# Modelos Estadísticos Interpretables
# Tópicos de Modelos Interpretables
## 2. Regresión Lineal Múltiple

María Eugenia Szretter Noste

Instituto de Cálculo
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

# Modelo de Regresión Lineal Múltiple

Tenemos $(Y, X_1, X_2, \ldots, X_{p-1})^T \in \mathbb{R}^p$ un vector aleatorio, todas variables medidas en el mismo individuo. El modelo de regresión lineal múltiple se ocupa de modelar la relación entre

$X_1, X_2, \ldots, X_{p-1}$ variables explicativas e $Y$ variable respuesta

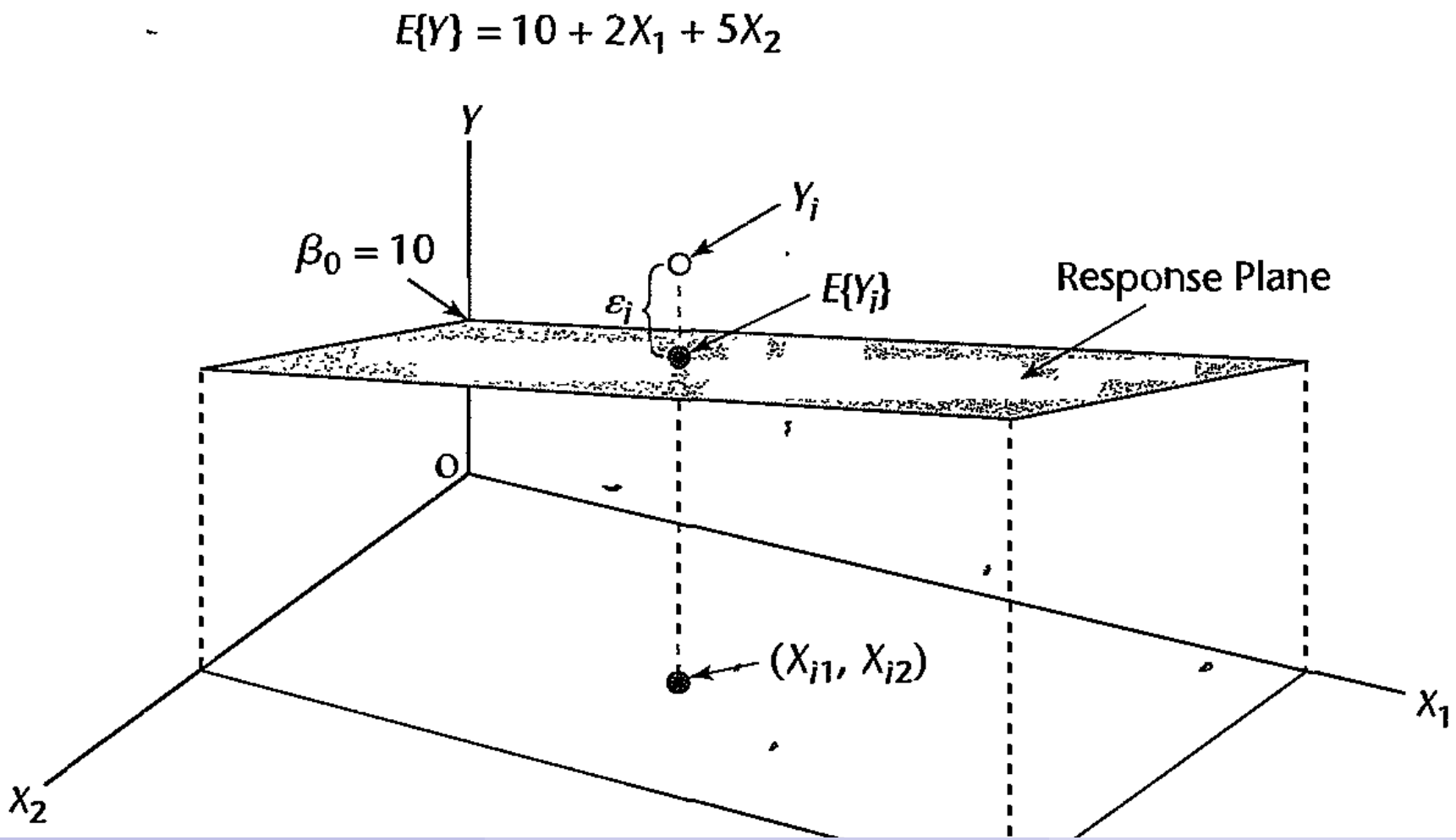
Notación: $\mathbf{X} = (X_1, X_2, \ldots, X_{p-1})^T$ . En negrita, los vectores, que los pensamos siempre como vectores columna. Escribimos el modelo en términos de la esperanza y varianza condicional:

$$E[Y \mid X_1, X_2, \ldots, X_{p-1}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} \quad (1)$$

$$Var[Y \mid X_1, X_2, \ldots, X_{p-1}] = \sigma^2 \quad (2)$$

# Función de respuesta

Figura 1: En regresión lineal con dos variables explicativas la función de respuesta es un plano.



# Modelo de Regresión Lineal Múltiple

Escribamos el modelo de regresión lineal múltiple cuando se tiene observaciones: $\{(Y_i, X_{i1}, \ldots, X_{i,p-1})\}_{1 \leq i \leq n}$, es un modelo para la variable aleatoria $Y$ cuando se conocen $X_1, X_2, \ldots, X_{p-1}$ las variables regresoras. El modelo es

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \tag{3}$$

donde $\beta_0, \beta_1, \ldots, \beta_{p-1}$ son parámetros (es decir, números) desconocidos, $X_{i1}, X_{i2}, \ldots, X_{i,p-1}$ son los valores de las variables predictoras medidas en el i-ésimo individuo (o i-ésima repetición del experimento) con $1 \leq i \leq n$, $n$ es el tamaño de muestra, $Y_i$ es la variable respuesta medida en el i-ésimo individuo y $\varepsilon_i$ es el error para el individuo i-ésimo, que no es observable. Asumimos

- $E[\varepsilon_i \mid X_{i1}, X_{i2}, \ldots, X_{i,p-1}] = 0$
- $V[\varepsilon_i \mid X_{i1}, X_{i2}, \ldots, X_{i,p-1}] = \sigma^2$

# Modelo de Regresión Lineal en notación matricial

Queremos expresar el modelo (25) de forma matricial. ¿Por qué? puesto que éste es el tratamiento estándar del tema, y además porque refleja los conceptos esenciales en el ajuste del modelo. Definimos

$$\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \underset{n\times p}{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} \tag{4}$$

$$\underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \underset{n\times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Cada fila de la matriz $X$ corresponde a las observaciones correspondientes a cada individuo (la fila i-ésima contiene las observaciones del individuo i-ésimo) y las columnas identifican a las variables.

# Modelo de Regresión Lineal en notación matricial

El modelo (25) que copiamos acá para tener presente

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i,$$

se escribe matricialmente en la siguiente forma

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{X}\ \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

donde

- $\mathbf{Y}$ es un vector de respuestas
- $\boldsymbol{\beta}$ es un vector de parámetros
- $X$ es una matriz de covariables
- $\boldsymbol{\varepsilon}$ es un vector de variables aleatorias

Puede ser cómodo notar con una notación vectorial al vector (o matriz de $1 \times p$) de covariables observado (algunos libros incluyen el 1 y otros no, ojo) $\boldsymbol{x}_i^T = [1\, X_{i1}\, X_{i2}\, \cdots\, X_{i,p-1}]$, o lo que es lo mismo

$$\boldsymbol{x}_i = \begin{bmatrix} 1 \\ X_{i1} \\ \vdots \\ X_{i,(p-1)} \end{bmatrix}$$

Entonces

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

Y también tenemos

$$X\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{x}_1^T\boldsymbol{\beta} \\ \boldsymbol{x}_2^T\boldsymbol{\beta} \\ \vdots \\ \boldsymbol{x}_n^T\boldsymbol{\beta} \end{bmatrix}$$

Tenemos: $E(\varepsilon \mid X) = \mathbf{0}$ y matriz de varianzas y covarianzas

$$Var(\varepsilon \mid X) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}.$$

Recordemos que operar condicional a las variables equis es como tomar las variables equis como fijas. Una formulación equivalente es:

$$E\left(\mathbf{Y} \mid \underbrace{X}_{\text{esta } X \text{ es una matriz}}\right) = X\beta$$

y la matriz de covarianza de las $\mathbf{Y}$ resulta ser la misma que la de $\varepsilon$

$$Var(\mathbf{Y} \mid X) = \sigma^2 \mathbf{I}.$$

Al igual que hicimos con el modelo de regresión simple, muchas veces omitiremos la condicionalidad a las equis en la notación, es decir, escribiremos $E(\mathbf{Y})$ en vez de $E(\mathbf{Y} \mid X)$.

# Regresión Lineal Múltiple: Estimación de los Parámetros

Usamos el método de mínimos cuadrados para ajustar el modelo. O sea, definimos el error cuadrático muestral cuando usamos a $b_0 + b_1 X_1 + \cdots + b_{p-1} X_{p-1}$ para predecir a $Y$, $\widehat{ECM}(Y, b_0, b_1, \ldots, b_{p-1})$ que notamos (escribimos $X_{i0} = 1$)

$$g(b_0, b_1, \ldots, b_{p-1}) = \sum_{i=1}^{n} (Y_i - b_0 X_{i0} - b_1 X_{i1} - b_2 X_{i2} - \cdots - b_{p-1} X_{ip-1})^2 \tag{5}$$

y los estimadores $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_{p-1}$ serán aquellos valores de $b_0, b_1, \ldots, b_{p-1}$ que minimicen a $g$. Los denominamos estimadores de mínimos cuadrados. Denotaremos al vector de coeficientes estimados por $\widehat{\boldsymbol{\beta}}$ :

$$\underset{p\times 1}{\widehat{\boldsymbol{\beta}}} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_{p-1} \end{bmatrix}$$

# Regresión Lineal Múltiple: Estimación de los Parámetros

Para hallar los estimadores de mínimos cuadrados derivamos a $g(b_0, b_1, \ldots, b_{p-1}) = g(\mathbf{b})$

$$g(\mathbf{b}) = \sum_{i=1}^{n} (Y_i - b_0 X_{i0} - b_1 X_{i1} - b_2 X_{i2} - \cdots - b_{p-1} X_{ip-1})^2$$

$$\frac{\partial}{\partial b_j} g(\mathbf{b}) = \sum_{i=1}^{n} 2(Y_i - b_0 X_{i0} - b_1 X_{i1} - b_2 X_{i2} - \cdots - b_{p-1} X_{ip-1})(-X_{ij})$$

para cada $j = 0, 1, \ldots, (p-1)$. Esto se puede escribir cómodamente de forma matricial:

$$\nabla g(\mathbf{b}) = \begin{bmatrix} \frac{\partial}{\partial b_0} g(\mathbf{b}) \\ \frac{\partial}{\partial b_1} g(\mathbf{b}) \\ \vdots \\ \frac{\partial}{\partial b_{p-1}} g(\mathbf{b}) \end{bmatrix} = -2X^T(\mathbf{Y} - X\mathbf{b})$$

# Regresión Lineal Múltiple: Estimación de los Parámetros

$$\text{(copiamos)} \qquad \nabla g(\mathbf{b}) = -2X^T(\mathbf{Y} - X\mathbf{b})$$

Igualamos a cero:

$$X^T\mathbf{Y} - X^TX\widehat{\boldsymbol{\beta}} = \mathbf{0}$$

Las **ecuaciones normales (o de estimación)** de mínimos cuadrados para el modelo de regresión lineal múltiple son

$$X^TX\widehat{\boldsymbol{\beta}} = X^T\mathbf{Y} \tag{6}$$

donde $X^T$ quiere decir la matriz traspuesta. Algunos autores lo notan $X^t$, o bien $X'$ (recordemos que la matriz traspuesta es aquella matriz $p \times n$ que tiene por filas a las columnas de $X$). Los estimadores de mínimos cuadrados son

$$\widehat{\boldsymbol{\beta}} = \left(X^TX\right)^{-1}X^T\mathbf{Y}$$

siempre que la inversa de la matriz $X^TX$ exista.

# Regresión Lineal Múltiple: Estimación de los Parámetros

La inversa de una matriz existe si la matriz tiene rango completo. Además $rg(X^TX) = rg(X)$ y también, tenemos $\dim(X) = n \times p$, por lo que $\dim\left(X^TX\right) = p \times p$. Por lo que vamos a poder hallar los estimadores de mínimos cuadrados de $\beta$ siempre que $rg(X) = p$.

## Ejercicio 2.1

*Probar, construyendo la matriz $X$ y los vectores $\mathbf{Y}$ y $\widehat{\beta}$ en el caso del modelo de regresión lineal simple, la expresión $\widehat{\beta} = \left(X^TX\right)^{-1}X^T\mathbf{Y}$ coincide con los estimadores de mínimos cuadrados previamente obtenida.*

# Cálculo de los estimadores

Aunque la expresión que encontramos para $\widehat{\boldsymbol{\beta}}$ involucra el cálculo de la inversa de una matriz, los paquetes estadísticos no suelen invertir a $X^T X$ para calcularla, ya que suele no ser lo más eficiente numéricamente. En cambio suele despejarse a $\widehat{\boldsymbol{\beta}}$ a partir del sistema de ecuaciones normales por alguno de los siguientes métodos:

1. Eliminación de Gauss
2. Método de Cholesky
3. Descomposición QR

# Modelo de Regresión Lineal Múltiple

Haremos supuestos sobre los errores

$$\varepsilon_i \sim N\left(0, \sigma^2\right), 1 \leq i \leq n, \quad \text{independientes entre sí.} \tag{7}$$

Es decir,

1. los $\varepsilon_i$ tienen media cero, $E\left(\varepsilon_i\right) = 0$.
2. los $\varepsilon_i$ tienen todos la misma varianza desconocida que llamaremos $\sigma^2$ y que es el otro parámetro del modelo, $Var\left(\varepsilon_i\right) = \sigma^2$.
3. los $\varepsilon_i$ tienen distribución normal.
4. los $\varepsilon_i$ son independientes entre sí, e independientes de las covariables $X_{i1}, X_{i2}, \ldots, X_{ip}$.

