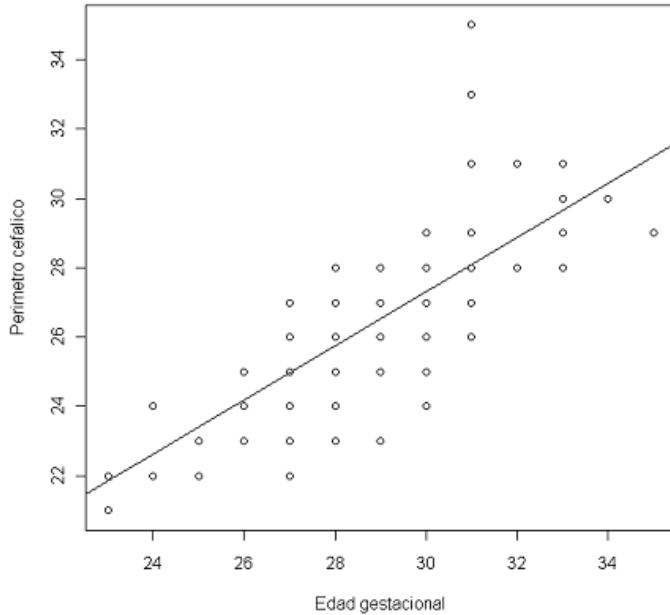
Si quisieramos predecir el valor del perímetro cefálico medio para un bebé con 29 semanas de gestación ($i = 1$), nuestro valor predicho sería

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

y el residuo sería

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

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

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

y el residuo sería

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

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)	$Y_i = \text{(headcirc)}$	$X_i = \text{(gestage)}$	\hat{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_regralinealsimple1.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 covariante. 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 $\log_{10}(\text{BrainWt})$ en función del $\log_{10}(\text{BodyWt})$. Como antes, escriba el modelo teórico y el ajustado. Una observación: en el `help` del `openintro` 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 superpongale 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

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

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

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i, \quad \text{Modelo ajustado.} \quad (14)$$

ya que los residuos se definen por $e_i = Y_i - \hat{Y}_i$. 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

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

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

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

La segunda igualdad de (16) se debe a que por (15) el promedio de los residuos \bar{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,

$$r = r((X_1, e_1), \dots, (X_n, e_n)) = \frac{\sum_{i=1}^n (e_i - \bar{e})(X_i - \bar{X})}{\sqrt{\sum_{i=1}^n (e_i - \bar{e})^2} \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}} = 0.$$

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

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (e_i - \bar{e})^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2. \quad (17)$$

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:

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

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 muestrados 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 $E(\hat{\beta}_1 | X_1, \dots, X_n)$. Para el estimador $\hat{\beta}_1$ 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

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

y² también

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

Un estimador de la varianza es

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

Finalmente, bajo los supuestos del modelo, puede probarse que

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

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

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

Esto es lo que se conoce como la distribución de muestreo de $\hat{\beta}_1$. 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 $se_{\hat{\beta}_1}$ para el error estándar de $\hat{\beta}_1$, o sea $se_{\hat{\beta}_1}$ es un estimador de la desviación estándar de la distribución de muestreo de $\hat{\beta}_1$.

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

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

²En realidad, $E(\hat{\beta}_1) = \beta_1$ y $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$, 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 , $Y_i | X_i \sim N(\beta_0, \sigma^2)$, 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 :

