Proceso de modelización estadística

lunes, 5 de agosto de 2024 9:29

1er. Especificación del modelo:
- Qué variables intervienen, que supuestos se deben cumplir, forma funcional, cuáles son los parámetros a estimar.

2do. Estimación del modelo:
- Construir estimaciones para los parámetros a partir de una muestra

3er. Inferencia del modelo
- Pruebas de hipótesis sobre los parámetros y la bondad de ajuste del modelo.

4to. Validación del modelo
- Diagnóstico y comprobación de supuestos del modelo.

Especificación de un modelo de regresión lineal simple

lunes, 5 de agosto de 2024 9:38

Variables: Dos variables (simple)
- X = variable independiente (numéricas continuas)
- Y = variable dependiente (numéricas continuas)

Forma funcional:
Y = F(X) + E
Donde:
F(X) = Función de regresión
E = término de error

Supuestos estructurales:
1. Linealidad: F(X) es una ecuación de una recta
2. Homocedasticidad: Var[E | X] = cte = σ^2
3. Normalidad: E sigue una distribución N(0, σ^2)
4. Independencia: No debe existir relación entre un error y otro

NOTA: A partir de ahora vamos a SUPONER que los supuestos se cumplen. Pero, una vez estimado el modelo, el cumplimiento de estos supuestos se deben VERIFICAR!!!

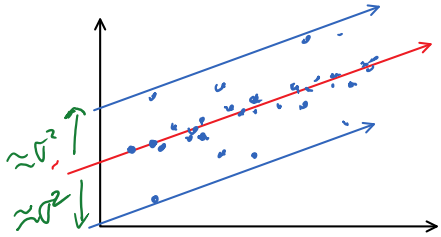
APLICACIÓN DE SUPUESTOS EN EL MODELO:

1. Linealidad: F(X) = a + bX
   a = Intersección,
   b = pendiente
   Como a y b son desconocidos, la idea es estimarlos a partir de una muestra.

PARAMETROS	INTERPRETACION EN EL MODELO
β_0	Intersección en el eje Y
β_1	Pendiente de la recta

   Forma funcional:
   Y = F(X) + E
   Y = β_0 + β_1 X + E

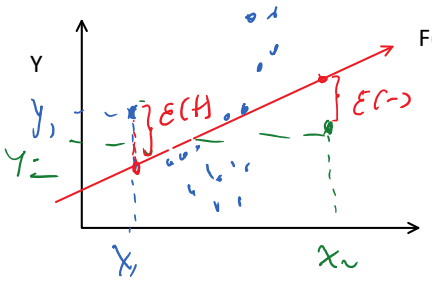
2. Homocedasticidad:
   Los errores al ser v.a. pueden tomar cualquier valor, independientemente del valor de X, pero, deben tener una magnitud de variabilidad constante
   Gráficamente, deben formar una "banda de ancho fijo" alrededor de la función de regresión



F(X) = β_0 + β_1 X

4. Independencia:
   Los errores deben ser independientes entre sí.
   Gráficamente, los errores no deben presentar ningún tipo de ordenamiento o patrón.

¿Qué representa el término de error E en el modelo?
- Si el término de error es CERO; los pares (x, y) estarían exactamente sobre la RECTA.
- Como asumimos que los errores no son cero, entonces los puntos pueden estar sobre (errores positivos) o bajo (errores negativos) la recta.
- Los errores pueden tomar cualquier valor (grande o pequeño)
- Entonces los Errores son una VARIABLE ALEATORIA!!