Calculemos esperanzas y varianzas de los estimadores usando las propiedades 1 y 2 sobre los errores. En lo que sigue, pensamos a las X's como fijas o condicionamos a ellas (todas las esperanzas y varianzas que siguen pueden pensarse condicionales a la matriz $X$ de observaciones)

$$\text{Modelo} \quad \boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\text{Tomando esperanza} \quad E(\boldsymbol{Y}) = E(X\boldsymbol{\beta}) + E(\boldsymbol{\varepsilon}) = X\boldsymbol{\beta} \tag{8}$$

$$\text{Tomando varianza} \quad Var(\boldsymbol{Y}) = Var(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = Var(\boldsymbol{\varepsilon}) = \sigma^2 I_n \tag{9}$$

Estimadores de $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T \mathbf{Y}$$

$$E\left(\widehat{\boldsymbol{\beta}}\right) = E\left(\left(X^T X\right)^{-1} X^T \mathbf{Y}\right) = \left(X^T X\right)^{-1} X^T E(\mathbf{Y})$$

$$= \left(X^T X\right)^{-1} X^T X\boldsymbol{\beta} = \boldsymbol{\beta} \tag{10}$$

Luego $\widehat{\boldsymbol{\beta}}$ es un estimador insesgado de $\boldsymbol{\beta}$.

$$
\begin{aligned}
Var\left(\widehat{\boldsymbol{\beta}}\right) &= Var\left(\left(X^TX\right)^{-1}X^T\mathbf{Y}\right) \\
&= \left(X^TX\right)^{-1}X^T Var\left(\mathbf{Y}\right)\left[\left(X^TX\right)^{-1}X^T\right]^T \\
&= \left(X^TX\right)^{-1}X^T Var\left(\mathbf{Y}\right)X\left(X^TX\right)^{-1} \\
&= \left(X^TX\right)^{-1}X^T\sigma^2 I_n X\left(X^TX\right)^{-1} \\
&= \sigma^2\left(X^TX\right)^{-1}X^TX\left(X^TX\right)^{-1} = \sigma^2\left(X^TX\right)^{-1} \qquad (11)
\end{aligned}
$$

Finalmente, de (10) y (11) tenemos, para $\boldsymbol{a} \in \mathbb{R}^p$ que

$$
E\left(\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{a}^T E\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{a}^T\boldsymbol{\beta} \in \mathbb{R} \qquad (12)
$$

$$
Var\left(\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{a}^T Var\left(\widehat{\boldsymbol{\beta}}\right)\boldsymbol{a} = \sigma^2\boldsymbol{a}^T\left(X^TX\right)^{-1}\boldsymbol{a} \in \mathbb{R} \qquad (13)
$$

# Si miramos las coordenadas de $\widehat{\boldsymbol{\beta}}$

Tomando $\boldsymbol{a}$ como un vector canónico

$\boldsymbol{a}^T = \boldsymbol{e}_{j+1}^T = \left(0, \ldots, 0, \underbrace{1}_{j+1}, 0, \ldots, 0\right)$ obtenemos

$$\boldsymbol{a}^T \widehat{\boldsymbol{\beta}} = \left(0, \ldots, 0, \underbrace{1}_{j+1}, 0, \ldots, 0\right) \cdot \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_{p-1} \end{bmatrix} = \widehat{\beta}_j$$

que, usando la ecuación (12) tiene esperanza

$$E\left(\widehat{\beta}_j\right) = E\left(\boldsymbol{a}^T \widehat{\boldsymbol{\beta}}\right) = E\left(\boldsymbol{e}_{j+1}^T \widehat{\boldsymbol{\beta}}\right) = \boldsymbol{e}_{j+1}^T \boldsymbol{\beta} = \beta_j$$

Las ecuación (13) nos dice,

$$
\begin{aligned}
Var\left(\widehat{\beta}_j\right) &= Var\left(\boldsymbol{e}_{j+1}^T\widehat{\boldsymbol{\beta}}\right) = \sigma^2\boldsymbol{e}_{j+1}^T\left(X^TX\right)^{-1}\boldsymbol{e}_{j+1} \\
&= \sigma^2\boldsymbol{e}_{j+1}^T \cdot \left(\text{columna (j+1) de } \left(X^TX\right)^{-1}\right) \\
&= \sigma^2\left(0,\ldots,0,\underbrace{1}_{j+1},0,\ldots,0\right) \cdot \left(\text{columna (j+1) de } \left(X^TX\right)^{-1}\right) \\
&= \sigma^2\left[\left(X^TX\right)^{-1}\right]_{(j+1),(j+1)}
\end{aligned}
$$

Si la comparamos con el caso regresión lineal simple ($p = 2$),

$$
Var\left(\hat{\beta}_1 \mid X_1,\ldots,X_n\right) = \frac{\sigma^2}{n\,s_X^2} = \frac{\sigma^2}{\sum_{i=1}^n\left(X_i-\overline{X}\right)^2} \tag{14}
$$

Se puede probar que coinciden.

Otra forma de escribir los supuestos es a través de la distribución de $Y$ condicional a $\boldsymbol{x} = (1, X_1, \ldots, X_{p-1})$. Está dada por

$$Y \mid \boldsymbol{x} \sim N\left(\boldsymbol{x}^{\boldsymbol{T}}\beta, \sigma^2\right). \tag{15}$$

También podemos pensar que las $X's$ son fijas, y no escribir el condicional. O seguir pensando que son aleatorias pero que todo lo que hacemos es condicional a $X's$ y no escribirlas.

Como las observaciones distintas son independientes, apelando a la distribución normal multivariada, de (15) podemos deducir la distribución del vector de errores:

$$\varepsilon \sim N_n\left(\mathbf{0}, \sigma^2\, I_n\right).$$

A partir de esto podemos deducir la distribución del vector $\boldsymbol{Y} \in \mathbb{R}^n$ condicional a conocer todas las observaciones $X \in \mathbb{R}^{n \times p}$

$$\boldsymbol{Y} \mid X \sim N_n\left(X^T\beta, \sigma^2\, I_n\right). \tag{16}$$

# Modelo ajustado, valores predichos y residuos

## Definición 2.1

*El* **modelo ajustado** *es*

$$\widehat{m}(\boldsymbol{x}) = \boldsymbol{x}^T\widehat{\boldsymbol{\beta}} = [1\ X_1\ X_2\ \cdots\ X_{p-1}] \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_{p-1} \end{bmatrix}$$

$$= \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + \cdots + \widehat{\beta}_{p-1} X_{p-1}$$

*El* **valor predicho iésimo** *es* $\widehat{m}(\boldsymbol{x}_i) = \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}$ *que también notamos* $\widehat{Y}_i$ *y el* **residuo iésimo** *es* $r_i = Y_i - \widehat{m}(\boldsymbol{x}_i)$. *Hablamos del vector de residuos* $\boldsymbol{r}$ *y del vector de predichos* $\widehat{\boldsymbol{Y}}$.

Observemos que, al igual que para el modelo lineal simple, el residuo está contenido en $\mathbb{R}$ (o sea, no es un vector).

# Estimador de $\sigma^2$

Como en el caso de regresión lineal simple, un estimador de $\sigma^2 = Var(\epsilon_i)$ es

$$S^2 = \frac{1}{n-p}\sum_{i=1}^{n} r_i^2 = \frac{1}{n-p}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}\right)^2 = \frac{RSS}{n-p} \qquad (17)$$

A partir del vector de residuos $\boldsymbol{r} = \boldsymbol{Y} - X\widehat{\boldsymbol{\beta}} \in \mathbb{R}^n$, podemos escribir

$$S^2 = \frac{\boldsymbol{r}^T\boldsymbol{r}}{n-p} = \frac{1}{n-p}\left(\boldsymbol{Y} - X\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{Y} - X\widehat{\boldsymbol{\beta}}\right) = \frac{RSS}{n-p}$$

Bajo el modelo lineal y si valen los supuestos 1) y 2) sobre los errores, resulta que $S^2$ es un estimador insesgado de $\sigma^2$.

# Distribución de los estimadores y tests, para el modelo lineal

Asumimos que las $X$ están fijas.

## Teorema 2.1

*Si* $\mathbf{Y} \sim N_n(X\beta, \sigma^2 I_n)$ *con* $X \in \mathbb{R}^{n\times p}$ *de rango* $p$, *entonces:*

i. $\widehat{\beta} \sim N_p(\beta, \sigma^2(X^T X)^{-1})$.

ii. $\frac{1}{\sigma^2}(\widehat{\beta} - \beta)^T (X^T X)\left(\widehat{\beta} - \beta\right) \sim \chi^2_p$.

iii. $\widehat{\beta}$ *es independiente de*

$$S^2 = \frac{\boldsymbol{r}^T \boldsymbol{r}}{n-p} = \frac{1}{n-p}\left(\boldsymbol{Y} - X\widehat{\beta}\right)^T \left(\boldsymbol{Y} - X\widehat{\beta}\right) = \frac{RSS}{n-p}.$$

iv. $\frac{RSS}{\sigma^2} = \frac{(n-p)S^2}{\sigma^2} \sim \chi^2_{n-p}$.

# Una combinación lineal de los estimadores

## Teorema 2.2

*Si* $\mathbf{Y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I_n)$ *con* $X \in \mathbb{R}^{n\times p}$ *de rango* $p$, *si* $\mathbf{a} \in \mathbb{R}^p$, *sea*

$$T = \frac{\mathbf{a}^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{S\sqrt{\mathbf{a}^T(X^TX)^{-1}\mathbf{a}}} \sim t_{n-p}$$

*Luego* $T$ *es una expresión pivote para obtener intervalos de confianza para* $\mathbf{a}^T\boldsymbol{\beta}$ *y tests para* $H_0 : \mathbf{a}^T\boldsymbol{\beta} = c$, *con* $c \in \mathbb{R}$.

# Caso particular: tests e intervalos para un $\beta_j$

Bajo el modelo lineal, asumiendo que $\varepsilon_1, \ldots, \varepsilon_n$ tienen distribución normal con media cero y varianza $\sigma^2$, el test de $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$, se basa en

$$T = \frac{\widehat{\beta}_j}{\widehat{sd}\left(\widehat{\beta}_j\right)} \sim t_{n-p}, \text{ bajo } H_0.$$

El test de nivel $\alpha$ rechaza $H_0$ cuando $|T| \geq t_{n-p}\left(\frac{\alpha}{2}\right)$ donde $t_{n-p}\left(\frac{\alpha}{2}\right)$ es el percentil $1 - \frac{\alpha}{2}$ de una t de Student con $n - p$ grados de libertad. El p-valor del test es

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

Finalmente, un intervalo de confianza de nivel $1 - \alpha$ para $\beta_j$ está dado por

$$\widehat{\beta}_j \pm t_{n-p}\left(\frac{\alpha}{2}\right) \sqrt{\widehat{var}\left(\widehat{\beta}_j\right)}$$

# Tests e intervalos para una combinación lineal de los $\beta$

Bajo $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ y asumiendo que $\varepsilon_1, \ldots, \varepsilon_n$ tienen distribución normal con media cero y varianza $\sigma^2$, el test de $H_0 : \mathbf{a}^T\boldsymbol{\beta} = c$ versus $H_1 : \mathbf{a}^T\boldsymbol{\beta} \neq c$, donde $\mathbf{a}$ es un vector de dimensión $p$ prefijado (habitualmente se usan los canónicos, compuestos de un cero y los restantes unos) se basa en

$$T = \frac{\mathbf{a}^T\widehat{\boldsymbol{\beta}} - c}{S\sqrt{\mathbf{a}^T(X^TX)^{-1}\mathbf{a}}} \sim t_{n-p}, \text{ bajo } H_0.$$

El test de nivel $\alpha$ rechaza $H_0$ cuando $|T| \geq t_{n-p}\left(\frac{\alpha}{2}\right)$ donde $t_{n-p}\left(\frac{\alpha}{2}\right)$ es el percentil $1 - \frac{\alpha}{2}$ de una t de Student con $n - p$ grados de libertad. El p-valor del test es

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

Finalmente, un intervalo de confianza de nivel $1 - \alpha$ para la combinación lineal $\mathbf{a}^T\boldsymbol{\beta}$ está dado por

$$\mathbf{a}^T\widehat{\boldsymbol{\beta}} \pm t_{n-p}(\alpha/2)\, S\sqrt{\mathbf{a}^T(X^TX)^{-1}\mathbf{a}}$$

# Ejemplo: *low birth weight data*

Datos publicados en

Leviton, A., Fenton, T., Kuban, K. C., y Pagano, M. (1991). Labor and deliver characteristics and the risk of germinal matrix hemorrhage in low birth weight infants. *Journal of child neurology*, 6 (1), 35-40.

Tratados en el libro de

Pagano, M., Gauvreau, K. (2018). *Principles of biostatistics*. Chapman and Hall/CRC.

(o en su versión anterior del año 2000).

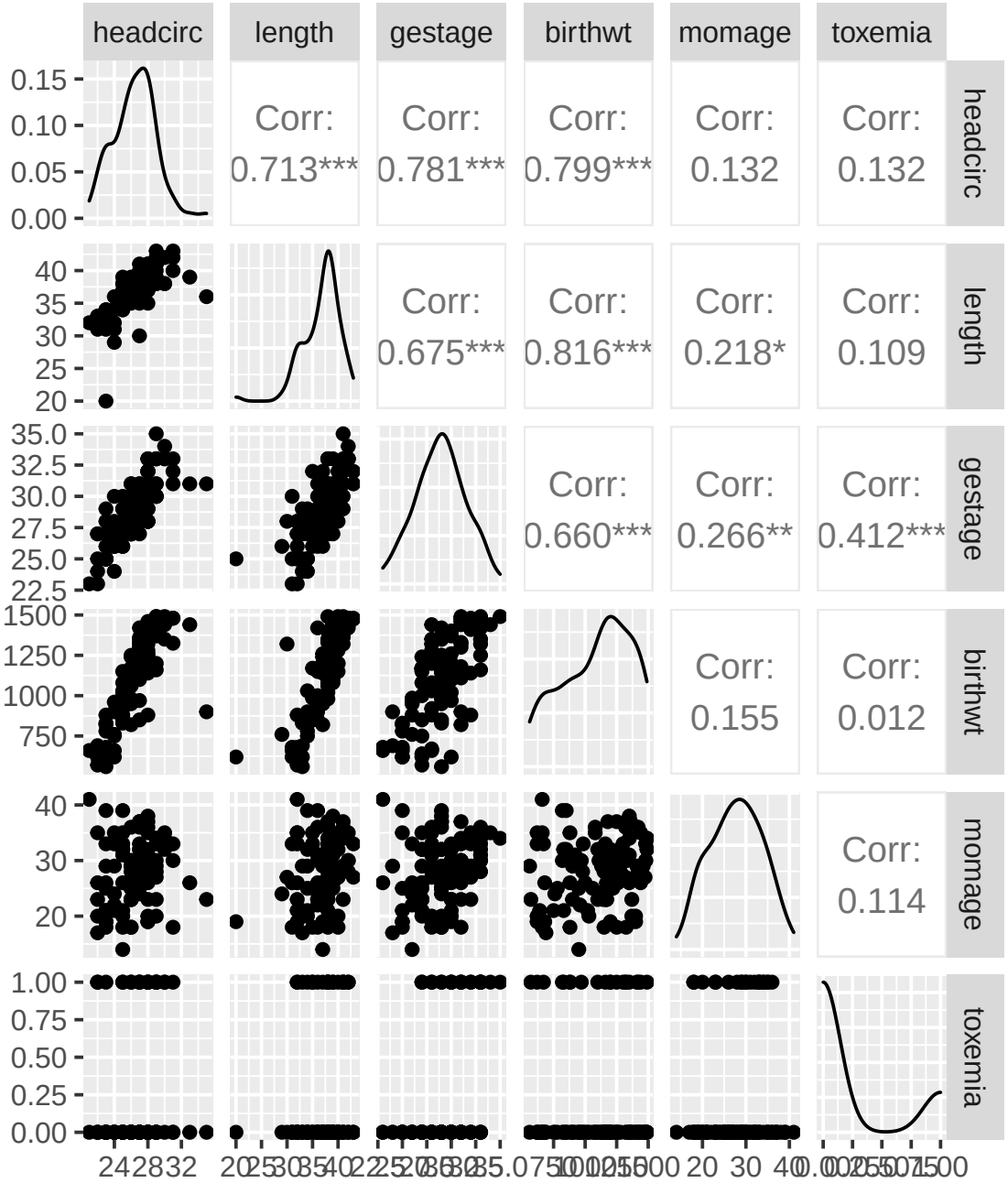
# Ejemplo: *low birth weight data*

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

- $Y =$ `headcirc`: el perímetro cefálico al nacer (medido en cm.)

