Repaso de variables aleatorias I

DEFINICIÓN (Distribución χ^2_1): Si $Z \sim \mathrm{N}(0,1)$ entonces la distribución de $Y = Z^2$ se denomina chi-cuadrada con un grado de libertad. La notaremos $Y \sim \chi^2_1$.

Vale que $\chi^2_1 = \Gamma(1/2, 1/2)$.

Recordemos que

PROPOSICIÓN: Sean $Y_1 \sim \Gamma(\alpha_1, \lambda)$ e $Y_2 \sim \Gamma(\alpha_2, \lambda)$ variables aleatorias independientes. Entonces $Y_1 + Y_2 \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$.

DEFINICIÓN (Distribución χ^2_n): Sean Z_i, $i = 1, 2, \ldots, n$ variables independientes con distribución $\mathrm{N}(0,1)$. La distribución de la variable aleatoria $Y = \sum_{i=1}^{n} Z_i^2$ la denominaremos distribución chi-cuadrada con n grados de libertad, que simbolizaremos por χ^2_n.

Vale que $\chi^2_n = \Gamma(n/2, 1/2)$, $n \in \mathbb{N}$.

$$E(Y) = n$$
$$var(Y) = 2n$$

Repaso de variables aleatorias II

DEFINICIÓN:(Distribución t de Student): Sean $U \sim \mathrm{N}(0,1)$ y V con distribución χ^2_n con U y V independientes. Luego se define la distribución t de Student con n grados de libertad, que simbolizaremos con t_n, como la distribución de

$$T = \frac{U}{\sqrt{V/n}}.$$

La densidad de T es

$$f_T(x) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

$$\begin{aligned} E(T) &= 0, \quad n \geq 2 \\ var(T) &= \frac{n}{n-2}, \quad n > 2 \end{aligned}$$

Repaso de variables aleatorias III

DEFINICIÓN: (Distribución F de Fisher) Sean $U \sim \chi^2_n$ y $V \sim \chi^2_m$ con U y V independientes. Luego se define la distribución de F de Fisher o de Snedecor con n grados de libertad en el numerador y m grados de libertad en el denominador que notaremos $F_{n,m}$, como la distribución de

$$F = \frac{U/n}{V/m}$$

Su función de densidad está dada por

$$f_F(x) = \frac{\Gamma((m+n)/2)}{\Gamma(n/2)\Gamma(m/2)} \left(\frac{n}{m}\right)^{n/2} x^{(n/2)-1} \left(1 + \frac{n}{m}x\right)^{-(n+m)/2} I_{(0,\infty)}(x).$$

Repaso de variables aleatorias IV

OBSERVACIONES:

1. Sea $T \sim t_n$, y definimos $W = T^2 = \left(\frac{U}{\sqrt{V/n}}\right)^2$. Resulta que $W \sim F_{1,n}$.
2. La distribución $t_1 = Cauchy(0, 1)$ y no tiene esperanza ni varianza finitas.
3. La distribución t_2 tiene esperanza finita (0), pero no tiene varianza finita.

Distribución de los estimadores y tests, para el modelo lineal

Asumimos que las X están fijas.

Teorema B1

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.

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

Una combinación lineal de los estimadores

Teorema B2

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

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

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

Tests e intervalos para una combinación lineal de los β

El test de $H_0 : \mathbf{a}^T\beta = c$ versus $H_1 : \mathbf{a}^T\beta \neq c$, se basa en

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

El test de nivel α 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 $\mathbf{a}^T\beta$ está dado por

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

Tests e intervalos para q combinaciones lineales de β

Teorema B3

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

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

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

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

con nivel simultáneo α, *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* $A\beta$ *estará dada por* $F \leq F_{q,n-p}\left(1-\alpha\right)$, *donde* F *está definido en (i.)*

Estimador de β con restricciones

Teorema B4

Bajo $\mathbf{Y} = X\beta + \varepsilon$ *con* $\varepsilon \sim N_p\left(\mathbf{0}, \sigma^2 I_n\right)$, *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* β *sujeto a las restricciones dadas por* H_0, *que llamaremos* $\widehat{\beta}_H$ *está dado por*

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

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* $\widehat{\mathbf{Y}}_H = X\widehat{\beta}_H$ *entonces,* $S\left(\widehat{\beta}_H\right) = RSS_H = \left\|\mathbf{Y} - \widehat{\mathbf{Y}}_H\right\|^2 = \left\|\mathbf{Y} - X\widehat{\beta}_H\right\|^2$