3. Normalidad:
   Los errores son v.a. que siguen una distribución normal, con media cero y varianza homocedástica σ^2

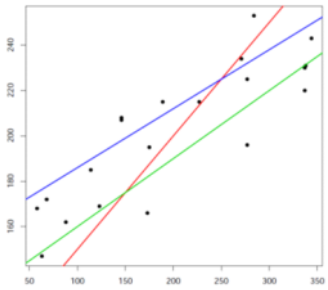
   La mayoría de los errores deben ser cero o muy cercanos a cero, una minoría pueden ser muy positivo o negativos, pero sin salirse de la banda homocedástica!
   Además, se espera que la mitad de los puntos tengan errores positivos, y la otra mitad tengan errores negativos

Estimación del modelo

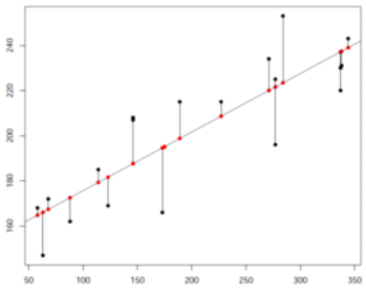
lunes, 5 de agosto de 2024 10:30

En todo modelo de RLS, se deben estimar tres parámetros: $\beta_0, \beta_1, \sigma^2$

Para estimar los parámetros, se toma una muestra de n pares observados de (X, Y) y se aplica un método de estimación: MINIMOS CUADRADOS ORDINARIOS.



Cada recta tiene sus propios residuos



Los residuos pueden ser (+) o (-)

Medida de ajuste = residuos al cuadrado

Residuos del modelo RLS

Supongamos el modelo $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, con $i = 1, \ldots, n$ y denotemos por $\hat{\beta}_0$ y $\hat{\beta}_1$ los estimadores de los parámetros. Los errores de predicción son:

$$\hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad i = 1, \ldots, n$$

también denominados residuos de la regresión.

Para un conjunto de datos (Xi, Yi) se pueden construir infinitos modelos (rectas)

OBJETIVO:

- Hallar el modelo (recta) que mejor se "ajuste" a los datos
- Es decir: que tenga los residuos más pequeños posibles.

$\min \sum \hat{\varepsilon}^2$

Problema de optimización numérica!!

OBJETIVO: Construir una recta que minimice la suma de residuos cuadrados!!

Para hallar la estimación de los parámetros, se debe derivar y luego igualar a cero, la primera derivada del criterio de optimización

Estimación por Mínimos Cuadrados

Consiste en seleccionar $\hat{\beta}_0$ y $\hat{\beta}_1$ tales que:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{(\beta_0, \beta_1)} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

Estimadores

$$\hat{\beta}_0 = \bar{Y} - \frac{S_{xY}}{S_x^2} \bar{x} \qquad \hat{\beta}_1 = \frac{S_{xY}}{S_x^2}$$

La recta de regresión estimada por mínimos cuadrados es la que pasa por el vector de medias o centro de gravedad, $(\bar{x}, \bar{Y})$, y tiene pendiente $\hat{\beta}_1 = \frac{S_{xY}}{S_x^2}$.

CONSULTA!!:
Desarrollar las ecuaciones de las estimaciones de β_0 y β_1 por mínimos cuadrados!

Estimación de σ^2

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

Ejm: Estimación

lunes, 12 de agosto de 2024 9:33

Values	
Beta0	141.703353881689
Beta1	0.294259414639639
n	10
S2x	7282.01
Sxy	2142.8
vecmed	num [1:2] 225 208
x	num [1:10] 175 189 344 88 114 338 2.
y	num [1:10] 195 215 243 162 185 231 .

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$

$\hat{y} =$ 141.70 + 0.29 X

Interpretación:
Proyección tiempo de uso de máquina = 141.70 + 0.29
*(unidades producidas en cada corrida de producción)

- $\hat{\beta}_0$= 141.70 min (tiempo promedio de uso de máquina, cuando no se produce ninguna unidad, X =0) - costos fijos
- $\hat{\beta}_1$= 0.29 min/unidad (tasa marginal unitaria). Si se produce una unidad adicional, el tiempo de uso de máquina AUMENTA en 0.29 minutos (relación directa)