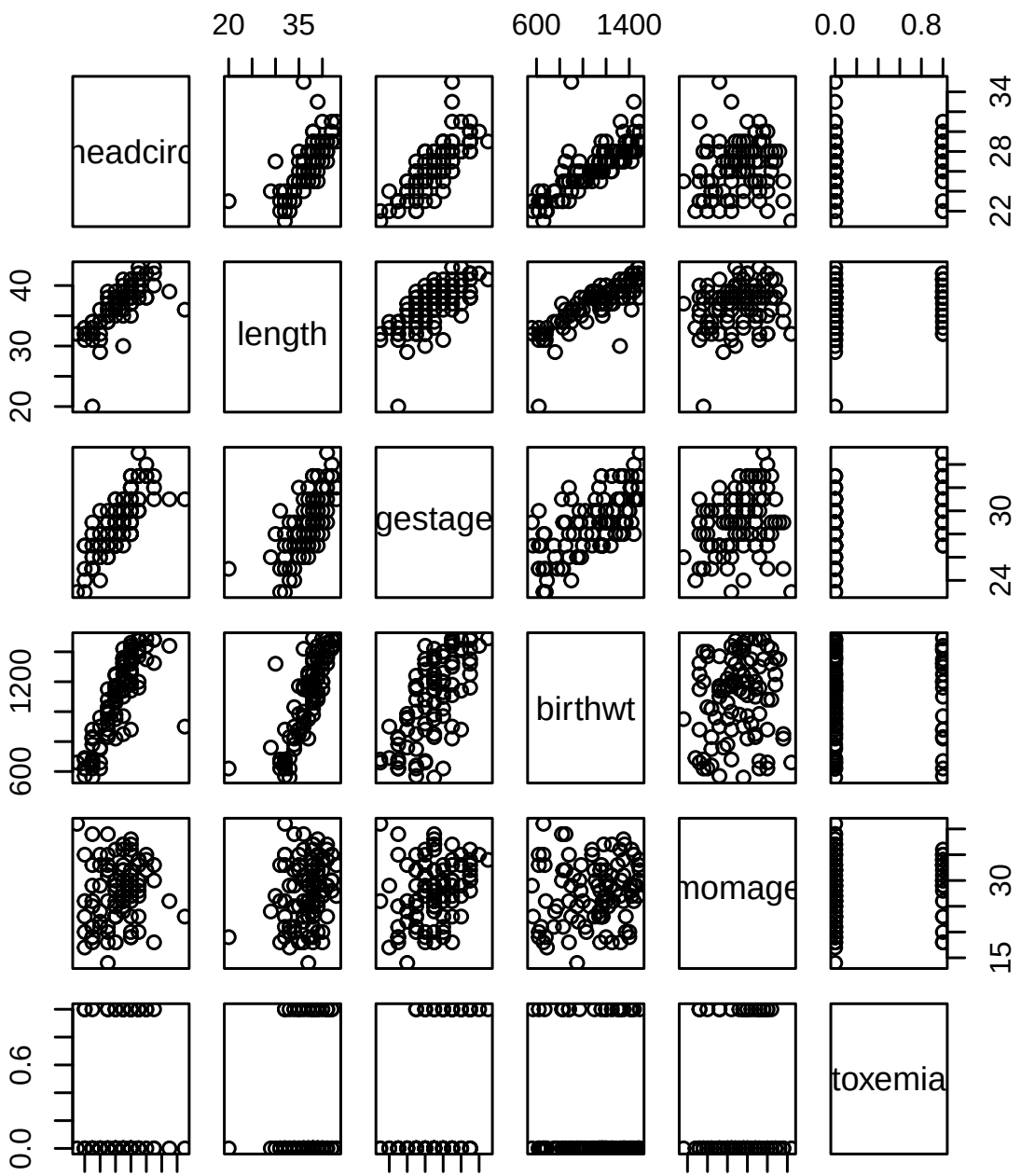
La base tiene varias covariables:

- $X_1 =$ `length`: longitud del bebé al nacer, en cm.
- $X_2 =$ `gestage`: edad gestacional o duración del embarazo
- $X_3 =$ `birthwt`: peso del bebé al nacer, en gramos
- $X_4 =$ `momage`: edad de la madre al nacimiento, en años
- $X_5 =$ `toxemia`: indicadora de que la madre padeció una patología durante el embarazo




```
> library(GGally)
> ggpairs(low)
```

```
>pairs(low)
```


headcirc
length
gestage
birthwt
momage
toxemia
20
35
600
1400
0.0
0.8
22
28
34
20
30
40
24
30
600
1200
15
30
0.0
0.6

Cuadro 1: Primeros 8 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 |

## Ejemplo 2.1 (Ajuste *low data* con dos covariables: gestage y birthwt)

Proponemos el modelo lineal múltiple con $p - 1 = 2$ covariables gestage y birthwt, es decir, $p = 3$ ($p$ es la cantidad de coeficientes $\beta$'s)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \qquad (18)$$

$$\texttt{headcirc}_i = \beta_0 + \beta_1 \cdot \texttt{gestage}_i + \beta_2 \cdot \texttt{birthwt}_i + \varepsilon_i$$

```
> ajuste2<-lm(headcirc~gestage+birthwt)
> summary(ajuste2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.3080154  1.5789429   5.262 8.54e-07 ***
gestage     0.4487328  0.0672460   6.673 1.56e-09 ***
birthwt     0.0047123  0.0006312   7.466 3.60e-11 ***
---
Residual standard error: 1.274 on 97 degrees of freedom
Multiple R-squared:  0.752,Adjusted R-squared:  0.7469
F-statistic: 147.1 on 2 and 97 DF,  p-value: < 2.2e-16
```

# Modelo ajustado

## Ejemplo 2.1

```
> coefficients(ajuste2)
(Intercept)     gestage     birthwt
8.308015388 0.448732848 0.004712283
```

Modelo (18) ajustado

$$\widehat{\texttt{headcirc}} = 8.308 + 0.449 \cdot \texttt{gestage} + 0.0047 \cdot \texttt{birthwt}$$

$$\widehat{Y} = 8.308 + 0.449 \cdot X_1 + 0.0047 \cdot X_2$$

# Interpretación de los parámetros estimados

## Ejemplo 2.1

La ordenada al origen, que es 8.3080 es, en teoría, el valor medio del perímetro cefálico para bebés de bajo peso con edad gestacional de 0 semanas y peso al nacer de 0 gramos, y por lo tanto carece de sentido.

El coeficiente estimado de edad gestacional (0.4487) no es el mismo que cuando la edad gestacional era la única variable explicativa en el modelo; su valor descendió de 0.7801 a 0.4487.

Esto implica que, si mantenemos el peso al nacer de un niño constante, cada incremento de una semana en la edad gestacional corresponde a un aumento de 0.4487 centímetros en su perímetro cefálico, en promedio. Una manera equivalente de decirlo es que dados dos bebés con el mismo peso al nacer pero tales que la edad gestacional del segundo de ellos es una semana más grande que la del primero, el perímetro cefálico esperado para el segundo bebé será 0.4487 centímetros mayor que el primero.

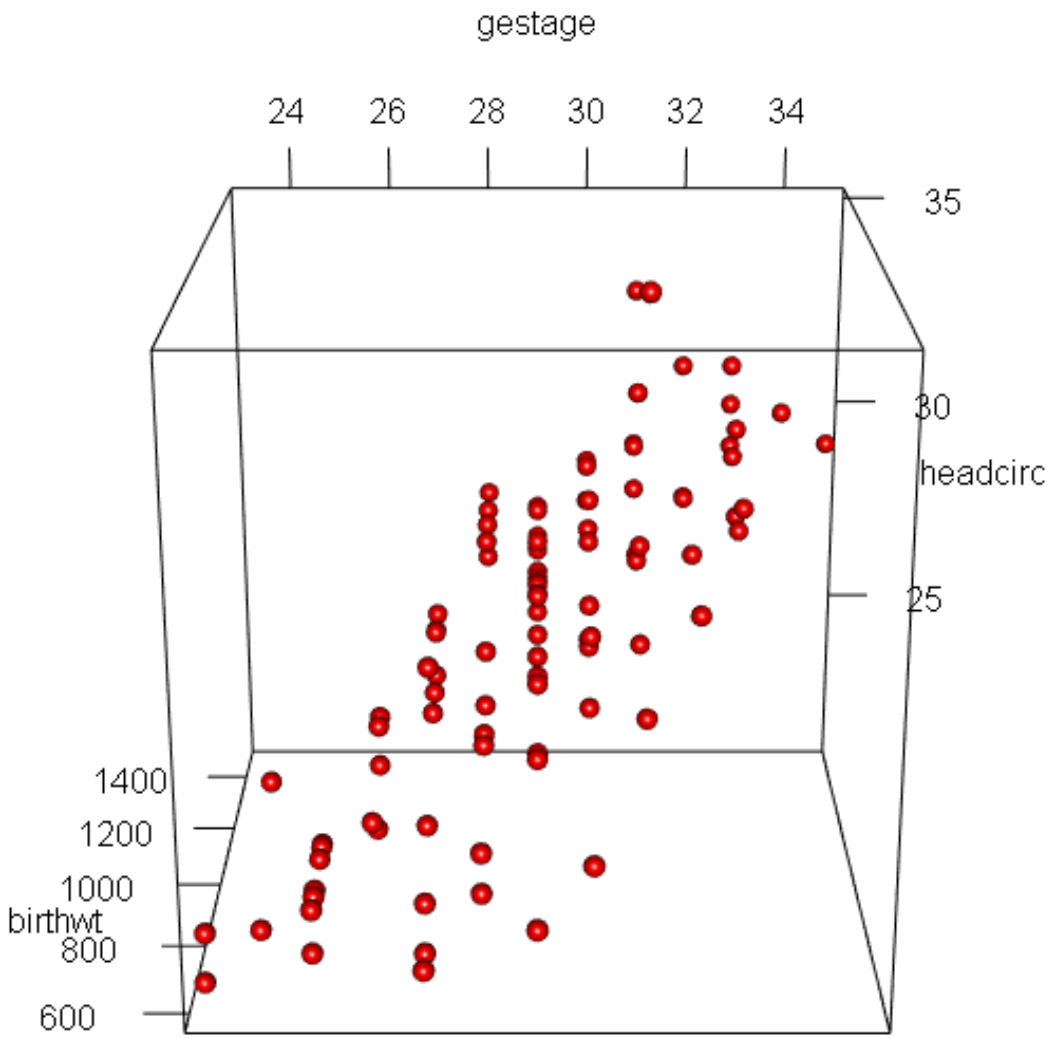
# Interpretación de los parámetros estimados

## Ejemplo 2.1

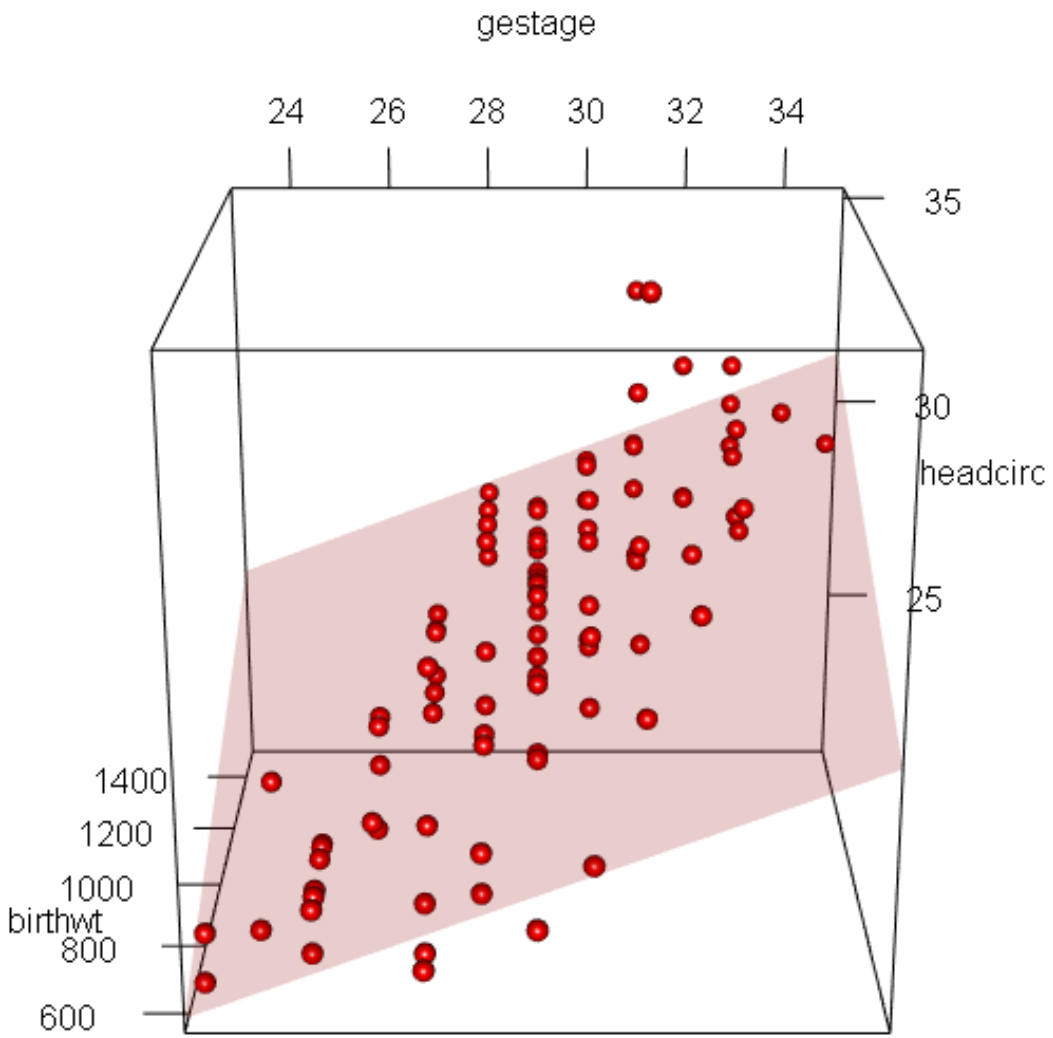
El coeficiente del peso al nacer indica que si la edad gestacional de un bebé no cambia, cada incremento de un gramo en el peso al nacer redunda en un aumento de 0.0047 centímetros en el perímetro cefálico, en promedio. En este caso en el que el valor del coeficiente estimado es tan pequeño, puede tener más sentido expresar el resultado aumentando las unidades involucradas, por ejemplo decir: si la edad gestacional no cambia, cada incremento de 10 g. en el peso al nacer redunda en un aumento de 0.047 cm. en el perímetro cefálico, en promedio.