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

donde

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

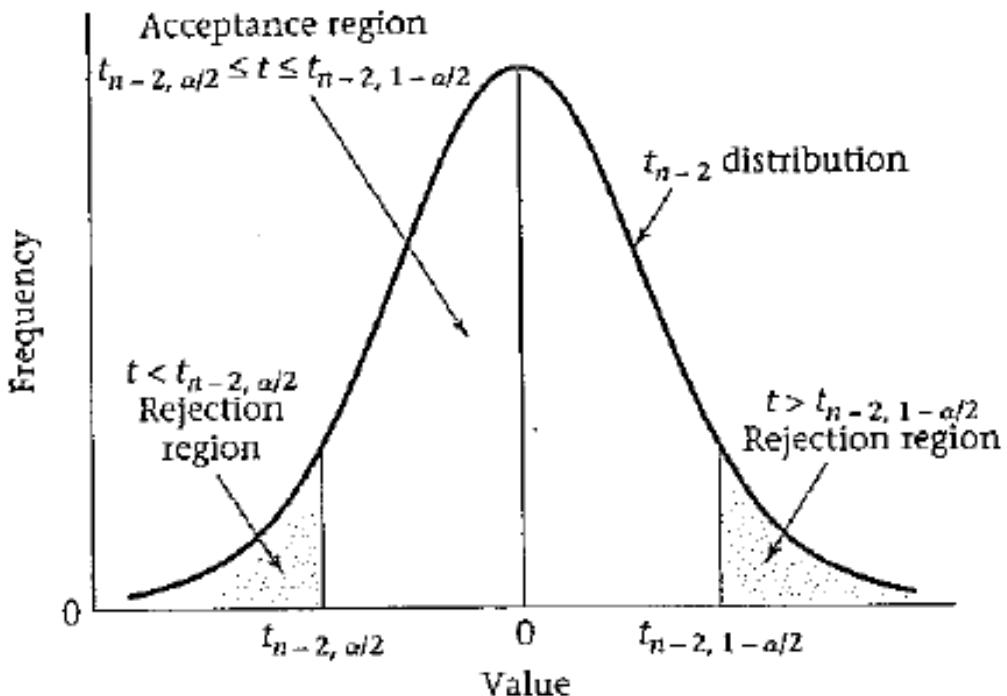
Entonces, la distribución de $\hat{\beta}_1$ será normal. Si el supuesto de normalidad de los errores no valiera, pero la muestra fuera suficientemente grande, y se cumpliera una condición sobre los c_i la distribución de $\hat{\beta}_1$ seguiría siendo aproximadamente normal. Luego, si $\beta_1 = 0$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

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

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

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

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

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

$$p\text{-valor} = 2P(T \geq |T_{obs}|),$$

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 $\hat{\beta}_1$, o sea

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

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

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

Finalmente,

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

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

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

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 $\hat{\beta}_1$ 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

$$se_{\hat{\beta}_1}^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\hat{\sigma}^2}{\left[\sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1) \right]} \cdot \frac{1}{(n - 1)}.$$

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 estandar del estimador de β_1 :

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

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

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

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

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

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

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

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

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

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

$$\hat{\beta}_0 \pm t_{n-2; \frac{\alpha}{2}} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}} \quad (23)$$

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 $E(Y_h | X = x_h)$ 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

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

El estimador puntual está dado por

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

La esperanza y la varianza de dicho estimador son

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

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

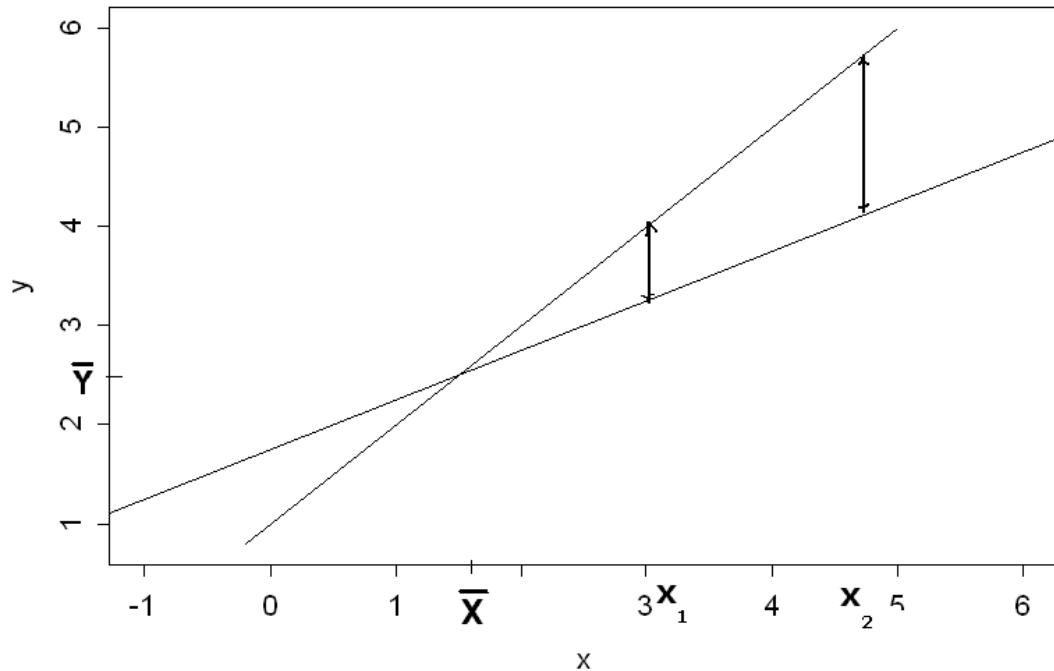
- por σ^2 , la variabilidad de las Y 's cuando conocemos el valor de X ,
- y por cuan lejos está ese valor particular de x_h (de X) del promedio observado en la muestra \bar{X} .