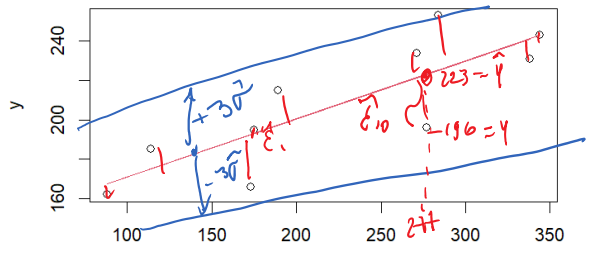
```
> Yproy = Beta0 + Beta1*x
> Yproy
 [1] 193.1988 197.3184 242.9286 167.5982 175.2489 241.1630 221.4477 192.6102
 [9] 225.2730 223.2132

> y
 [1] 195 215 243 162 185 231 234 166 253 196
> resid = y - Yproy
> resid
 [1]   1.80124856  17.68161675   0.07140748  -5.59818237   9.75107285
 [6] -10.16303603  12.55234475 -26.61023261  27.72697236 -27.21321174
```

$\hat{\varepsilon} = y - \hat{y}$

$\hat{\varepsilon}_1 = 195 - 193,1988 = +1,80$

$\hat{\varepsilon}_{10} = 196 - 223,21 = -27,213$




```
> sum(resid) # practicamente CERO
[1] 2.842171e-14
```

$\sum \hat{\varepsilon} \approx 0 \rightarrow \frac{\sum \hat{\varepsilon}}{n} = \bar{\hat{\varepsilon}} = 0$

```
> SCR = sum(resid^2)
> SCR
[1] 2920.609
```

$\sum \hat{\varepsilon}^2 \leftarrow$ min $\leftarrow$ MCO

```
> sigma2 = SCR / (n-2)
> sigma2
[1] 365.0762
```

$\hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}^2}{n-2} = 365,0762$

```
> sigma = sqrt(sigma2)
> sigma
[1] 19.10697
```

$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = 19,10$

Inferencia del modelo: Pruebas T individuales

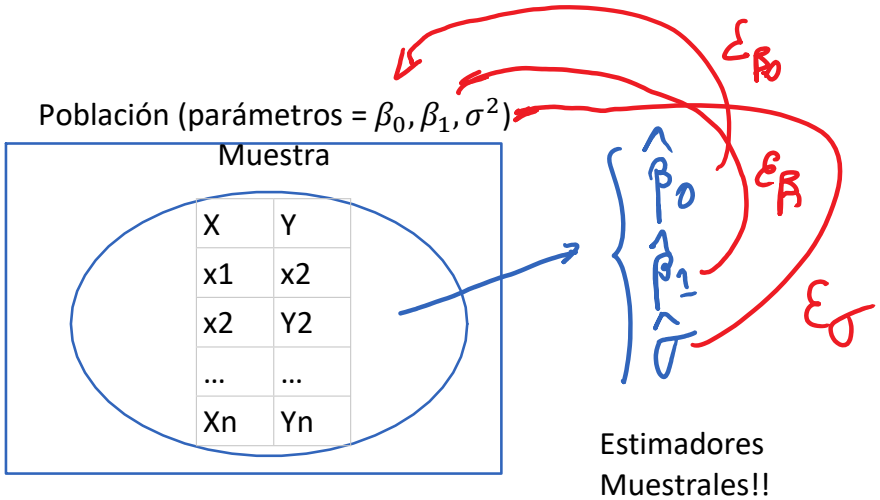
lunes, 12 de agosto de 2024 9:52

- Una cuestión es indicar cuánto valen las estimaciones del modelo, otra cosa es que esas estimaciones sean válidas. Porque estas estimaciones se obtuvieron a partir de una muestras, y toda muestra tiene ERRORES!!

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}^2}{n-2}$$



Parámetro	Estimador	Error muestral	Error típico	Estadístico Pivote
β_0	$\hat{\beta}_0$	$\lvert\hat{\beta}_0 - \beta_0\rvert$	$\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{nS_x^2}}$	$\frac{(\hat{\beta}_0 - \beta_0)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{nS_x^2}}} = t_{(n-2)}$
β_1	$\hat{\beta}_1$	$\lvert\hat{\beta}_1 - \beta_1\rvert$	$\hat{\sigma}/(S_x\sqrt{n})$	$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/(S_x\sqrt{n})} = t_{(n-2)}$
σ^2	$\hat{\sigma}^2$	$\hat{\sigma}^2/\sigma^2$		$\chi^2 = (n-2)\hat{\sigma}^2/\sigma^2$