Grafiquemos los datos y el ajuste, usando el paquete `rgl`.

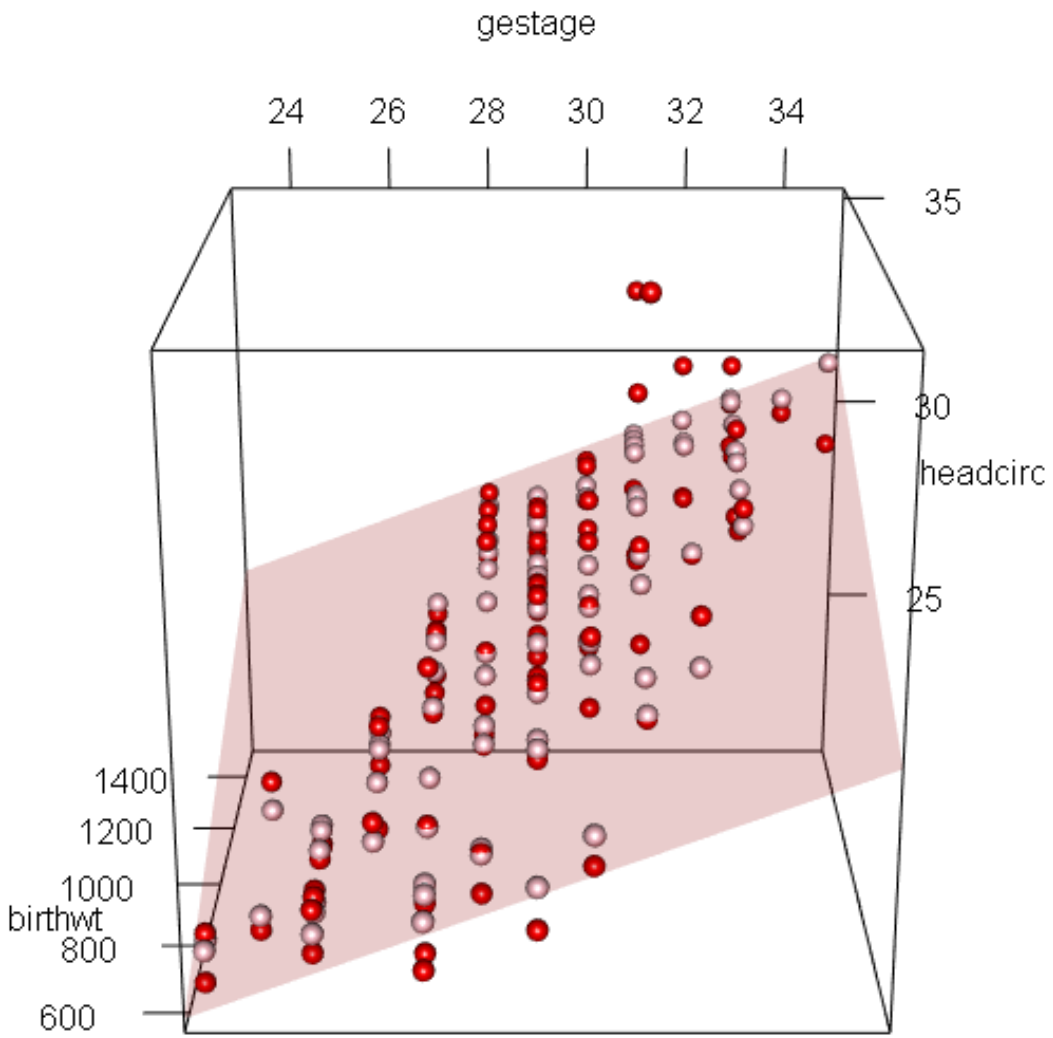
# Observaciones



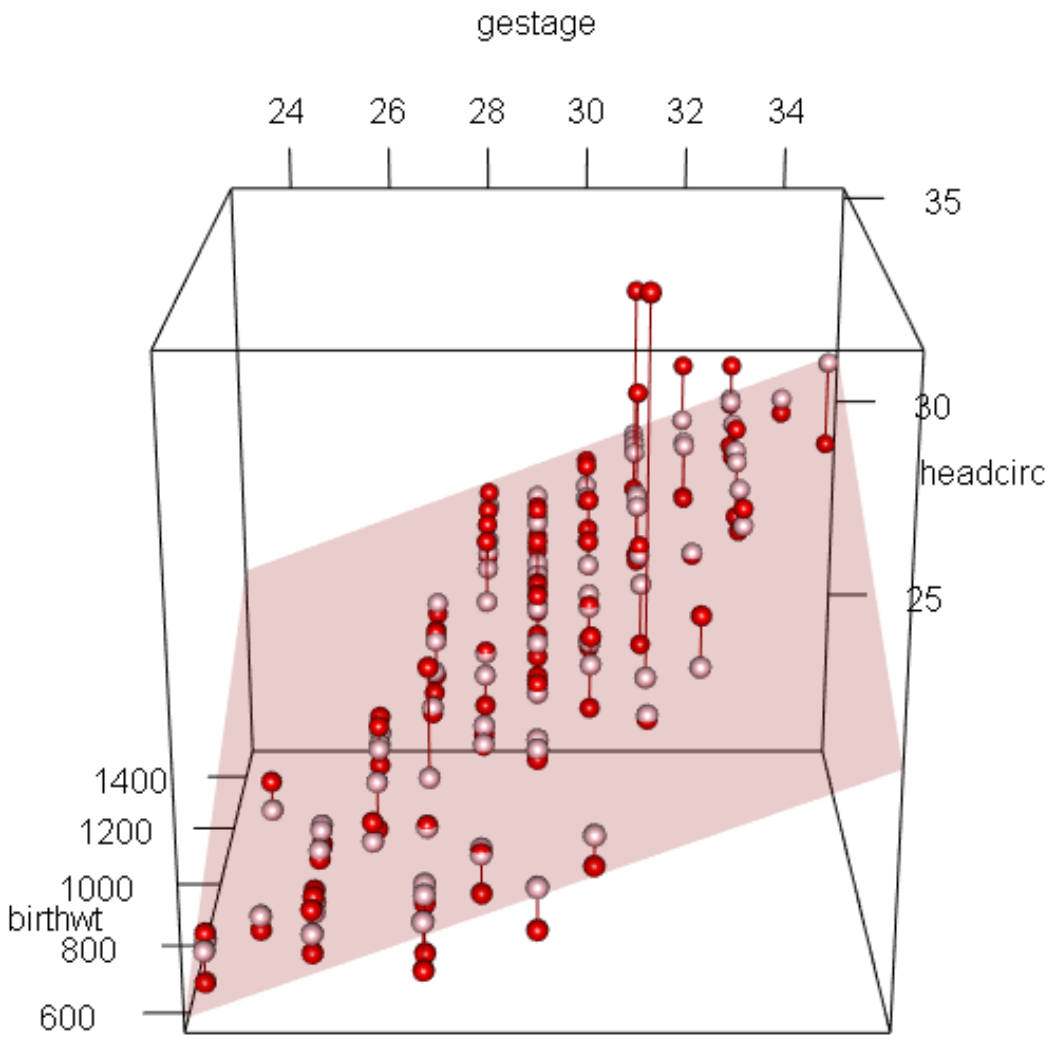
# Observaciones + plano ajustado



# Observaciones + plano ajustado + predichos (en rosa)



# Observaciones + plano ajustado + predichos + residuos



# ¿Son significativos los coeficientes del modelo?

Ejemplo 2.1

Queremos testear si tiene sentido complicarse con un modelo de regresión lineal múltiple para explicar a `headcirc`, o tenemos resultados parecidos con el modelo lineal simple al que habíamos llegado antes, que sólo utiliza a la edad gestacional como explicativa. ¿Qué test podemos hacer para responder a esta pregunta?

En el modelo (18)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i,$$

$$\texttt{headcirc}_i = \beta_0 + \beta_1 \cdot \texttt{gestage}_i + \beta_2 \cdot \texttt{birthwt}_i + \varepsilon_i$$

Usualmente se resume diciendo que queremos testear si la variable `birthwt` *es significativa* para explicar al perímetro cefálico cuando en el modelo está incluida la variable `gestage`. Las hipótesis a testear son

$$H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 \neq 0$$

# ¿Son significativos los coeficientes del modelo?

Las hipótesis a testear en el modelo (18) son

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

Para probar la hipótesis nula (o para no descartarla), necesitamos determinar si $\widehat{\beta}_2$, nuestra estimación de $\beta_2$, está lo suficientemente lejos de cero como para que podamos tener la certeza de que $\beta_2$ no es cero. ¿Qué tan lejos es lo suficientemente lejos? Por supuesto, esto depende de la precisión de $\widehat{\beta}_2$, es decir, depende de la raíz cuadrada de la varianza de $\widehat{\beta}_2$, que no conocemos, pero podemos estimar por $\widehat{sd}\left(\widehat{\beta}_2\right)$. ¿Cuánto es *chiquito o grande* en este contexto? Esto lo controla la distribución $t$ de Student con $n - p$ grados de libertad, ya que, si $H_0$ es verdadera (y bajo los supuestos del modelo lineal) tenemos que

$$T = \frac{\widehat{\beta}_2}{\widehat{sd}\left(\widehat{\beta}_2\right)} \sim t_{n-p}, \text{ bajo } H_0.$$

Miremos la salida del R

```
> summary(ajuste2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.3080154  1.5789429   5.262 8.54e-07 ***
gestage     0.4487328  0.0672460   6.673 1.56e-09 ***
birthwt     0.0047123  0.0006312   7.466 3.60e-11 ***
```

El estadístico del test es $T = \frac{\widehat{\beta}_2}{\widehat{sd}(\widehat{\beta}_2)} \sim t_{n-p}$, bajo $H_0$. Su valor observado es $T_{obs} = \frac{0.0047123}{0.0006312} = 7.466$. Los grados de libertad son $n - p = 100 - 3 = 97$. El test de nivel $\alpha = 0.05$ rechaza $H_0$ cuando $|T| \geq t_{n-p}\left(\frac{\alpha}{2}\right)$ donde $t_{n-p}\left(\frac{\alpha}{2}\right) =$ `qt(0.975,df = 97)` $= 1.9847$ es el percentil $1 - \frac{\alpha}{2} = 0.975$ de una t de Student con $n - p$ grados de libertad.

# ¿Son significativos los coeficientes del modelo?

¿Rechazamos o no rechazamos? Como $T_{obs} = 7.466 > t_{97}\,(0.025)$ entonces rechazo $H_0$, y concluyo que la variable `birthwt` es significativa para predecir al perímetro cefálico cuando la edad gestacional figura en el modelo. El p-valor del test es

$$\begin{aligned} \text{p-valor} &= 2P(T \geq |T_{obs}|) = P(|T| \geq |T_{obs}|) \\ &= 2P(T \geq 7.466) = 3.6 \cdot 10^{-11} < 0.05 \end{aligned}$$

Repitiendo el razonamiento, la variable edad gestacional también es significativa a nivel 0.05 en el modelo para explicar al perímetro cefálico en el modelo lineal múltiple que también incluye al peso al nacer, pues el p-valor del test es $1.56 \cdot 10^{-9} < 0.05$.

¿Cuánto vale el estimador de $\sigma$? $S = 1.274$ (en el modelo simple con edad gestacional como explicativa teníamos $S = 1.59$ Desde este punto de vista, también mejoramos.

¿Podemos incluir otras covariables?

## Ejemplo 2.2 (Ajuste *low data* con `length`)

Entre las variables medidas se encuentra `length`: la longitud del bebé al nacer, en cm. Para practicar mirar la significatividad en el modelo lineal simple proponemos el modelo

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i, \qquad (19)$$

$$\texttt{headcirc}_i = \beta_0 + \beta_1 \cdot \texttt{length}_i + \varepsilon_i$$

```
> ajuste3<-lm(headcirc ~ length)
> summary(ajuste3)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.84406    1.85829   4.221 5.44e-05 ***
length       0.50532    0.05024  10.059  < 2e-16 ***
---
Residual standard error: 1.785 on 98 degrees of freedom
Multiple R-squared:  0.508,Adjusted R-squared:  0.503
F-statistic: 101.2 on 1 and 98 DF,  p-value: < 2.2e-16
```

Si testeamos en el modelo (19) las hipótesis

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

vemos que el p-valor es menor a $2 \cdot 10^{-16}$ indicando que hay asociación lineal entre `length` y `headcirc`.

El valor de $S$ que surge como estimación del desvío estándar del error es $S = 1.785$ indicando que es, por lo menos desde este punto de vista, mejor el ajuste del modelo lineal simple con `gestage` como explicativa.

Nos quedan covariables que aún no miramos.

## Ejemplo 2.3 (Ajuste *low data* con `toxemia`)

Medimos la variable `toxemia`: indicadora de que la madre padeció una patología durante el embarazo, es una indicadora, toma los valores 0 y 1. ¿Qué significa el modelo lineal que tiene solamente a esta como covariable?

$$Y_i = \beta_0 + \beta_5 X_{i5} + \varepsilon_i, \tag{20}$$

$$\texttt{headcirc}_i = \beta_0 + \beta_5 \cdot \texttt{toxemia}_i + \varepsilon_i$$

O bien,

$$E\left[Y \mid \texttt{toxemia} = x\right] = \beta_0 + \beta_5 \cdot x \tag{21}$$

Entonces

$$E\left[Y \mid \texttt{toxemia} = 0\right] = \beta_0 \tag{22}$$

$$E\left[Y \mid \texttt{toxemia} = 1\right] = \beta_0 + \beta_5 \tag{23}$$

Entonces los parámetros tienen interpretación propia: $\beta_0$ es la esperanza del perímetro cefálico en la población de las madres sin `toxemia` y $\beta_5$ es el incremento (o disminución) esperada en el perímetro cefálico al comparar las madres que tuvieron la patología con las que no la tuvieron.

# Ejemplo 2.3: Ajuste *low birth weight data* con `toxemia`

¿Y qué representa, para el modelo (20) el test de las hipótesis que siguen?

$$H_0 : \beta_5 = 0$$
$$H_1 : \beta_5 \neq 0$$

$$H_0 : \beta_5 = 0 \Leftrightarrow E\left[Y \mid \texttt{toxemia} = 0\right] = E\left[Y \mid \texttt{toxemia} = 1\right]$$

O sea, que es un test de igualdad de medias de dos poblaciones diferentes, versus que sean distintas, basada en muestras independientes. ¿Con qué distribución?

$$\texttt{Y}_i = \beta_0 + \beta_1 \cdot \texttt{toxemia}_i + \varepsilon_i$$

Como $\varepsilon_i \sim N(0, \sigma^2)$, tenemos que

- si `toxemia=0` entonces $Y_i \sim N(\beta_0, \sigma^2)$
- si `toxemia=1` entonces $Y_i \sim N(\beta_0 + \beta_1, \sigma^2)$

Ya conocemos un test para comparar las medias de dos poblaciones basado en muestras independientes para este caso.