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

Estimador de mínimos cuadrados con restricciones

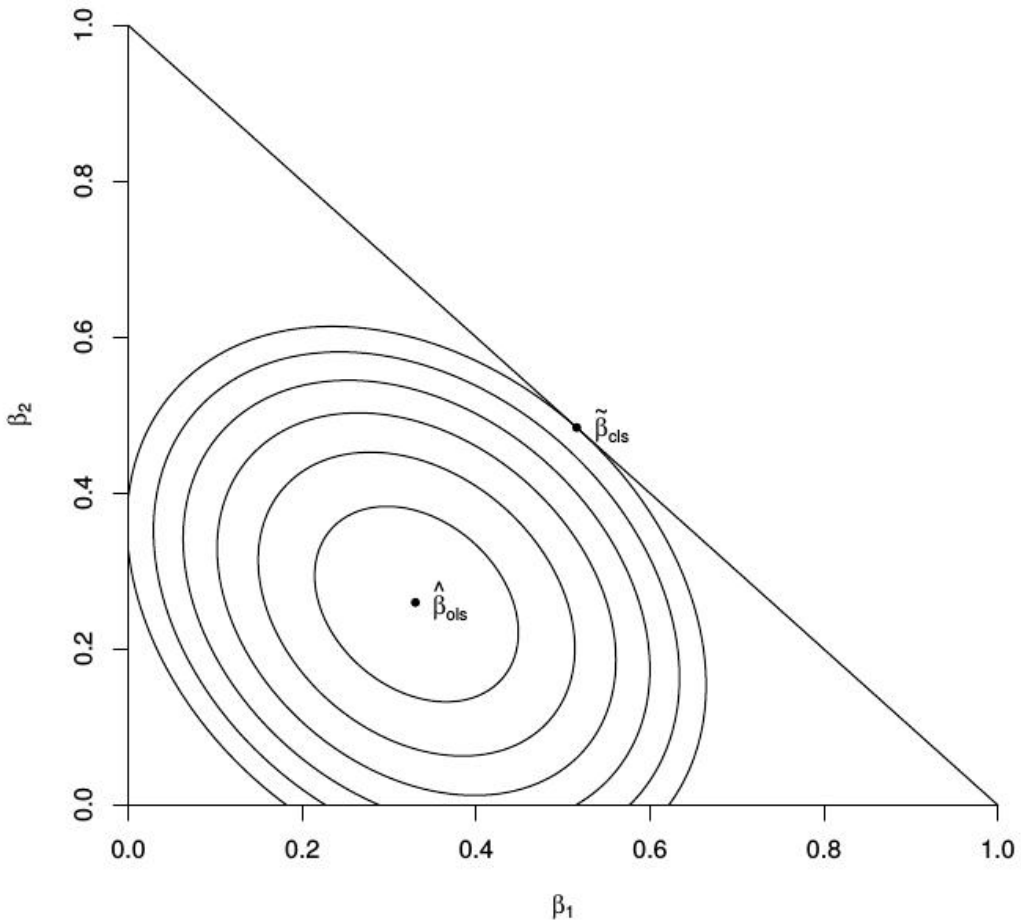

Figure 8.1: Imposing a Constraint on the Least Squares Criterion

Fuente: Econometrics. Bruce E.

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

Test F y proyecciones

Teorema B5

Bajo $\mathbf{Y} = X\beta + \varepsilon$ *con* $\varepsilon \sim N_p\left(\mathbf{0}, \sigma^2 I_n\right)$, *rango*$(X) = p$. *Sea* $A \in \mathbb{R}^{q \times p}$ *de rango* q. *Sea* $H : A\beta = \mathbf{c} \in \mathbb{R}^q$. *Escribimos* H *en vez de* H_0 *para simplificar la notación.*

i.

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

ii. *Bajo* H,

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

Test F y proyecciones II

Teorema B5 (cont.)

iii. *Cuando* $\mathbf{c} = \mathbf{0}$, *podemos definir los subespacios*

$$\begin{aligned} \Omega &= gen(X) = \{X\beta : \beta \in \mathbb{R}^p\} \\ \omega &= \{X\beta : \beta \in \mathbb{R}^p \text{ y } A\beta = \mathbf{0}\} \\ \gamma &= gen\left\{X\left(X^TX\right)^{-1}A^T\right\} \end{aligned}$$

de dimensiones $p, p-q, q$ *respectivamente. Además* $\gamma \subset \Omega$ *y* $\omega \subset \Omega$. *Luego,* $\omega = \Omega \cap \gamma^{\perp}$ *y*

$$F = \left(\frac{n-p}{q}\right)\frac{Y^T\left(P_\Omega - P_\omega\right)Y}{Y^T\left(I_n - P_\omega\right)Y}$$

donde P_Ω *y* P_ω *son las matrices de proyección a los subespacios* Ω *y* ω *de* $\mathbb{R}^n$ *respectivamente.*