Notar que la variabilidad de la estimación de $E(Y_h | X = x_h)$ será menor cuanto más cercano a la media muestral \bar{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, \dots, 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 (\bar{X}, \bar{Y}) , ya que

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{X} = \underbrace{\bar{Y} - \hat{\beta}_1 \bar{X}}_{\hat{\beta}_0} + \hat{\beta}_1 \bar{X} = \bar{Y}$$

Luego, si tomamos muchas muestras de observaciones $(X_1, Y_1), \dots, (X_n, Y_n)$ con los mismos valores X_1, \dots, X_n , resultará que el valor \bar{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 (\bar{X}, \bar{Y}) , se observa la variabilidad mayor en el valor predicho (o ajustado) para $E(Y | X = x_2)$ que para $E(Y | X = x_1)$ si la distancia al \bar{X} es mayor para x_2 que para x_1 .



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

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

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

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

$$\hat{Y}_h \pm t_{n-2; 1-\frac{\alpha}{2}} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}. \quad (24)$$

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 incertezza 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

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

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 , $E(Y_h | X = x_h)$): 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 $Y_{h(\text{nueva})}$ 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

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

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

$$\hat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

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

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

2.12.1. Aplicación al ejemplo

Calculemos los intervalos de confianza de nivel 0,95 para $E(Y_h | X = x_h)$ 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 $E(Y_h | X = x_h)$ se obtiene por

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

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

$$\bar{X} = 28,89$$

$$S_{XX} = 635,69$$

$$n = 100$$

La varianza estimada por la regresión es

$$\hat{\sigma}^2 = \frac{SSRes}{n-2} = 2,529$$

de dónde surge

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

y

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

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

$$\begin{aligned} \hat{Y}_h &\pm t_{n-2;\frac{\alpha}{2}} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \\ 26,537 &\pm 1,984467 \cdot 1,5903 \cdot \sqrt{\frac{1}{100} + \frac{(29 - 28,89)^2}{635,69}} \\ 26,537 &\pm 0,3159 \\ [26,22; & 26,85] \end{aligned}$$

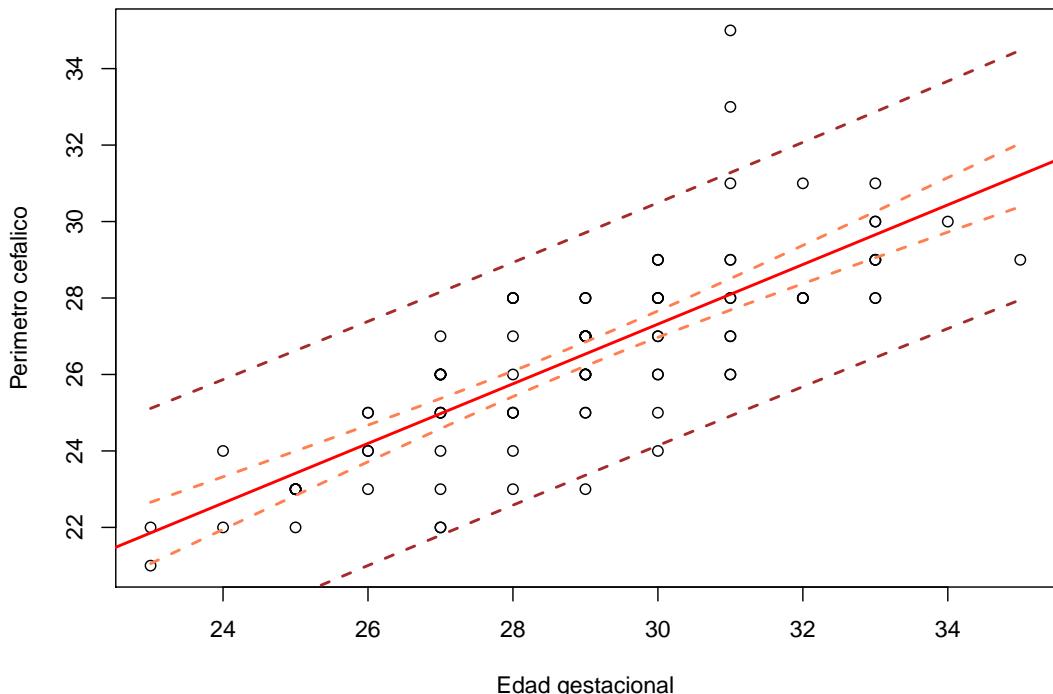
que coincide con lo hallado por el R: [26,21989; 26,85172].