# Test para comparar dos medias bajo normalidad

Sean $W_1, \ldots, W_{n_1}$ variables aleatorias independientes idénticamente distribuidas con $E(W_i) = \mu_0$ e independientes de $Z_1, \ldots, Z_{n_2}$ que a su vez son variables aleatorias independientes entre sí e idénticamente distribuidas con $E(Z_i) = \mu_1$. El test $t$ permite decidir entre las hipótesis

$$H_0 : \mu_0 = \mu_1$$
$$H_1 : \mu_0 \neq \mu_1$$

Recordemos que el estadístico del test es

$$\frac{\left(\overline{W}_{n_1} - \overline{Z}_{n_2}\right)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \underset{\text{Bajo } H_0}{\sim} t_{n_1+n_2-2}$$

donde $n_1$ y $n_2$ son los tamaños de las muestras respectivas, y

$$S_p^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_{i=1}^{n_1}\left(W_i - \overline{W}_{n_1}\right)^2 + \sum_{j=1}^{n_2}\left(Z_j - \overline{Z}_{n_2}\right)^2\right]$$

es la varianza *pooleada* o combinada de ambas muestras. Por otra parte, para el modelo (20), el test de $H_0 : \beta_5 = 0$ es también un test $t$

# Test para testear si $\beta_5$ es cero

## Ejemplo 2.3

```
> ajuste4<-lm(headcirc ~ toxemia)
> summary(ajuste4)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.2785     0.2838  92.586   <2e-16 ***
toxemia       0.8168     0.6194   1.319     0.19
---

Residual standard error: 2.523 on 98 degrees of freedom
Multiple R-squared:  0.01744,Adjusted R-squared:  0.007409
F-statistic: 1.739 on 1 and 98 DF,  p-value: 0.1903
```

# Test para comparar dos medias bajo normalidad

## Ejemplo 2.3

```
> t.test(headcirc ~ toxemia,var.equal = TRUE)
Two Sample t-test

data:  headcirc by toxemia
t = -1.3187, df = 98, p-value = 0.1903
alternative hypothesis: true difference in means between
group 0 and group 1 is not equal to 0
95 percent confidence interval:
-2.0458641  0.4123499
sample estimates:
mean in group 0 mean in group 1
26.27848        27.09524
```

Observemos que tanto el estadístico calculado como el p-valor son los mismos. Ambos indican que la variable `toxemia` no es significativa para explicar al perímetro cefálico.

## Ejercicio 2.2

*Comprobar que ambos tests dan los mismos resultados. Para ello es un buen ejercicio calcular cuánto valen* $\widehat{\mathbf{Y}}$ *y RSS para el modelo lineal (20).*

Ahora proponemos el modelo con todas las covariables como explicativas (aun cuando conjeturamos que `toxemia` no resultará significativa).

## Ejemplo 2.4 (Ajuste *low data* con cinco covariables)

Proponemos el modelo con $p - 1 = 5$, es decir, $p = 6$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i, \tag{24}$$

```
> ajuste<-lm(headcirc~.,data = low)
> summary(ajuste)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.2097216  2.1285705   3.387  0.00103 **
length       0.0082711  0.0653434   0.127  0.89954
gestage      0.5261922  0.0835553   6.298 9.62e-09 ***
birthwt      0.0042555  0.0008867   4.799 5.99e-06 ***
momage      -0.0300651  0.0222312  -1.352  0.17950
toxemia     -0.5160581  0.3696445  -1.396  0.16597

Residual standard error: 1.269 on 94 degrees of freedom
Multiple R-squared:  0.7615,Adjusted R-squared:  0.7488
F-statistic: 60.03 on 5 and 94 DF,  p-value: < 2.2e-16
```

# Ejemplo: Ajuste *low birth weight data*: interpretación

Ejemplo 2.4

Entonces si queremos testear si la variable `length` es significativa para explicar al perímetro cefálico cuando en el modelo están incluidas las variables $X_2$, $X_3$, $X_4$, $X_5$ el test de

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

tiene p–valor igual a $0.8995 > 0.05$, por lo cual no hay evidencia suficiente en la muestra para rechazar $H_0$, y podemos pensar que no es necesaria esta variable para explicar a $Y$ cuando ya incluimos las otras 4.

También nos podría interesar hacer el test de si la edad de madre (`momage`) es significativa en el modelo (24), es decir cuando se incluyen las variables $X_1$, $X_2$, $X_3$, $X_5$. Para ese test el p–valor resulta ser $0.1795 > 0.05$ por lo cual, esta variable no resulta significativa en el modelo (24). Tampoco resulta significativa `toxemia`)(p–valor 0.165) cuando se incluyen las otras cuatro covariables en el modelo.

¿Tiene sentido que la regresión múltiple sugiera que no hay relación entre headcirc y length, mientras que la regresión lineal simple implica la opuesto? ¿Cómo lo entendemos? Consideremos la matriz de correlación muestral entre las variables predictoras

| | headcirc | length | gestage | birthwt | momage | toxemia |
|---|---|---|---|---|---|---|
| headcirc | 1.00 | 0.71 | 0.78 | 0.80 | 0.13 | 0.13 |
| length | 0.71 | 1.00 | 0.68 | 0.82 | 0.22 | 0.11 |
| gestage | 0.78 | 0.68 | 1.00 | 0.66 | 0.27 | 0.41 |
| birthwt | 0.80 | 0.82 | 0.66 | 1.00 | 0.15 | 0.01 |
| momage | 0.13 | 0.22 | 0.27 | 0.15 | 1.00 | 0.11 |
| toxemia | 0.13 | 0.11 | 0.41 | 0.01 | 0.11 | 1.00 |

Vemos que la correlación entre birthwt y length es de 0.82. Esto indica que los bebés largos tienden a tener también un mayor peso al nacer. Es decir, que ambas variables tienen información de alguna manera **repetida** sobre el bebé recién nacido. La relación entre el perímetro cefálico y el largo del chico se ve claramente en el modelo lineal simple. Pero cuando ponemos toda la información en el modelo múltiple, el hecho de que el peso del bebé resulte significativo y la longitud del bebé no, es porque el modelo asigna a birthwt ese efecto y la variable length no aporta nueva información a esa relación.

Lo mismo pasa con `gestage` y `length`, cuya correlación muestral resulta también alta, 0.68.

¿Por qué no resultan significativas `toxemia` ni `momage`? Vemos que tienen baja correlación muestral con la variable respuesta.

**Ejercicio 2.3**

*Compruebe, ajustando un modelo lineal simple, que `momage` no resulta significativa para explicar a `headcirc`.*

## Ejercicio 2.4

*Proponga un modelo lineal con dos variables explicativas:* `toxemia` *y* `gestage` *para explicar a* `headcirc`.

1. *Ajústelo con los datos. Evalúe la significatividad de los coeficientes. ¿Puede interpretar el modelo propuesto? Para hacerlo, escriba ecuaciones para la esperanza condicional para este caso, como las dadas en (21), (22) y (23), condicionando a*
   - `toxemia` *y* `gestage` *como en (21),*
   - `toxemia` $= 0$ *y* `gestage` *como en (22),*
   - `toxemia` $= 1$ *y* `gestage` *como en (23),*
2. *Grafique en 2D* `headcirc` *versus* `gestage`*, con dos colores distintos (azul y rojo) según* `toxemia`*. Superpóngale al gráfico los valores predichos (usando* `fitted.values`*), en colores celeste y rosa de acuerdo al valor de* `toxemia`*. Interprete.*

# Ejemplo: Ajuste *low data*, test de nivel simultáneo

Ejemplo 2.4

Ahora nos interesa hacer un test simultáneo de las hipótesis

$$H_0 : \beta_1 = 0, \beta_4 = 0, \beta_5 = 0$$
$$H_1 : \text{ alguno de los tres es } \neq 0$$

Esto no lo sabemos resolver aún. Necesitamos el siguiente resultado.

# Tests e intervalos para $q$ combinaciones lineales de $\boldsymbol{\beta}$

## Teorema 2.3

*Bajo* $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ *con* $\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}, \sigma^2 I_n\right)$, *rango*$(X) = p$. *Sea* $A \in \mathbb{R}^{q \times p}$ *de rango* $q$. *Entonces*

i. *El estadístico* $F = \dfrac{\left(A\widehat{\boldsymbol{\beta}} - A\boldsymbol{\beta}\right)^T \left[A\left(X^TX\right)^{-1}A^T\right]^{-1}\left(A\widehat{\boldsymbol{\beta}} - A\boldsymbol{\beta}\right)}{qS^2} \sim F_{q,n-p}$

ii. *Para testear* $H_0 : A\boldsymbol{\beta} = \mathbf{c} \in \mathbb{R}^q$ *versus* $H_1 : A\boldsymbol{\beta} \neq \mathbf{c}$ *puede usarse el test basado en*

$$F = \frac{\left(A\widehat{\boldsymbol{\beta}} - \mathbf{c}\right)^T \left[A\left(X^TX\right)^{-1}A^T\right]^{-1}\left(A\widehat{\boldsymbol{\beta}} - \mathbf{c}\right)}{qS^2} \sim F_{q,n-p}, \text{ bajo } H_0$$

*con nivel simultáneo* $\alpha$, *que rechaza la hipótesis nula cuando* $F_{obs} \geq F_{q,n-p}\left(1-\alpha\right)$, *siendo* $F_{q,n-p}\left(1-\alpha\right)$ *el percentil* $1-\alpha$ *de una distribución* $F_{q,n-p}$.

iii. *Una región de confianza de nivel simultáneo* $1-\alpha$ *para* $A\boldsymbol{\beta}$ *estará dada por* $F \leq F_{q,n-p}\left(1-\alpha\right)$, *donde* $F$ *está definido en (i.)*

# Estimador de $\beta$ con restricciones

## Teorema 2.4

*Bajo* $\mathbf{Y} = X\beta + \varepsilon$ *con* $\varepsilon \sim N_n\left(\mathbf{0}, \sigma^2 I_n\right)$, *rango*$(X) = p$. *Sea* $A \in \mathbb{R}^{q\times p}$ *de rango* $q$. *Queremos testear* $H_0 : A\beta = \mathbf{c} \in \mathbb{R}^q$

i. *El estimador de mínimos cuadrados de* $\beta$ *sujeto a las restricciones dadas por* $H_0$, *que llamaremos* $\widehat{\beta}_{H_0}$ *está dado por*

$$\widehat{\beta}_{H_0} = \left(X^T X\right)^{-1} A^T \left[A\left(X^T X\right)^{-1} A^T\right]^{-1} \left(\mathbf{c} - A\widehat{\beta}\right) + \widehat{\beta}$$

ii. *El mínimo valor de la suma de cuadrados es*
   - *Sin restricciones,* $S\left(\widehat{\beta}\right) = RSS = \left\|\mathbf{Y} - \widehat{\mathbf{Y}}\right\|^2 = \left\|\mathbf{Y} - X\widehat{\beta}\right\|^2$
   - *Con la restricción* $A\beta = \mathbf{c}$, *llamemos* $\widehat{\mathbf{Y}}_{H_0} = X\widehat{\beta}_{H_0}$ *entonces,* $S\left(\widehat{\beta}_{H_0}\right) = RSS_{H_0} = \left\|\mathbf{Y} - \widehat{\mathbf{Y}}_{H_0}\right\|^2 = \left\|\mathbf{Y} - X\widehat{\beta}_{H_0}\right\|^2$

iii. $\left\|\mathbf{Y} - \widehat{\mathbf{Y}}_{H_0}\right\|^2 = \left\|\mathbf{Y} - \widehat{\mathbf{Y}}\right\|^2 + \left\|\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{H_0}\right\|^2$

# Estimador de mínimos cuadrados con restricciones

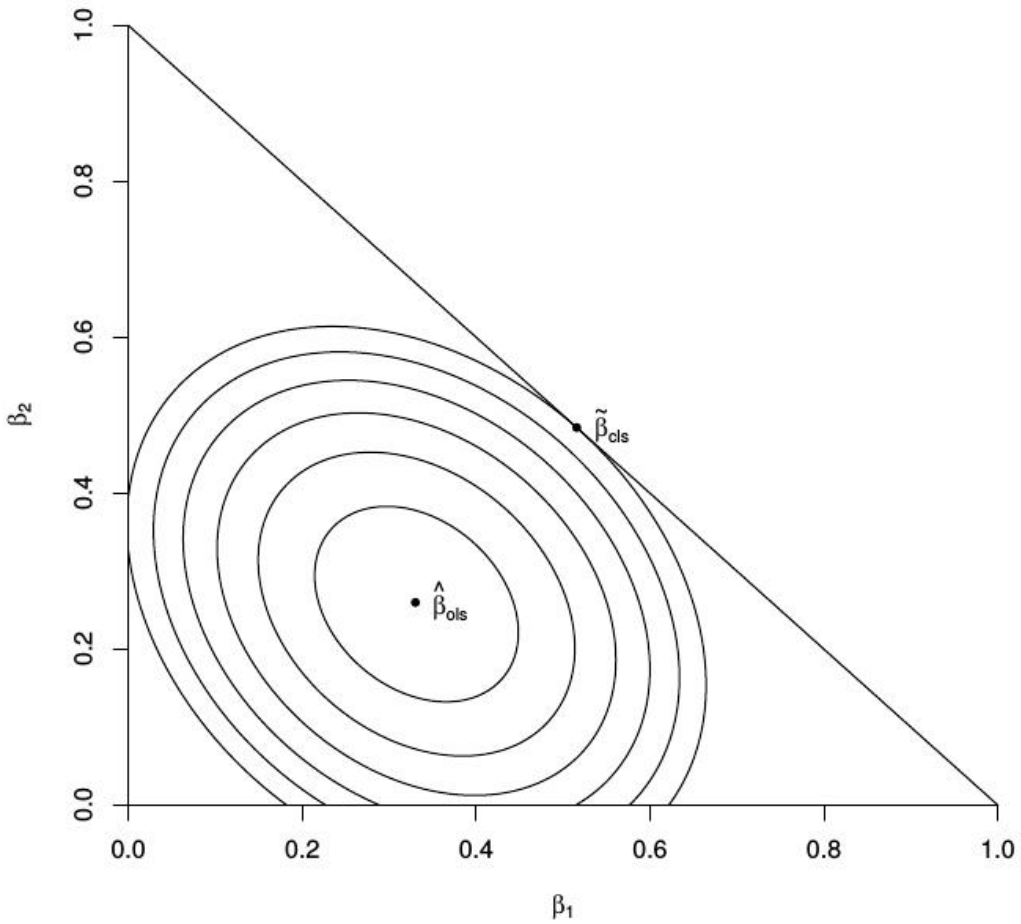



Figure 8.1: Imposing a Constraint on the Least Squares Criterion

Fuente: Econometrics. Bruce E.