Ejm: Inferencia sobre los betas

lunes, 12 de agosto de 2024 10:07

Inferencia para el Beta Cero

```
> etB0 = sigma * sqrt((1/n)+((mean(x)^2)/(n*S2x)))
> etB0 # estimacion del error tipico del beta cero
[1] 17.05838

> tB0 = Beta0 / etB0
> tB0 # estadistico de contraste de Beta 0
[1] 8.306967
```

Parámetro	Estimador	Error muestral	Error típico	Estadístico Pivote
β_0	$\hat{\beta}_0$	$\lvert\hat{\beta}_0 - \beta_0\rvert$	$\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{nS_x^2}}$	$\frac{(\hat{\beta}_0 - \beta_0)}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{nS_x^2}}} = t_{(n-2)}$

$\sigma_{\hat{\beta}_0}$ = 17,05

$t = \frac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}} = 8,30$ con $gl = (n-2) = 8$

$H_0: \beta_0 = 0$

$H_1: \beta_0 \neq 0$

P_Valor ↓

```
> PvalorB0 = 2*(1 - pt(tB0, n-2))
> PvalorB0 # es menor al 5% , H0 se
[1] 3.326772e-05
```

$< 5\%$ →

H0 se rechaza (Rechazamos que el Beto cero = 0)
Si existen costos fijos en el modelo!!

Inferencia para el Beta 1:

β_1	$\hat{\beta}_1$	$\lvert\hat{\beta}_1 - \beta_1\rvert$	$\hat{\sigma}/(S_x\sqrt{n})$	$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/(S_x\sqrt{n})} = t_{(n-2)}$
			0,0708	

$\hat{\sigma}_{\hat{\beta}_1}$

$\hat{\sigma}/\sqrt{n S_x^2}$

```
> etB1 = sigma / sqrt(n*S2x)
> etB1
[1] 0.07080535
```

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$t = \frac{\hat{\beta}_1 - \beta_0}{\sigma_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\sigma_{\hat{\beta}_1}} = \frac{0,2942}{0,0708} = 4,155$

```
> tB1 = Beta1 / etB1
> tB1
[1] 4.155892
```

P_Valor = 0,003 ($< 5\%$)

```
> PvalorB1 = 2*(1 - pt(tB1, n-2))
> PvalorB1 # es menor al 5% , H0 s
[1] 0.003183112
```

H0 se rechaza
La tasa marginal unitaria de X es distinta a cero!

