Ayudantía N°1

Ignacio Sepulveda.

2022-09-03

Respuestas

Primero nos aseguramos de que estamos trabajando en el directorio correcto

```
rm(list=ls()) ## Remueve
getwd() ## Obtiene tu directorio
```

```
## [1] "C:/Users/IgnacioSepulveda/Documents/Ayudantias/econometrics"
```

```
setwd("C:/Users/IgnacioSepulveda/Documents/Ayudantias/econometrics")## Setealo
```

Ejercicio 1

1. Para estudiar la relación entre la altura de los hijos y la altura de los padres, un estudiante considera el siguiente modelo de regresión lineal:

$$Y_i = \beta_0 + \beta_1 X_i + \mu_i$$

donde Yi es la altura de un individuo en metros y Xi es la altura de su padre en metros. El estudiante pregunta a 40 individuos y obtiene la siguiente estimación:

$$\underset{(0,5)}{\widehat{Y_i} = 0{,}1434} + \underset{(0,2828)}{0{,}9342 X_i}$$

Suponemos que se cumplen las hipótesis del modelo de regresión lineal con errores normales.

a) Contraste la hipótesis nula de que la altura del padre no es significativa.

Respuesta 1

Queremos testear la hipótesis si $H_0 : \beta_1 \neq 0 \, vs \, H_1 : \beta_1 = 0$.

La formula para el estadístico t viene dada por:

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\beta_1)} \sim t_{\alpha/2, N-k-1}$$

Reemplazamos,

```
beta_1_estimado=0.9342
beta_1_poblacional=0
SE=0.2828
### Calculamos el t estadístico
(beta_1_estimado-beta_1_poblacional)/SE
```

```
## [1] 3.303395
```

Ahora debemos calcular el t critico,

```
alpha=0.05
N=40
K=1
abs(qt(alpha/2,N-K-1))
```

```
## [1] 2.024394
```

Notamos que nuestro t calculado es mayor al t critico, por lo que rechazamos nuestra hipótesis nula, por lo tanto nuestra variable es significativa. Osea que tenemos evidencia de que la altura del padre es significativa.

Ejercicio 2

2. Un investigador quiere analizar el efecto que tiene la edad de una persona (age) sobre la cantidad de cigarrillos que fuma diariamente (cigs).

a) Proponga un modelo econométrico que represente la investigación a realizar.
b) Basado en el fichero SMOKE, estime el modelo propuesto en a) y presente los resultados en forma de ecuación.
c) Interprete los parámetros del modelo y el coeficiente de determinación.
d) Si una persona tiene 20 años, ¿Cuánto estima el modelo que fumará diariamente?
e) Verifique si existe evidencia suficiente para afirmar que, a mayor edad, mayor es la cantidad de cigarrillos diarios consumidos (asumiendo que se cumplen los supuestos del MRLC).

Respuesta 2

a)

El modelo que nos permite hacer la investigación es,

$$cigs_i = \beta_0 + \beta_1 agei + \mu_i$$

Lo que estamos proponiendo en este caso es que la cantidad de cigarros consumidos por una persona es una función lineal de la edad mas un componente aleatorio, que es independiente de la edad, tiene media igual a 0 y varianza constante.

b)

Librerías ocupadas.

```
library(haven) ## Para leer dta
library(tidyverse) ## Manipular datos
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Leemos los datos.