Hansen University of Wisconsin. Copyright 2000, 2019. https://www.ssc.wisc.edu/ bhansen/econometrics/

# Test F y proyecciones

## Teorema 2.5

*Bajo* $\mathbf{Y} = X\beta + \varepsilon$ *con* $\varepsilon \sim N_n\left(\mathbf{0}, \sigma^2 I_n\right)$, *rango*$(X) = p$. *Sea* $A \in \mathbb{R}^{q \times p}$ *de rango* $q$. *Sea* $H : A\beta = \mathbf{c} \in \mathbb{R}^q$.

i.

$$RSS_{H_0} - RSS = \left\| \widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{H_0} \right\|^2 = \left(A\widehat{\beta} - \mathbf{c}\right)^T \left[A\left(X^T X\right)^{-1} A^T\right]^{-1} \left(A\widehat{\beta} - \mathbf{c}\right).$$

ii. *Bajo* $H$,

$$F = \frac{(RSS_{H_0} - RSS)/q}{RSS/(n-p)} = \frac{\left(A\widehat{\beta} - \mathbf{c}\right)^T \left[A\left(X^T X\right)^{-1} A^T\right]^{-1} \left(A\widehat{\beta} - \mathbf{c}\right)}{qS^2} \sim F_{q,n-p}$$

# Interpretación geomótrica del test F

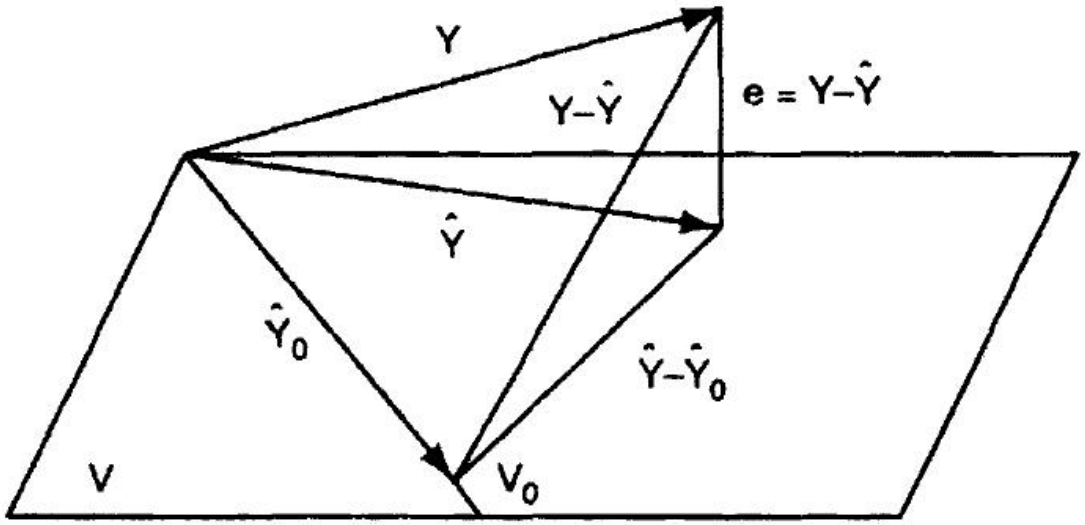



**FIGURE 3.10** Illustration for the $F$-test of $H_0$: $\theta \in V_0$.

Fuente: Stapleton, J. H. (2009).

Linear statistical models (Vol. 719). John Wiley & Sons.

# Ejemplo: Ajuste *low birth weight data*

Ajustamos el modelo con $p - 1 = 5$, es decir, $p = 6$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i, \qquad (25)$$

```
> ajuste<-lm(headcirc~.,data = low)
> summary(ajuste)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.2097216  2.1285705   3.387  0.00103 **
length       0.0082711  0.0653434   0.127  0.89954
gestage      0.5261922  0.0835553   6.298 9.62e-09 ***
birthwt      0.0042555  0.0008867   4.799 5.99e-06 ***
momage      -0.0300651  0.0222312  -1.352  0.17950
toxemia     -0.5160581  0.3696445  -1.396  0.16597

Residual standard error: 1.269 on 94 degrees of freedom
Multiple R-squared:  0.7615,Adjusted R-squared:  0.7488
F-statistic: 60.03 on 5 and 94 DF,  p-value: < 2.2e-16
```

# Ejemplo: Ajuste *low birth weight data*

Ahora nos interesa hacer un test simultáneo de las hipótesis

$$\begin{aligned} H_0 &: \beta_1 = 0, \beta_4 = 0, \beta_5 = 0 \\ H_1 &: \text{alguno de los tres es } \neq 0 \end{aligned}$$

entonces podemos hacer el test F que presentamos la clase pasada, Teorema 3.3 Para ello, sea

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \in \mathbb{R}^{3\times 6} = \mathbb{R}^{q\times p}$$

de rango $q = 3$, $p = 6$. Entonces, $H_0 : A\boldsymbol{\beta} = 0$. El tamaño de muestra es $n = 100$. Una manera de hacer esta cuenta en R es ajustar el modelo sin `length`, sin `momage`, sin `toxemia`, y pedirle a R que haga las cuentas de comparación de las sumas de los cuadrados de los residuos de ambos modelos, mediante el comando `anova`.

# Ejemplo: Ajuste *low birth weight data*

```
> ajuste <- lm(headcirc ~., data = low)
> ajuste2 <- lm(headcirc ~ gestage + birthwt)
> anova(ajuste2,ajuste)
Analysis of Variance Table
Model 1: headcirc ~ gestage + birthwt
Model 2: headcirc ~ length + gestage + birthwt
+ momage + toxemia
Res.Df    RSS   Df Sum of Sq   F   Pr(>F)
1     97 157.42
2     94 151.38  3    6.0452 1.2513 0.2957
```

En esta salida vemos que $RSS = 151.38$, $RSS_H = 157.42$ y $RSS_H - RSS = 157.42 - 151.38 = 6.0452$ Luego,

$$F = \frac{RSS_H - RSS}{RSS}\left(\frac{n-p}{q}\right) = \frac{6.0452}{151.38}\cdot\frac{94}{3} = 1.2513$$

# Ejemplo: Ajuste *low birth weight data*

Como el p–valor del test F es 0.2957, no rechazamos la hipótesis nula, y podemos quedarnos con el modelo lineal que tiene solamente a `gestage` y `birthwt` como explicativas.

```
> ajuste2<-lm(headcirc~gestage+birthwt)
> summary(ajuste2)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.3080154  1.5789429   5.262 8.54e-07 ***
gestage     0.4487328  0.0672460   6.673 1.56e-09 ***
birthwt     0.0047123  0.0006312   7.466 3.60e-11 ***

Residual standard error: 1.274 on 97 degrees of freedom
Multiple R-squared:  0.752,Adjusted R-squared:  0.7469
F-statistic: 147.1 on 2 and 97 DF,  p-value: < 2.2e-16
```

# Modelos con interacción entre variables

Como ya dijimos, cuando proponemos un modelo de regresión lineal múltiple del estilo de

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \tag{26}$$

estamos asumiendo que los efectos de las variables $X_1$ y $X_2$ sobre la respuesta $Y$ no interactúan entre sí: es decir, que el efecto de $X_1$ en $Y$ no depende del valor que tome $X_2$ (y al revés, cambiando $X_1$ por $X_2$, el efecto de $X_2$ en $Y$ no depende del valor que tome $X_1$). Cuando esto no sucede, es inadecuado proponer el modelo (26), y es necesario agregarle a dicho modelo un témino que intente dar cuenta de la *interacción* entre $X_1$ y $X_2$ en su relación con $Y$, es decir, del hecho de que el efecto de un predictor sobre la respuesta difiere de acuerdo al nivel de otro predictor. La manera estándar de hacerlo es agregarle al modelo (26) un término de interacción, es decir

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} \cdot X_{i2} + \varepsilon_i. \tag{27}$$

El modelo (27) es un caso particular del modelo de regresión lineal múltiple

Sea $X_{i3} = X_{i1} \cdot X_{i2}$ el producto entre las variables $X_1$ y $X_2$ medidas en el iésimo individuo, entonces el modelo (27) puede escribirse de la forma

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i,$$

que es un caso particular del modelo de regresión lineal múltiple. Algunas veces al coeficiente de la interacción se lo nota con los subíndices $1:2$, es decir $\beta_{1:2} = \beta_3$ para explicitar que es el coeficiente asociado a la interacción. Tiene interpretación distinta si las dos variables involucradas son:

- una categórica y una continua
- las dos continuas
- las dos categóricas

Comencemos por el caso de una categórica y una continua. Comencemos por un ejemplo.

# ANCOVA: Ejemplo **colesterol**

(Ejemplo simulado siguiendo a Faraway, J. *Linear Models with R*, 2009)

## Ejemplo 2.5 (Colesterol)

Estamos interesados en medir el efecto de un nuevo medicamento para bajar el nivel de colesterol en sangre. Nos interesa saber si este medicamento produce mejores resultados que el medicamento estándar. Tenemos dos grupos de pacientes:

- uno recibe la medicación nueva (grupo tratamiento)
- el otro recibe el tratamiento estándar (grupo control)

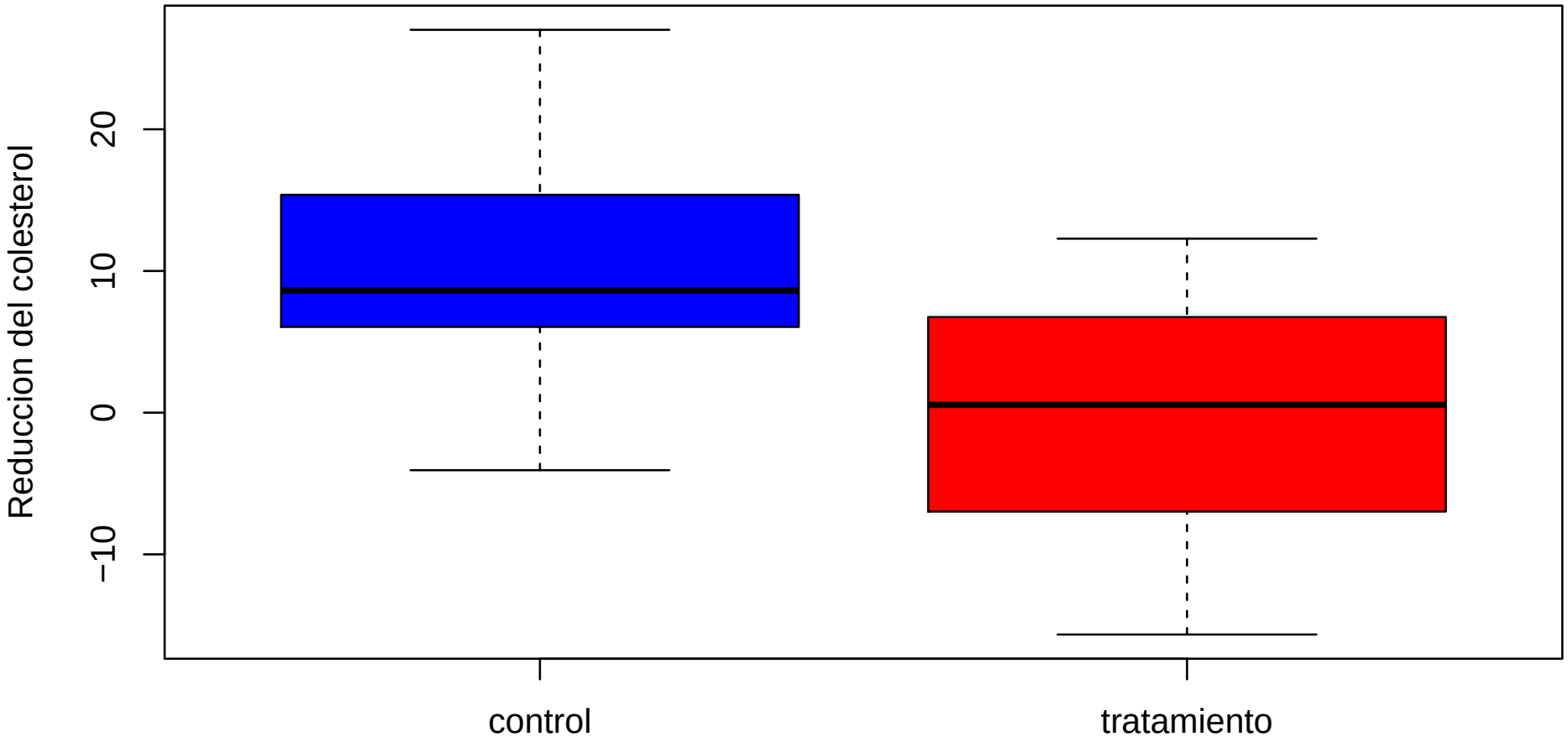
Para cada paciente se registra la edad y la variable respuesta: reducción porcentual en el nivel de colesterol luego del tratamiento, `reducccol`

`reducccol` = reducción porcentual de colesterol

Si supiéramos que los dos grupos difieren con respecto a la edad, no podríamos tratar esto como un simple problema de comparación de dos muestras ya que la edad afecta el nivel de colesterol. Veamos un ejemplo (simulado) en lo que sigue. Datos `colesterol.txt`. Script `ancovasimu2.R`

# Ejemplo **colesterol**: Boxplot de `reduccol` por grupos

20 pacientes en cada grupo



# ANCOVA: Ejemplo **colesterol**

Promedio (sd) de cada variable en la muestra de 20 pacientes por grupo

| | control | tratamiento |
|---|---|---|
| `reducccol` | 10.58 (7.75) | 0.01 (8.65) |
| `edad` | 39.50 | 59.80 |

Para los pacientes que recibieron la medicación, la reducción media en el nivel de colesterol fue del 0 %, mientras que para los que no lo hicieron, la reducción media fue del 10 %. Luego, concluiríamos que es mejor no ser tratado. Sin embargo, la edad promedio de los pacientes del grupo tratamiento fue de 60 años, mientras que la edad promedio de los pacientes del grupo control fue de 40 años. Queremos hacer una comparación entre tratamientos *que tenga en cuenta la edad*. O, que *controle por edad*.