Interpretación geométrica del test F

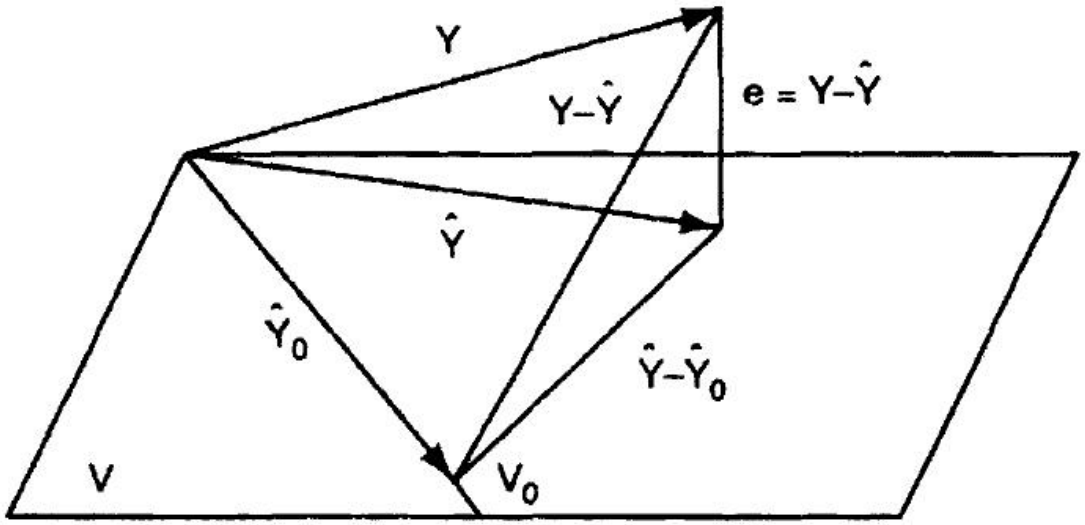

FIGURE 3.10 Illustration for the F-test of H_0: $\theta \in V_0$.

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

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

Test F en el caso de rango no completo

Bajo $\mathbf{Y} = X\beta + \varepsilon$ con $\varepsilon \sim N_p\left(\mathbf{0}, \sigma^2 I_n\right)$, rango$(X) = r \leq n$. Queremos generalizar lo anterior. Podemos escribir el modelo:

$$\mathbf{Y} = \boldsymbol{\theta} + \varepsilon,\ \boldsymbol{\theta} \in \Omega = \text{gen}(X) = \{X\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p\}$$

Sea $H : \boldsymbol{\theta} \in \omega$ con ω un subespacio $(r - q)$ dimensional de $\mathbb{R}^n$, $\omega \subset \Omega$.

Teorema B6

Si H es verdadera y $\varepsilon \sim N_p\left(\mathbf{0}, \sigma^2 I_n\right)$, rango$(X) = r$, entonces

$$F = \frac{(RSS_H - RSS)/q}{RSS/(n-r)} = \left(\frac{n-r}{q}\right) \frac{\varepsilon^T (P_\Omega - P_\omega)\,\varepsilon}{\varepsilon^T (I_n - P_\omega)\,\varepsilon},$$

$F \sim F_{q,n-r}$ donde P_Ω y P_ω son las matrices de proyección a los subespacios Ω y ω de $\mathbb{R}^n$ respectivamente.

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

Ejemplo: Ajuste *low birth weight data*

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

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i, \qquad (1)$$

```
> 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*

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

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

tiene p–valor igual a 0.8995, 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 (1), 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 por lo cual, esta variable no resulta significativa en el modelo (1). Tampoco resulta significativa `toxemia`) (p–valor 0.165) cuando se incluyen las otras cuatro covariables en el modelo.

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 venimos discutiendo en clase. 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, $\Omega = gen(X)$ y tiene dimensión 6 y $H_0 : A\beta = 0$. Luego

$$\omega = gen(\begin{bmatrix} X(0) & X(2) & X(3) \end{bmatrix})$$

donde $X(j)$ es la columna j–ésima de X, por lo que $dim(\omega) = 3$. El tamaño de muestra es $n = 100$

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
```

Ejemplo: Ajuste *low birth weight data*

En este último ajuste vemos que los coeficientes de las dos covariables resultan significativamente distintas de cero, para explicar al perímetro cefálico cuando la otra covariable es incluida en el modelo.

La superficie (el plano) ajustada resulta ser

$$\widehat{Y}_i = 8.308 + 0.44873 X_{i2} + 0.00471 X_{i3}, \qquad \text{Modelo ajustado}$$

En el script `regmultiple.R` pueden verse las instrucciones de los gráficos 2D y 3D de este conjunto de datos que vimos en clase.