```
df=read_dta('SMOKE.dta')
```

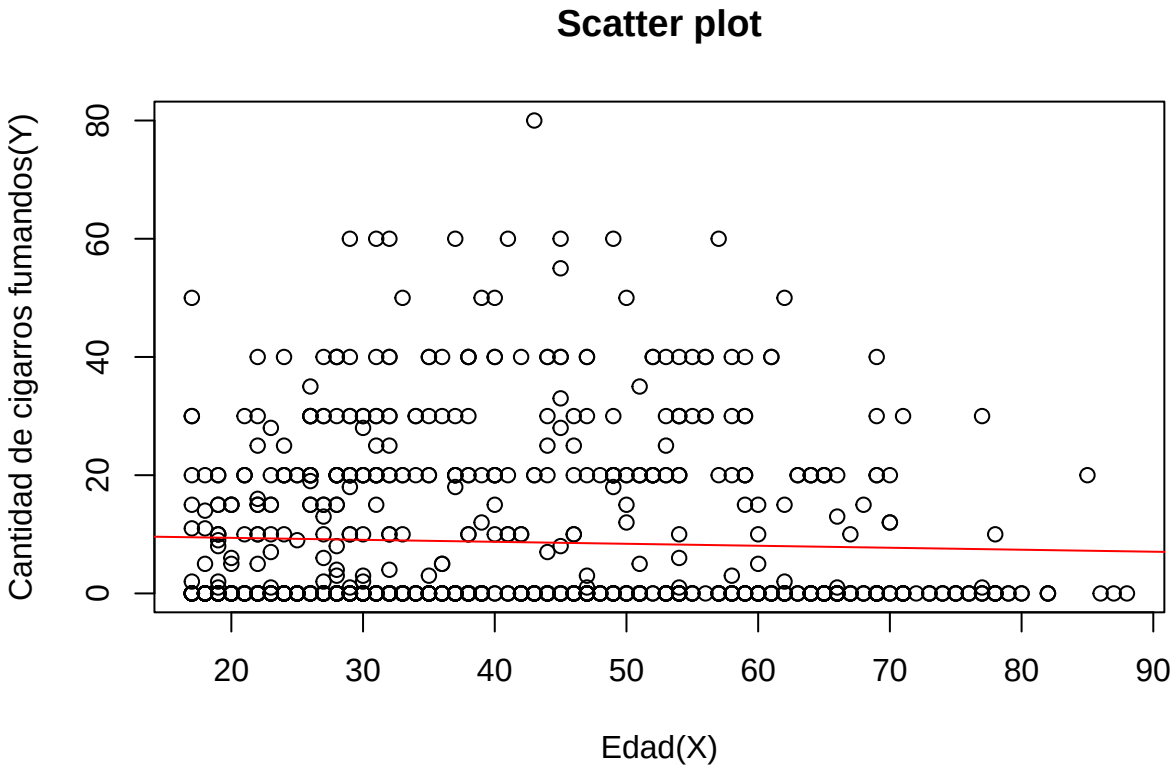
Ajustamos el modelo propuesto en los datos.

```
reg=lm(cigs~age,data=df)
reg %>% summary()
```

```
##
## Call:
## lm(formula = cigs ~ age, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.498 -8.929 -7.991 10.669 71.372
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.06698    1.26597   7.952  6.2e-15 ***
## age         -0.03348    0.02838  -1.180    0.238
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.72 on 805 degrees of freedom
## Multiple R-squared:  0.001726,   Adjusted R-squared:  0.0004856
## F-statistic: 1.392 on 1 and 805 DF,  p-value: 0.2385
```

El R^2 es bastante bajo, una inspección al scatter plot nos mostrara el espacio donde varía la cantidad de cigarros y la edad lo que nos permitirá tener una intuición visual de cual es la relación, o bien ver si esta se acerca a una lineal.

```r
plot(x=df$age,
     y=df$cigs,
     ylab='Cantidad de cigarros fumandos(Y)',
     xlab='Edad(X)',
     main='Scatter plot')
abline(reg,col='red') ## Ajusta la regresion estimada
```



c)

- β_0: Nos dice que si una persona tiene 0 años, en promedio fumara 10 cigarros.
- β_1: Nos dice que por cada año de edad mas, la cantidad de cigarrillos disminuye en promedio en 0.033 unidades.
- R^2: Nos dice que la variabilidad en la edad explica el 0.17% de la variabilidad en la cantidad de cigarros.

d)

Para saber cuantos cigarros fuma una persona de 20 años debemos reemplazarlo en la regresión lineal muestral, que cuenta con nuestros parámetros estimados en el punto b). Entonces partimos definiendo los parámetros y la función.