# ANCOVA: Ejemplo **colesterol**

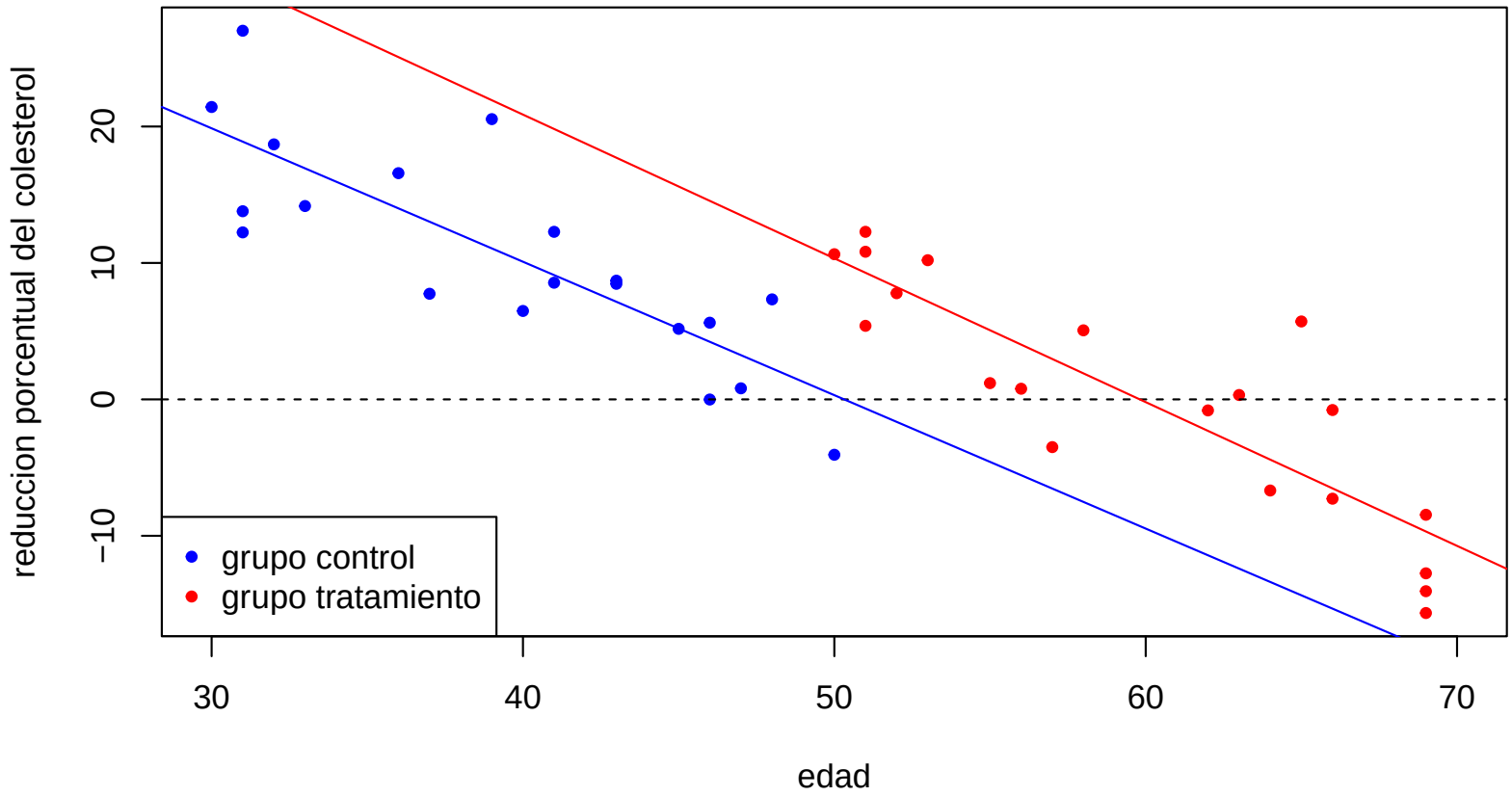
Comparamos las medias de la respuesta para ambos grupos **sin** tener en cuenta la edad.

```
> t.test(reducccol[trat==1],reducccol[trat==0],var.equal = T)
Two Sample t-test
data:  reducccol[trat == 1] and reducccol[trat == 0]
t = -4.0679, df = 38, p-value = 0.0002307
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
-15.82639  -5.30861
sample estimates:
mean of x mean of y
0.0095   10.5770
```

Grafiquemos la respuesta por edad, en colores distintos segun el grupo al que pertenece el paciente. Superponemos la recta de ajuste por mínimos cuadrados que se obtiene con los 20 pacientes de cada grupo.

# ANCOVA: Ejemplo **colesterol**

Figura 2: Datos de reducción de colesterol versus edad, con rectas estimadas por dos modelos de regresión lineal simple, cada uno ajustado a 20 datos



# Dos ajustes simples

Grupo control

```
> summary(lm(reducccol[tratamiento==0]~ edad[tratamiento==0]))
Estimate Std. Error t value Pr(>|t|)
(Intercept)              49.2062     6.3794   7.713 4.11e-07 ***
edad[tratamiento == 0]  -0.9780     0.1595  -6.133 8.59e-06 ***
---
Residual standard error: 4.53 on 18 degrees of freedom
Multiple R-squared:  0.6763,Adjusted R-squared:  0.6584
F-statistic: 37.61 on 1 and 18 DF,  p-value: 8.591e-06
```

Grupo tratamiento

```
> summary(lm(reducccol[tratamiento==1]~ edad[tratamiento==1]))
Estimate Std. Error t value Pr(>|t|)
(Intercept)              63.0178     8.6726   7.266 9.38e-07 ***
edad[tratamiento == 1]  -1.0537     0.1441  -7.314 8.58e-07 ***
---
Residual standard error: 4.462 on 18 degrees of freedom
Multiple R-squared:  0.7482,Adjusted R-squared:  0.7342
F-statistic: 53.49 on 1 and 18 DF,  p-value: 8.578e-07
```

# ANCOVA: Ejemplo **colesterol**

Definimos una variable indicadora del tratamiento ($\texttt{trat} = 1$ si el paciente pertenece al grupo tratamiento, 0 sino). El modelo más general que proponemos es (con interacción)

$$Y_i = \beta_0 + \beta_1 \texttt{edad}_i + \beta_2 \texttt{trat}_i + \beta_{1:2} \left(\texttt{trat}_i \cdot \texttt{edad}_i\right) + \varepsilon_i, \tag{28}$$

Es decir, que para los pacientes del grupo control (i.e. $\texttt{trat}_i = 0$) tenemos

$$Y_i = \beta_0 + \beta_1 \texttt{edad}_i + \varepsilon_i,$$

y que para los pacientes tratados ($\texttt{trat}_i = 1$) tenemos

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \texttt{edad}_i + \beta_2 + \beta_{1:2} \texttt{edad}_i + \varepsilon_i \\ &= \beta_0 + \beta_2 + \left(\beta_1 + \beta_{1:2}\right) \texttt{edad}_i + \varepsilon_i \end{aligned}$$

¡rectas no necesariamente paralelas! ¿Cómo se relacionan las ordenadas al origen entre sí? O sea, 2 rectas libres (sin ninguna atadura entre ambos modelos).

# ANCOVA: Ejemplo **colesterol**

Escribimos todo en términos de esperanza condicional

$$E\left(Y_i \mid \texttt{edad}_i, \texttt{trat}_i\right) = \beta_0 + \beta_1 \texttt{edad}_i + \beta_2 \texttt{trat}_i + \beta_{1:2}\left(\texttt{trat}_i \cdot \texttt{edad}_i\right), \tag{29}$$

Es decir, que para los pacientes del grupo control (i.e. $\texttt{trat}_i = 0$) tenemos,

$$\begin{aligned} E\left(Y_i \mid \texttt{edad}_i, \texttt{trat}_i = 0\right) &= \beta_0 + \beta_1 \texttt{edad}_i + \beta_2 \cdot 0 + \beta_{1:2}\left(0 \cdot \texttt{edad}_i\right) \\ &= \beta_0 + \beta_1 \texttt{edad}_i \end{aligned}$$

y que para los pacientes tratados ($\texttt{trat}_i = 1$) tenemos

$$\begin{aligned} E\left(Y_i \mid \texttt{edad}_i, \texttt{trat}_i = 1\right) &= \beta_0 + \beta_1 \texttt{edad}_i + \beta_2 \cdot 1 + \beta_{1:2}\left(1 \cdot \texttt{edad}_i\right) \\ &= \beta_0 + \beta_2 + (\beta_1 + \beta_{1:2})\texttt{edad}_i \end{aligned}$$

¡rectas no necesariamente paralelas!

# Colesterol: Ajuste del modelo completo

```
> trat.f <- factor(tratamiento)
> modelo1 <- lm(reducccol ~ edad*trat.f)
> summary(modelo1)

Estimate Std. Error t value Pr(>|t|)
(Intercept)   49.2062     6.3317   7.771 3.28e-09 ***
edad          -0.9780     0.1583  -6.179 4.01e-07 ***
trat.f1       13.8116    10.7917   1.280    0.209
edad:trat.f1  -0.0757     0.2148  -0.352    0.727
---
Residual standard error: 4.496 on 36 degrees of freedom
Multiple R-squared:  0.8023,Adjusted R-squared:  0.7858
F-statistic:  48.7 on 3 and 36 DF,  p-value: 9.36e-13
```

El coeficiente de la interacción no es significativo. ¿Qué quiere decir?

Los coeficientes en el modelo no resultan ser todos significativos. De hecho, el test de

$$H_0 : \beta_{1:2} = 0 \text{ versus } H_0 : \beta_{1:2} \neq 0$$

asumiendo que el modelo contiene a la edad y a la indicadora de tratamiento, tiene por estadístico $t_{obs} = -0.352$ y $p-\text{valor} = 0.727$. Esto nos dice que esta muestra no provee evidencia suficiente de que la edad tenga **un efecto diferente en el colesterol dependiendo del tratamiento recibido**.

Como la interacción no es estadísticamente significativa, no la retendremos en el modelo de regresión.

# Colesterol: Ajuste del modelo completo

```
> interac <- (edad*tratamiento)
> cor(cbind(edad,tratamiento,interac))
edad       tratamiento   interac
edad        1.0000000   0.8366529 0.8908923
tratamiento 0.8366529   1.0000000 0.9868524
interac     0.8908923   0.9868524 1.0000000
```

La interacción y el tratamiento están muuuy correlacionadas (repiten información). Probemos ajustar un modelo sin interacción

| pac | edad | trat.f1 | inter | pac | edad | trat.f1 | inter |
|---|---|---|---|---|---|---|---|
| 1 | 33 | 0 | 0 | 21 | 69 | 1 | 69 |
| 2 | 40 | 0 | 0 | 22 | 52 | 1 | 52 |
| 3 | 32 | 0 | 0 | 23 | 66 | 1 | 66 |
| 4 | 31 | 0 | 0 | 24 | 69 | 1 | 69 |
| 5 | 48 | 0 | 0 | 25 | 64 | 1 | 64 |
| 6 | 45 | 0 | 0 | 26 | 58 | 1 | 58 |
| 7 | 30 | 0 | 0 | 27 | 51 | 1 | 51 |
| 8 | 31 | 0 | 0 | 28 | 51 | 1 | 51 |
| 9 | 50 | 0 | 0 | 29 | 56 | 1 | 56 |
| 10 | 43 | 0 | 0 | 30 | 66 | 1 | 66 |
| 11 | 47 | 0 | 0 | 31 | 50 | 1 | 50 |
| 12 | 41 | 0 | 0 | 32 | 63 | 1 | 63 |
| 13 | 37 | 0 | 0 | 33 | 69 | 1 | 69 |
| 14 | 36 | 0 | 0 | 34 | 53 | 1 | 53 |
| 15 | 31 | 0 | 0 | 35 | 65 | 1 | 65 |
| 16 | 41 | 0 | 0 | 36 | 62 | 1 | 62 |
| 17 | 46 | 0 | 0 | 37 | 51 | 1 | 51 |
| 18 | 39 | 0 | 0 | 38 | 69 | 1 | 69 |
| 19 | 43 | 0 | 0 | 39 | 55 | 1 | 55 |
| 20 | 46 | 0 | 0 | 40 | 57 | 1 | 57 |

# Colesterol: modelo aditivo

El modelo que proponemos es

$$Y_i = \beta_0 + \beta_1 \texttt{edad}_i + \beta_2 \texttt{trat}_i + \varepsilon_i, \tag{30}$$

Es decir, que para los pacientes del grupo control (i.e. $\texttt{trat}_i = 0$) tenemos

$$Y_i = \beta_0 + \beta_1 \texttt{edad}_i + \varepsilon_i,$$

y que para los pacientes tratados ($\texttt{trat}_i = 1$) tenemos

$$Y_i = \beta_0 + \beta_2 + \beta_1 \texttt{edad}_i + \varepsilon_i$$

¡rectas paralelas! ¿Cómo son sus ordenadas al origen? ¿Cuál es la interpretación del coeficiente $\beta_2$?

# **Colesterol**: ajuste del modelo aditivo

```
> modelo2 <- lm(reducccol ~ edad+trat.f)
> summary(modelo2)
Estimate Std. Error t value Pr(>|t|)
(Intercept)  50.8301     4.2919  11.843 3.76e-14 ***
edad         -1.0191     0.1057  -9.641 1.23e-11 ***
trat.f1      10.1195     2.5648   3.946 0.000342 ***
---
Residual standard error: 4.442 on 37 degrees of freedom
Multiple R-squared:  0.8016,Adjusted R-squared:  0.7909
F-statistic: 74.76 on 2 and 37 DF,  p-value: 1.007e-13
```

Todos los coeficientes son significativos. Vemos que una vez que se tiene en cuenta la edad, la diferencia entre tratamiento y control nuevamente es del 10 %, pero esta vez a favor del tratamiento.

# Colesterol: modelo aditivo

El modelo ajustado es

$$\widehat{Y}_i = 50.83 - 1.02 \cdot \texttt{edad}_i + 10.12 \cdot \texttt{trat}_i,$$

Es decir, que para los pacientes del grupo control (i.e. $\texttt{trat}_i = 0$) tenemos

$$\text{(controles)} \qquad \widehat{Y}_i = 50.83 - 1.02 \cdot \texttt{edad}_i$$

y que para los pacientes tratados ($\texttt{trat}_i = 1$) tenemos

$$\text{(tratados)} \qquad \widehat{Y}_i = 60.95 - 1.02 \cdot \texttt{edad}_i$$

# ¿Colesterol: por qué usar un modelo lineal?

El análisis de la covarianza, utiliza la relación entre la variable de respuesta (reducción en el nivel de colesterol, en nuestro ejemplo) y una o más variables cuantitativas para las cuales hay información disponible (edad, en nuestro ejemplo) para reducir la variabilidad del término del error y permitir que la comparación entre los grupos sea más poderosa. Se lo suele denominar *controlar por la variable edad*.

El interés está puesto en la comparación de la respuesta en ambos grupos, pero los grupos difieren en caracterósticas que pueden ser tenidas en cuenta en el modelo.

Esto sucedió en el ejemplo del colesterol, en el cual si nos quedamos con el análisis sin covariables (comparación de medias con el `t.test` y boxplot) concluimos equivocadamente que conviene quedarse con el tratamiento estándar, sin embargo, cuando incluimos a la edad como covariable ("controlamos por la edad") podemos llegar a la conclusión correcta: el nuevo tratamiento reduce más el colesterol que el estándar.