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

$$\begin{aligned} \hat{Y}_h &\pm t_{n-2;\frac{\alpha}{2}} \cdot \hat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \\ 26,537 &\pm 1,984467 \cdot 1,5903 \cdot \sqrt{1 + \frac{1}{100} + \frac{(29 - 28,89)^2}{635,69}} \\ 26,537 &\pm 3,1717 \\ [23,365; & 29,709] \end{aligned}$$

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 $E(Y_h | X = \bar{X})$ 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 $\frac{(x_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$ 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, $E(Y | X = x) = \beta_0 + \beta_1 x$. 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)

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

donde

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

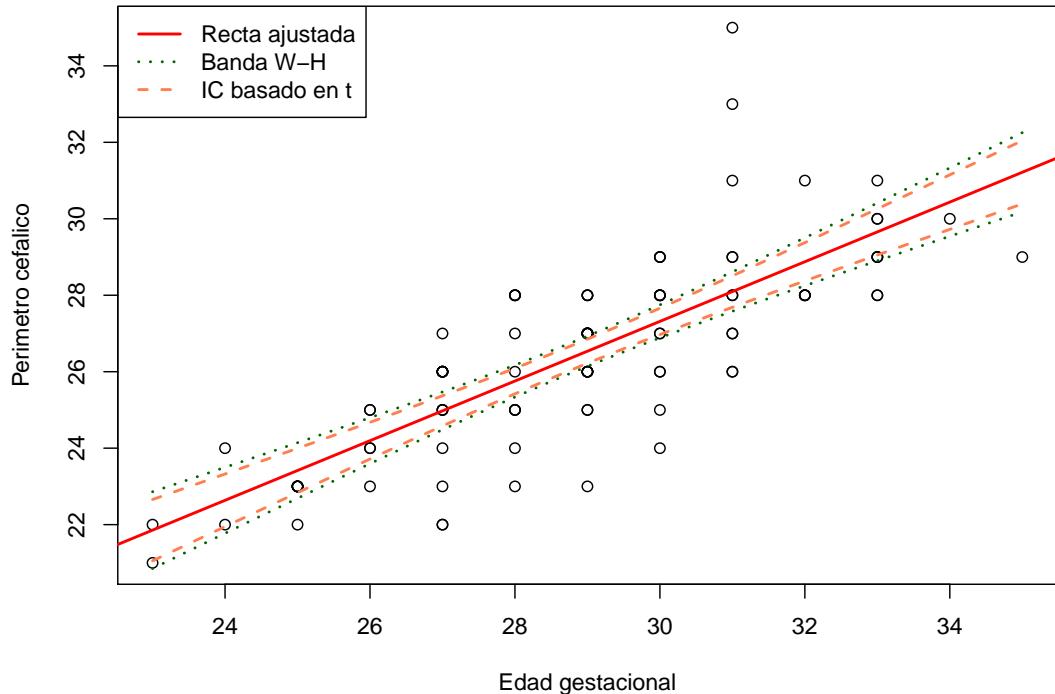
donde $F_{1-\alpha,2,n-2}$ es el cuantil $1 - \alpha$ de una distribución F de Fisher con 2 grados de libertad en el numerador y $(n - 2)$ en el denominador, que en R se calcula usando el comando `qf(1 - alpha, df1 = 2, df2 = n - 2)`. Observemos que la fórmula (25) tiene la misma forma que (24) para 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 \bar{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, \dots, Y_n . A falta de algo mejor proponemos el modelo

Modelo A: $E(Y | X) = \mu$, o escrito de otro modo

Modelo A: $Y_i = \mu + u_i$ con $u_i \sim N(0, \sigma_Y^2)$, $1 \leq i \leq n$,
independientes entre sí.