Inferencia del modelo: Prueba F global

lunes, 12 de agosto de 2024 10:20

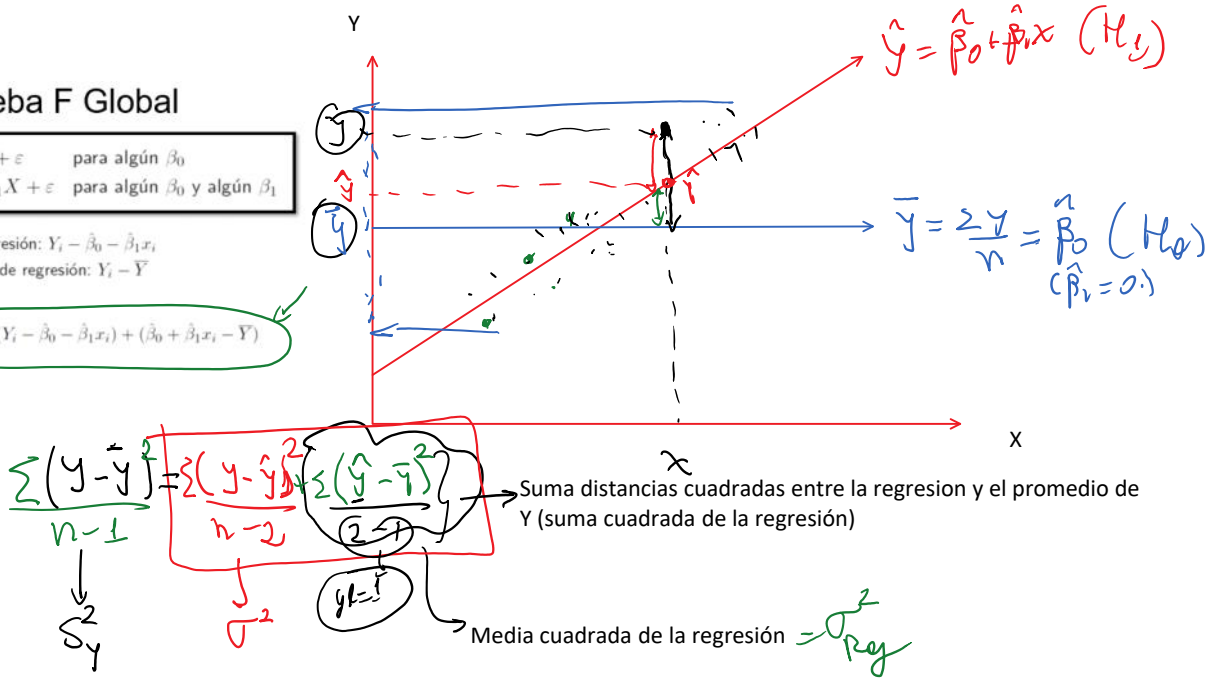
Prueba F Global

$H_0: \quad Y = \beta_0 + \varepsilon \quad$ para algún β_0

$H_a: \quad Y = \beta_0 + \beta_1 X + \varepsilon \quad$ para algún β_0 y algún β_1

- Residuos de la regresión: $Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$
- Residuos sin recta de regresión: $Y_i - \bar{Y}$
- Descomposición:

$$Y_i - \bar{Y} = (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) + (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{Y})$$



$\sum (y - \bar{y})^2 = \sum (y - \hat{y})^2 + \sum (\hat{y} - \bar{y})^2$

$n-1$ → S_y^2 ; $n-2$ → σ^2 ; $gl = 1$

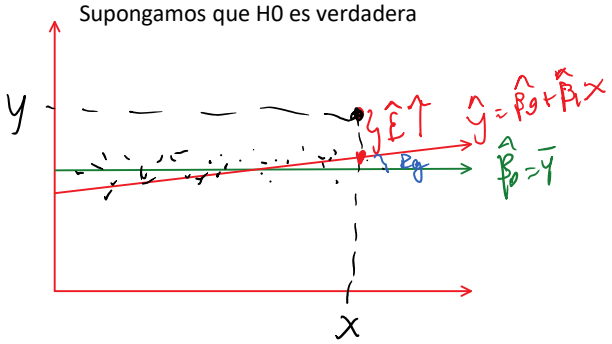
Suma distancias cuadradas entre la regresion y el promedio de Y (suma cuadrada de la regresión)

Media cuadrada de la regresión $= \sigma^2_{Reg}$

TABLA DE ANALISIS DE VARIANZAS DE LA REGRESION LINEAL

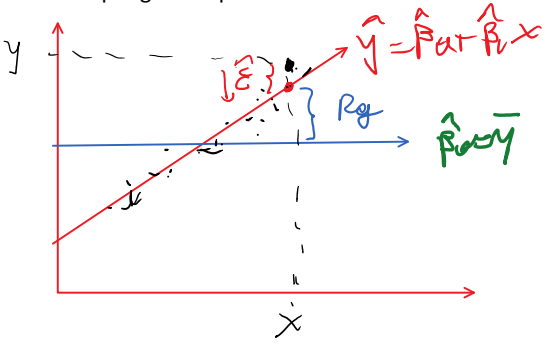
Fuente de variabilidad	Sumas cuadradas	Grados de libertad	Medias cuadradas	Estadístico F
Regresión	$\sum (\hat{y} - \bar{y})^2$	$2-1 = 1$	$= \sigma^2_{Reg}$	
Error	$\sum (y - \hat{y})^2 = \sum \hat{\varepsilon}^2$	$n-2$	$= \sigma^2$	
Total	$\sum (y - \bar{y})^2$	$n-1$	$= S_y^2$	

Supongamos que H0 es verdadera



El residuo es más grande que la distancia entre la recta y el promedio de Y

Supongamos que H0 es falsa



La recta de regresión se aleja del promedio de Y, y los residuos se hacen más pequeños!!