```r
# Creamos la función
regresion_muestral=function(x,beta_0,beta_1){
  return(beta_0+beta_1*x)
}
```

```r
# Parametros estimados
beta_0=10.06
beta_1=-0.033
x=20 ## Valor a predecir
regresion_muestral(x,beta_0,beta_1) ## Predecimos
```

```
## [1] 9.4
```

Entonces para una edad de 20 años predecimos un promedio de 9.4 cigarrillos.

e)

Lo que queremos hacer es testear si el coeficiente es igual mayor a cero, por lo tanto nuestras hipótesis son las siguientes $H_0 : \beta_1 > 0 \, vs \, H_1 : \beta_1 \leq 0$.

El estadístico a ocupar es el t, el cual tiene la siguiente formula.

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{\alpha, n-k-1}$$

. Es α dado que es test de un cola. Si hubiera sido de dos colas sería $\alpha/2$,

Ya definimos una variable que tiene guardado al $\hat{\beta}_1$, por lo tanto nos falta $SE(\hat{\beta}_1)$

```r
SE=0.028 ## lo sacamos de la regresión en el punto b)
t_calculado=beta_1/SE ## Sera nuestro t critico
t_calculado
```

```
## [1] -1.178571
```

Notas que es el mismo valor t que se encuentra en la tabla de regresión.

Ahora nos falta el $t^c = t_{0.05,807-1-1}$

```r
alpha=0.05 # Significancia
N=length(df$cigs) #Total datos
K=1 #Q de parametros
t_critico=qt(alpha,N-K-1) ## Quantil de la distribucion de densidad de probabilidad
t_critico
```

```
## [1] -1.646749
```

Como el t calculado es mayor que el t critico, no podemos rechazar H_0 o no podemos rechazar que el efecto es distinto de 0. Otra forma de hacerlo es calcular el p-value.

```
round(pt(t_calculado,N-K-1)*100,5) ## p-value, nos dice la significancia minima desde la cual podemos r
```

```
## [1] 11.94586
```

```
## En caso de que hubieran sido dos colas debiese haber sido *2.
```

Comprobemos.

```
nuevo_t_critico=qt(0.1194586,N-K-1) ### Debería dar lo mismo que el t-calculado.
paste('t critico nuevo:',round(nuevo_t_critico,4))
```

```
## [1] "t critico nuevo: -1.1786"
```

```
paste('t calculado:',round(t_calculado,4))
```

```
## [1] "t calculado: -1.1786"
```

Ejercicio 3

3. Se propone el siguiente modelo para determinar el efecto de la educación en el salario de una persona

$$wage_i = \beta_0 + \beta_1 educ_i + \mu_i$$

donde "*wage*" es el ingreso mensual en dólares y "*educ*" son los años de educación.

a) ¿Qué signo se espera que tenga B1? ¿Por qué?
b) Con el fichero WAGE2, estime el modelo propuesto y presente los resultados en forma de ecuación.
c) Si la educación aumenta en 5 años, ¿Cuánto predice el modelo que variará el salario de una persona?
d) Con solo ver la información en la tabla de R, determine la significancia individual de la variable educación.
e) Realice el contraste de significancia individual basado en las zonas de rechazo.

a)

Si aumentan los años de educación, se espera que el ingreso mensual también aumente, por lo que β_1 debería ser positivo.

b)

Leemos los datos

```
rm(list=ls()) ## Removemos
df=read_dta('WAGE2.dta') ## Leemos
```

Ajustamos el modelo