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

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

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 $\sum_{i=1}^n (Y_i - \bar{Y})^2$ 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 $(Y_1 - \bar{Y}), \dots, (Y_{n-1} - \bar{Y})$ puede deducir el valor de $(Y_n - \bar{Y})$ pues todos ellos suman 0.

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

gresión tenemos el modelo

$$\text{Modelo B: } E(Y | X) = \beta_0 + \beta_1 X, \text{ o escrito de otro modo} \quad (26)$$

$$\text{Modelo B: } Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{con } \varepsilon_i \sim N(0, \sigma^2), \quad 1 \leq i \leq n, \\ \text{independientes entre sí.}$$

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

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

es decir, la variación de las observaciones alrededor de la recta ajustada. Como ya comentamos, a la cantidad $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ 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 $(Y_1 - \hat{Y}_1), \dots, (Y_{n-2} - \hat{Y}_{n-2})$ puede deducir el valor de $e_{n-1} = (Y_{n-1} - \hat{Y}_{n-1})$ y $e_n = (Y_n - \hat{Y}_n)$ 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 incluido 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. Cuanto mayor sea la variación de las Y_i alrededor de la recta ajustada, mayor será la SSResiduos.

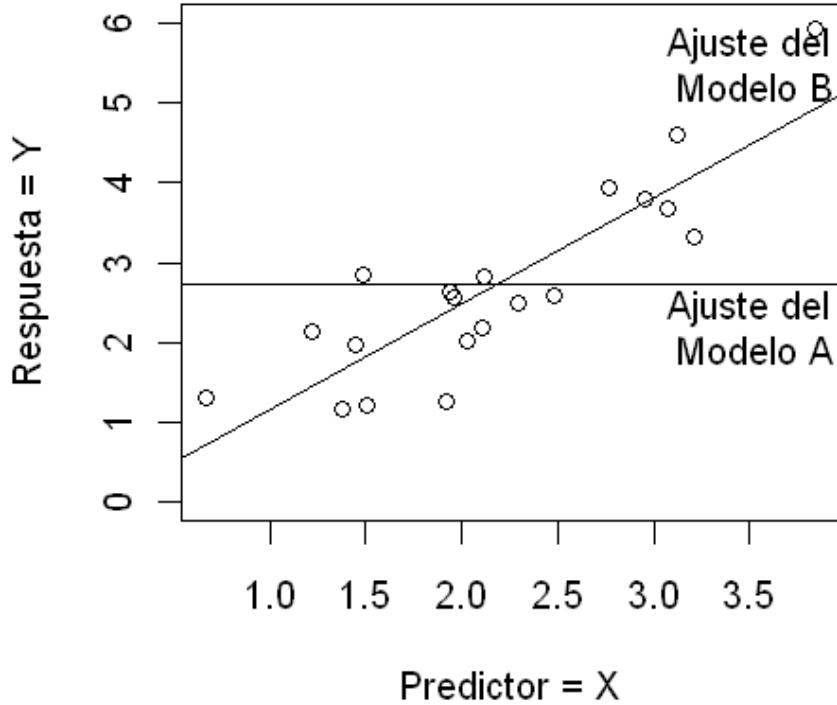
¿Cuál de las dos será mayor: SSTotal o SSRes? Vale que

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

pues $\hat{\beta}_0$ y $\hat{\beta}_1$ 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