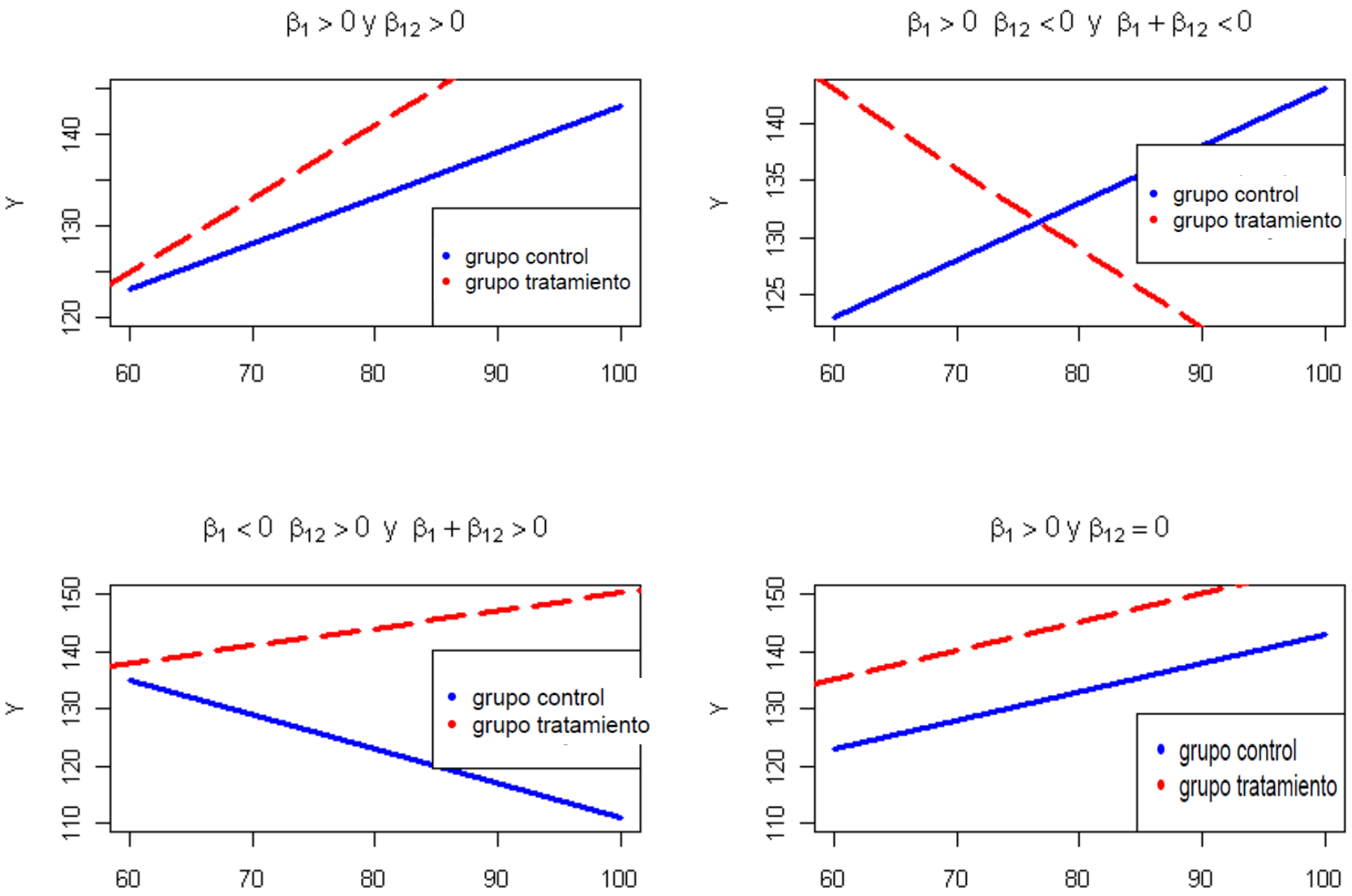
# Comparación con dos modelos lineales por separado

¿Qué diferencia hay con usar dos modelos lineales, uno para cada grupo de pacientes?

1. El modelo aditivo (30) asume pendientes iguales y la misma varianza del error de para cada observación. En consecuencia, la pendiente común $\beta_1$ se puede estimar mejor usando la información en la muestra conjunta. Ojo, este modelo no debería usarse si no se cree que este supuesto sea correcto para los datos a analizar.
2. Se puede testear si el modelo de rectas paralelas es correcto
3. Usando el modelo aditivo otras inferencias, como por ejemplo las realizadas sobre $\beta_0$ y $\beta_2$ resultarán más precisas pues se dispone de más observaciones para estimarlas y para estimar a $\sigma^2$ la varianza del error (lo que se traduce en más grados de libertad para estimarlos).

El coeficiente de la interacción ($\beta_{1:2}$) representa el aumento (o la disminución) de la pendiente en un grupo de observaciones (en este caso, los tratados) con respecto al otro (no tratados). Si $\beta_{1:2} = 0$ esto significaría que ambas rectas son paralelas. Los distintos valores que pueden tomar $\beta_1$ y $\beta_{1:2}$ dan lugar a distintos posibles tipos de interacción entre las variables, según se ve en la Figura 3.

Figura 3: Gráfico de posibles combinaciones de valores de $\beta_1$ y $\beta_{1:2}$ para el modelo con interacción (29).



# Interacción entre dos variables cuantitativas

Vimos que el modelo aditivo propone que cuando la covariable $X_j$ aumenta una unidad, la media de $Y$ aumenta en $\beta_j$ unidades **independientemente de cuáles sean los valores de las otras variables**. Esto implica paralelismo de las rectas que relacionan a $Y$ y $X_j$, cualesquiera sean los valores que toman las demás variables.

En nuestro ejemplo de los bebés de bajo peso, propusimos un modelo de regresión lineal múltiple con dos variables predictoras. Recordemos que habíamos definido

$$
\begin{aligned}
Y_i &= \text{ perímetro cefálico del iésimo niño, en centímetros (}\texttt{headcirc}\text{)} \\
X_{i1} &= \text{ edad gestacional del iésimo niño, en semanas (}\texttt{gestage}\text{)} \\
X_{i2} &= \text{ peso al nacer del iésimo niño, en gramos (}\texttt{birthwt}\text{)}
\end{aligned}
$$

Propusimos el siguiente modelo,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon. \tag{31}$$

La superficie ajustada resultó ser

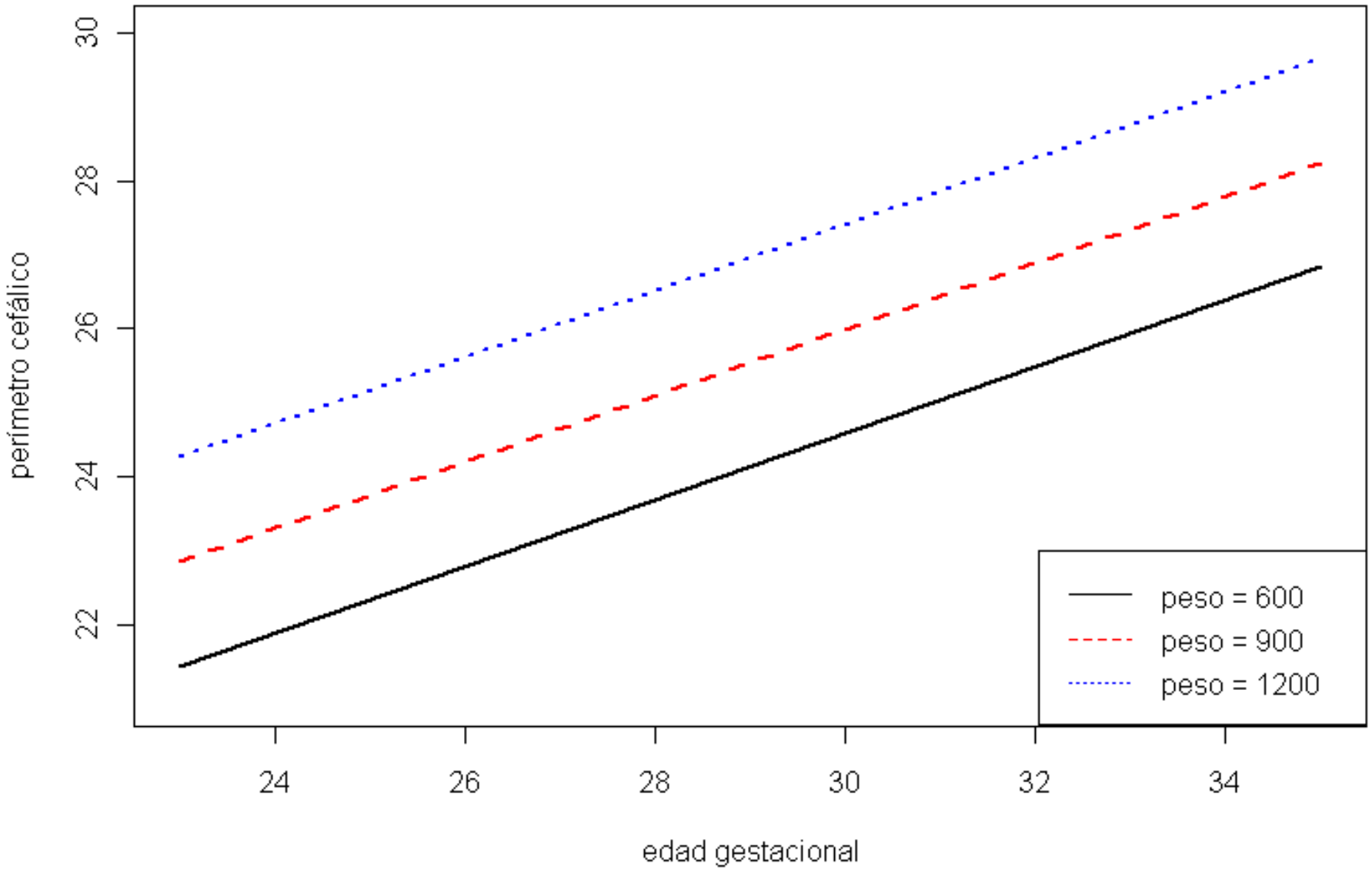
$$\widehat{Y} = 8.3080 + 0.4487 X_1 + 0.0047 X_2.$$

Cuando controlamos por $X_2$ (peso al nacer), la ecuación (parcial) ajustada que relaciona el perímetro cefálico y la edad gestacional es

$$\begin{aligned}
X_2 &= 600, \quad \widehat{Y} = 8.3080 + 0.4487 X_1 + 0.0047 \cdot 600 = 11.128 + 0.4487 X_1 \\
X_2 &= 900, \quad \widehat{Y} = 8.3080 + 0.4487 X_1 + 0.0047 \cdot 900 = 12.538 + 0.4487 X_1 \\
X_2 &= 1200, \quad \widehat{Y} = 8.3080 + 0.4487 X_1 + 0.0047 \cdot 1200 = 13.948 + 0.4487 X_1
\end{aligned}$$

Para cada nivel posible de peso al nacer, por cada unidad de aumento en la edad gestacional se espera un aumento de 0.448 unidades (cm.) en el perímetro cefálico al nacer. Gráficamente, esto se ve representado en la Figura 4. Lo mismo sucedería si controláramos por $X_1$ en vez de $X_2$ : tendríamos rectas paralelas, de pendiente 0.0047.

Figura 4: Perímetro cefálico esperado en función de la edad gestacional,



Este modelo fuerza a que los efectos de las covariables en la variable dependiente sean aditivos, es decir, el efecto de la edad gestacional será el mismo para todos los valores del peso al nacer, y viceversa, porque el modelo no le permitirá ser de ninguna otra forma.

Cuando esto no suceda, es decir, cuando pensemos que tal vez la forma en que el perímetro cefálico varíe con la edad gestacional dependa del peso al nacer del bebé, será necesario descartar (o validar) esta conjetura. Una manera de investigar esta posibilidad es incluir un **término de interacción en el modelo**. Para ello, creamos la variable artificial que resulta de hacer el producto de las otras dos: $X_3 = X_1 \cdot X_2 = \texttt{gestage} \cdot \texttt{birthwt}$, y proponemos el modelo

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \\ Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{1:2} X_1 \cdot X_2 + \varepsilon \end{aligned} \tag{32}$$

Este es un caso especial de un modelo de regresión con tres variables regresoras.

¿Cómo se interpreta este modelo para dos variables cuantitativas? En este caso decimos que existe interacción estadística cuando **la pendiente** de la relación entre la variable respuesta y una variable explicativa **cambia** para distintos niveles de las otras variables. Para entenderlo, escribamos el modelo propuesto cuando controlamos el valor de $X_2$.

$$
\begin{aligned}
E\left(Y \mid X_1, X_2 = 600\right) &= \beta_0 + \beta_1 X_1 + \beta_2 600 + \beta_{1:2} X_1 \cdot 600 \\
&= \underbrace{\beta_0 + \beta_2 600}_{\text{ordenada al origen}} + \underbrace{(\beta_1 + \beta_{1:2} 600)}_{\text{pendiente}} X_1 \\
E\left(Y \mid X_1, X_2 = 900\right) &= \beta_0 + \beta_1 X_1 + \beta_2 900 + \beta_{1:2} X_1 900 \\
&= \underbrace{\beta_0 + \beta_2 900}_{\text{ordenada al origen}} + \underbrace{(\beta_1 + \beta_{1:2} 900)}_{\text{pendiente}} X_1 \\
E\left(Y \mid X_1, X_2 = 1200\right) &= \beta_0 + \beta_1 X_1 + \beta_2 1200 + \beta_{1:2} X_1 1200 \\
&= \underbrace{\beta_0 + \beta_2 1200}_{\text{ordenada al origen}} + \underbrace{(\beta_1 + \beta_{1:2} 1200)}_{\text{pendiente}} X_1
\end{aligned}
$$

$$
\begin{aligned}
E\left(Y \mid X_1, X_2\right) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{1:2} X_1 X_2 \\
&= \underbrace{\beta_0 + \beta_2 X_2}_{\text{ordenada al origen}} + \underbrace{(\beta_1 + \beta_{1:2} X_2)}_{\text{pendiente}} X_1
\end{aligned} \tag{33}
$$

Luego, en el modelo (32), la pendiente de la relación entre $X_1$ e $Y$ depende de $X_2$, decimos entonces que existe interacción entre las variables.

Entonces, cuando $X_2$ aumenta en una unidad, la pendiente de la recta que relaciona $Y$ con $X_1$ aumenta en $\beta_{1:2}$. En este modelo, al fijar $X_2$, $E(Y \mid X_1, X_2)$ es una función lineal de $X_1$, pero la pendiente de la recta depende del valor de $X_2$. Del mismo modo, $E(Y \mid X_1, X_2)$ es una función lineal de $X_2$, pero la pendiente de la relación varía de acuerdo al valor de $X_1$.

Si $\beta_{1:2}$ no fuera estadísticamente significativa, entonces los datos no avalarían la hipótesis de que el cambio en la respuesta con un predictor dependa del valor del otro predictor, y podríamos ajustar directamente un modelo aditivo, que es mucho más fácil de interpretar.