```
reg=lm(wage~educ,data=df)
reg %>% summary()
```

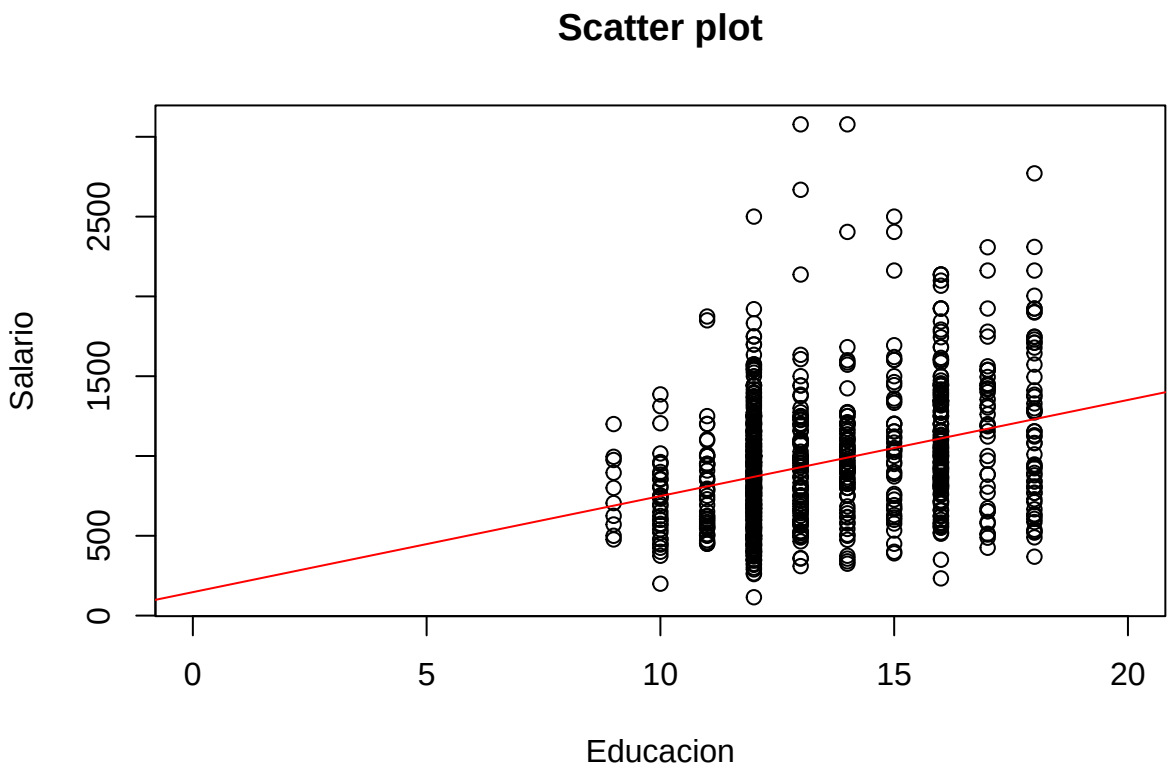
```
##
## Call:
## lm(formula = wage ~ educ, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -877.38 -268.63  -38.38  207.05 2148.26
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  146.952     77.715   1.891   0.0589 .
## educ          60.214      5.695  10.573   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 382.3 on 933 degrees of freedom
## Multiple R-squared:  0.107,  Adjusted R-squared:  0.106
## F-statistic: 111.8 on 1 and 933 DF,  p-value: < 2.2e-16
```

Si armamos la ecuación tenemos que,

$$\widehat{Wage}_i = 146.95 + 60.21\widehat{Educ}_i$$

Notemos que el R^2 es bastante mejor la regresión del ejercicio anterior. Lo cual lo podemos ver reflejado en el scatter plot.

```
plot(
     x=df$educ,
     y=df$wage,
     ylab='Salario',
     xlab='Educacion',
     main='Scatter plot',xlim=c(0,20))
abline(a=146.95,b=60.21,col='red',xlim=c(0,20))
```



c)

Recordemos que β_1 representa cuanto aumenta en promedio el salario cuando aumenta en un año la educación. Por lo tanto si aumenta 5 años basta multiplicar el β_1 por 5 y obtendremos el aumento total del salario. Esta es la gracia de la función lineal.

```
beta_1=60.21
beta_1*5
```

```
## [1] 301.05
```

Entonces un aumento de 5 años de educación aumenta el salario en promedio en 301.05 dolares.

d)

La tabla nos muestra que el p-value es muy chico, lo que nos indica que podemos rechazar la hipótesis nula, por lo tanto educación es significativa explicando el salario.

e)

Queremos testear la siguiente hipótesis $H_0 : \beta_1 \neq 0 \, vs \, H_1 : \beta_1 = 0$.

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\beta_1)} \sim t_{\alpha/2, N-k-1}$$

.

Reemplazamos y obtenemos.

```
SE=5.695
t_calculado=beta_1/SE
t_calculado
```

```
## [1] 10.57243
```

El estadístico t es grande por lo que es evidente que rechazaremos. **Dado que es de dos colas!**

```
alpha=0.05 # Significancia
N=length(df$wage) #Total datos
K=1 #Q de parametros
t_critico=qt(alpha/2,N-K-1) ## Quantil de la distribucion de densidad de probabilidad
abs(t_critico)
```

```
## [1] 1.96251
```

Queda mas claro cuando vemos el t-critico.