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

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



Por lo tanto,

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

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

$$\text{SSRes} \leq \sum_{i=1}^n (Y_i - \bar{Y})^2 = \text{SSTo}. \quad (28)$$

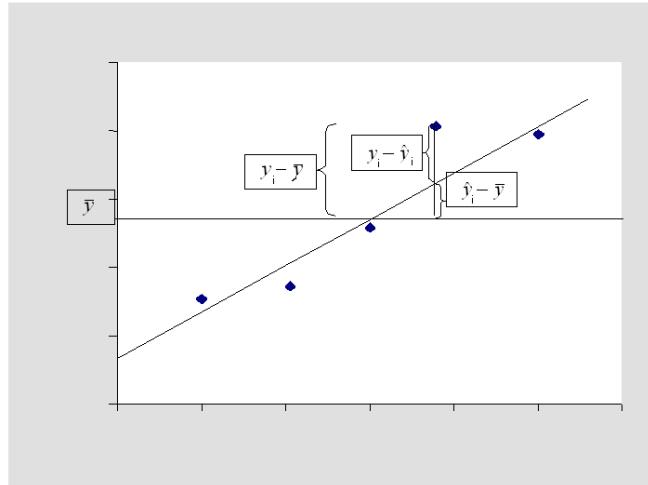
Podemos interpretar a SSTo como una medida de la variabilidad de las Y que no queda explicada por el modelo A. Es una medida del desajuste del modelo A a los datos. Lo mismo puede decirse de SSRes: es una medida de la variabilidad

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

$$\underbrace{Y_i - \bar{Y}}_{\text{desviación total}} = \underbrace{Y_i - \hat{Y}_i}_{\substack{\text{desvío alrededor} \\ \text{de la recta de regresión}}} + \underbrace{\hat{Y}_i - \bar{Y}}_{\substack{\text{desvío de los predichos} \\ \text{respecto de la media} \\ \text{ajustada}}} \quad (29)$$

En la Figura 23 vemos estas diferencias graficadas para una observación. La desviación total $Y_i - \bar{Y}$ mide la distancia vertical (con signo) de la observación a la recta horizontal que corta al eje vertical en \bar{Y} , $Y_i - \hat{Y}_i$ 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 $\hat{Y}_i - \bar{Y}$ 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.



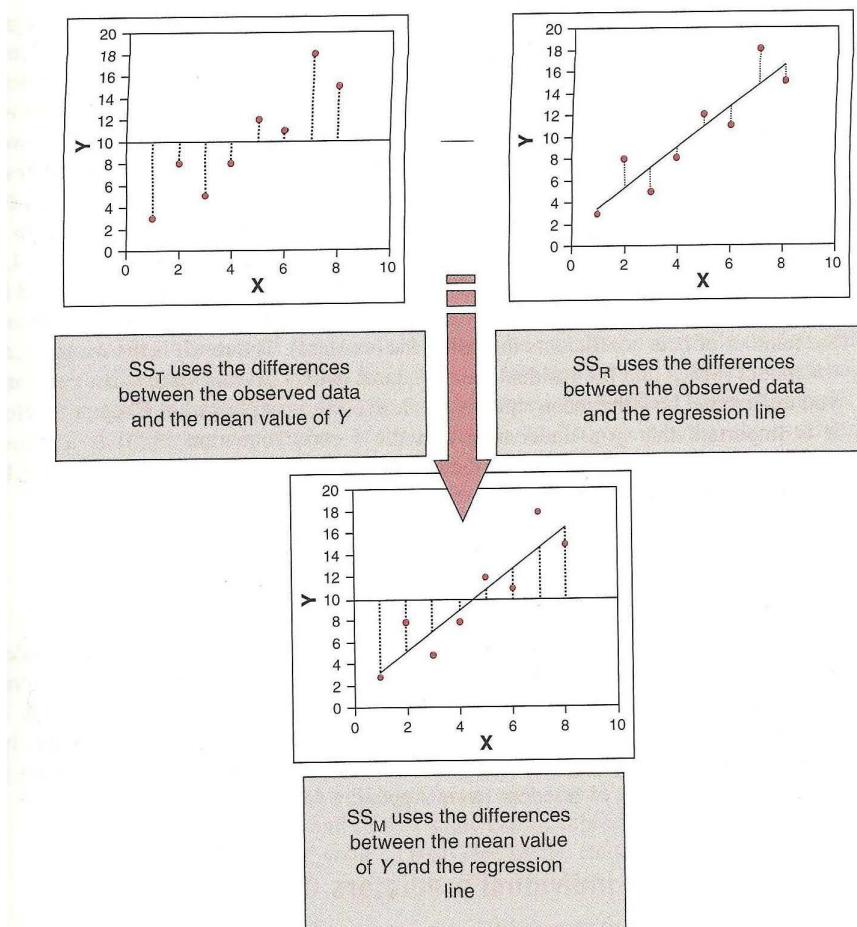
Obviamente es falso que el cuadrado del término de la izquierda en la igualdad (29) anterior sea igual a la suma de los cuadrados de los términos de la derecha es decir,

$$(Y_i - \bar{Y})^2 \neq (Y_i - \hat{Y}_i)^2 + (\hat{Y}_i - \bar{Y})^2 \quad \text{para cada } i.$$