Luego, para que el modelo tenga sentido, se requiere que:

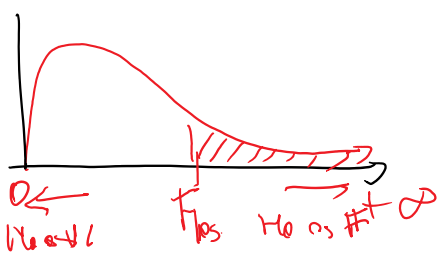
- Los residuos sea pequeños.
- La regresión se aleje del promedio de Y

A partir de esto, se construye un estadístico de contraste F:

- Si la H0 es verdadera, F tiende a cero
- Si la H0 es falsa, F tiende a + infinito

$$F = \frac{\sigma^2_{Reg}}{\sigma^2_{\hat{\varepsilon}}} = \frac{\frac{\sum (\hat{y} - \bar{y})^2}{1}}{\frac{\sum \hat{\varepsilon}^2}{n-2}}$$

$1 \leftarrow gl$ (num) ; $n-2 \leftarrow gl$ (denom)



$p\_Valor < \alpha \rightarrow$ Ho es F

$p\_Valor \geq \alpha \rightarrow$ Ho es V

Ejm: Inferencia: Prueba F global

lunes, 12 de agosto de 2024 10:48

```
> SCT = sum((y - mean(y))^2) # suma cuadrada total/global
> SCT
[1] 9226
```

$\rightarrow SCT = \sum (y - \bar{y})^2 = 9226$

```
> SCReg = sum((Yproy - mean(y))^2) # suma cuadrada de la regresion
> SCReg
[1] 6305.391
```

$\rightarrow SCReg = \sum (\hat{y} - \bar{y})^2 = 6305{,}391$

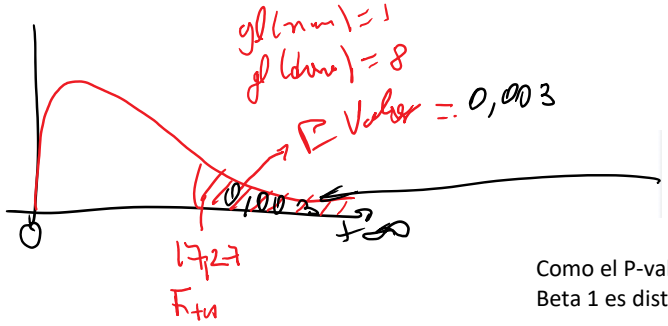
```
> SCR = sum((y - Yproy)^2) # suma cuadrada del residuo
> SCR
[1] 2920.609
```

$\rightarrow SCE = \sum (y - \hat{y})^2 = \sum \hat{\varepsilon}^2 = 2920{,}609$

```
> SCReg + SCR #esta suma debe ser igual a la SCT
[1] 9226
```

```
> # PRUEBA GLOBAL F (contraste F)
> Ftest = (SCReg/1) / (SCR/(n-2)) # estadistico F
> Ftest
[1] 17.27144
```

$$F_{tnt} = \frac{\frac{SCReg}{1}}{\frac{SCE}{n-2}} = \frac{\left(\frac{6305{,}391}{1}\right)}{\left(\frac{2920{,}609}{8}\right)} = 17{,}27144$$




```
> PvalorF = 1 - pf(Ftest, 1, n-2)
> PvalorF # Este valor es menor al 5%, por tanto la H0 se rechaza
[1] 0.003183112
> # Beta1 es distinto de cero, luego Y si depende de X
```

Como el P-valor es menor al 5%, la H0 se rechaza!!
Beta 1 es distinto de cero (Y depende de X)