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

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2. \quad (30)$$

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 \bar{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 \bar{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

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

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	$\frac{\text{MSReg}}{\text{MSRes}}$	$P(F_{1,n-2} \geq F_{obs})$
Residuos	SSRes	$n - 2$	MSRes		
Total	SSTo	$n - 1$			

donde

$$\begin{aligned} \text{SSReg} &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 & \text{MSReg} &= \frac{\text{SSReg}}{1} \\ \text{SSRes} &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 & \text{MSRes} &= \frac{\text{SSRes}}{n-2} \\ \text{SSTo} &= \sum_{i=1}^n (Y_i - \bar{Y})^2 & F &= \frac{\text{MSReg}}{\text{MSRes}} = \frac{\text{SSReg}(n-2)}{\text{SSRes}} \end{aligned}$$

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

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

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

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

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:

$$\begin{array}{ccc} 100\% \text{ de variabilidad} & \text{---} & SSTo \\ \% \text{ de variabilidad explicada} & \text{---} & SSTo - SSRes \end{array}$$

Luego el porcentaje de variabilidad explicada es

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

A la cantidad

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

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 $\{\hat{(Y_i, Y_i)}\}_{1 \leq i \leq n}$, 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 $\text{MSRes} = \text{SSRes}/(n - 2)$ tiene esperanza σ^2 .
2. La distribución de muestreo de $\text{MSReg} = \text{SSReg}/1$ tiene esperanza $\sigma^2 + \beta_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$.

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

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

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

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

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

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

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 $\hat{\beta}_0$ (12) tenemos*

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

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

$$T = \frac{\hat{\beta}_1 - 0}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} = \frac{\hat{\beta}_1}{se_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\sqrt{\frac{SSRes/(n-2)}{\sum_{i=1}^n (X_i - \bar{X})^2}}}$$

y el estadístico F que resulta ser

$$F = \frac{MSReg}{MSRes} = \frac{\frac{\hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2}{n-2}}{\frac{SSRes}{n-2}} = \frac{\hat{\beta}_1^2}{\frac{SSRes/(n-2)}{\sum_{i=1}^n (X_i - \bar{X})^2}},$$

vemos que

$$F = T^2$$

y el *p*-valor del test *t* se calculaba

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

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_regrlinealsimple2.R`

Ejercicio 2.5 *Medidas del cuerpo, Parte IV. Base de datos `bdimis` 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 covariante, 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 covariante, 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 contendrá 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 intervalos entre sí. Compárelas 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.
- 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).
 - Si cada madre e hija tuvieran exactamente la misma altura que su hija, ¿cómo luciría este scatterplot? Resuma lo que observa en este gráfico. Superpongale 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. Superpongala 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 pendiente. Testee la hipótesis $E(Dheight | Mheight) = \beta_0$ versus la alternativa que $E(Dheight | Mheight) = \beta_0 + \beta_1 Mheight$. 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*

$$Y = 10 + 5X + \varepsilon, \quad (31)$$

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, $\hat{\beta}_1$? ¿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) Superpongale 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 $\hat{\beta}_1$ 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, \dots, 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, \dots, 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 $\log_{10}(\text{BrainWt})$ en función del $\log_{10}(\text{BodyWt})$,

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

Una observación: en el `help` del `openintro` 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 $\log_{10}(\text{BodyWt})$ 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 $\log_{10}(\text{BrainWt})$ 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 $\log_{10}(\text{BrainWt})$. Prediga el peso del cerebro de dicho animal. Ahora queremos convertir el intervalo de predicción del $\log_{10}(\text{BrainWt})$ en un intervalo de predicción para el `BrainWt`. Para eso, observemos que si el intervalo (a, b) es un intervalo de predicción de nivel 95 % para $\log_{10}(\text{BrainWt})$, entonces, un intervalo para el `BrainWt` está dado por $(10^a, 10^b)$. ¿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 100kg, no es posible hacer la conversión del ítem anterior de manera automática, ya que para cualquier función 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 (*small-mouth 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 escamas, 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?