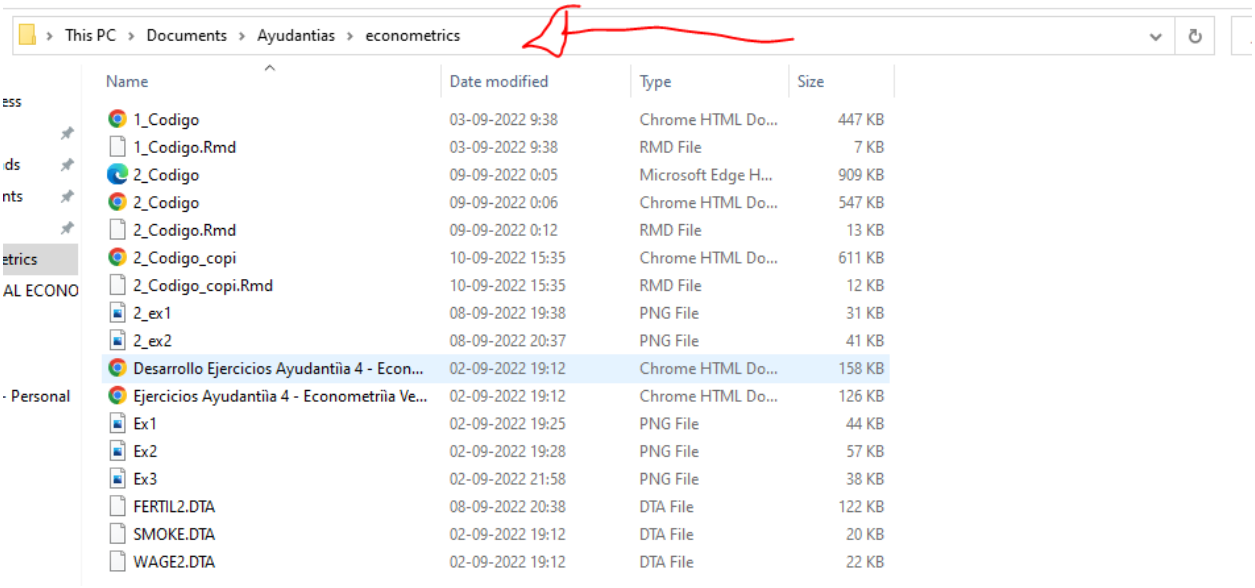
Codigo ayudantia 2

2022-09-09

Codigo de ayudantia de hoy, cualquier cosa me escriben *mail:* [ignacio.sepulveda.l@usach.cl] *cel:*+569 68484764

Siempre que comencemos en R studio, dos cosas que es bueno tener claro. La primera es que el ambiente de trabajo se encuentra limpio y la otra es que tener claridad en que el directorio en el que trabajamos es el correcto.

Para saber en que directorio esta los datos, una forma es ingresar al explorador de archivos(file explorer) dirigirse a la carpeta donde se encuentra nuestro archivo y damos click derecho sobre donde apunta la flecha roja.



Si esta trabajando en archivo .R, que es el común, copie solo el código para obtener los resultados.

```
## Directorio

## Limpiamos el ambiente
rm(list=ls())

## Nos da el directorio actual
getwd()
```

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

```
### Escogemos el directorio, aquí deberían estar los datos,
### esto cambiara de pc en pc.Ver la imagen para saber como
### reconocer en cual directorio esta actualmente el archivo.
```

```r
setwd("C:/Users/IgnacioSepulveda/Documents/Ayudantias/econometrics")
## Checkeamos que efectivamente sea el directorio
getwd()
```

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

```r
## Librerias a usar

## En caso de que no esten instalados los paquetes, corre sin el # inicial
#install.packages(c("haven","tidyverse","car"))
library("haven")
library("tidyverse")
library("car")
```

Respuesta 1

1. Ramiro le envía a su jefe los resultados de una regresión de salarios en función de la variable regresora "educación" a partir de una muestra de 500 personas, reportándole que el estadístico t para el coeficiente de la variable era 0.75, mientras que el error estándar de este coeficiente era 0.2. Su jefe necesitaba contrastar si este coeficiente era estadísticamente distinto de 1 pero no pudo hacerlo porque no tenía los datos originales. Ramiro al otro día le envía la información, pero le advierte a su jefe que SI tenía los datos para poder hacer esta prueba de hipótesis. Explique cómo podría haberlo hecho y haga el test para la hipótesis nula $H_0 : \beta_1 = 1$ con un nivel de significación del 5 %.

Primero definimos las variables relevantes,

```r
muestra=500 #muestra de personas
estadistico_t=0.75 # estadistico t
error_estandar=0.2 # error estandar
```

Antes de hacer cualquier test necesitamos saber cual es nuestro $\hat{\beta}1$ estimado a partir de la muestra.

Sabemos que el estadístico t es igual a,

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

Cuando $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$, entonces basta con reemplazar y hacer el calculo.

$$0.75 = \frac{\beta_1}{0.2} \rightarrow \beta_1 = 0.75 * 0.2$$

Lo que nos da,

```
beta_1=0.75*0.2
beta_1
```

```
## [1] 0.15
```

Cuando es un test de dos colas, rechazamos cuando el valor absoluto del estadístico t es mayor al t critico.

Definimos las variables que nos faltan,

```
beta_jefe=1 # beta a testear
alpha=0.05 # significancia
```

Finalmente evaluamos.

```
# La funcion
t_calculado=(beta_1-beta_jefe)/error_estandar
t_critico=qt(1-alpha/2,muestra-2)
print(paste("t_critico",round(t_critico,3),",t_caculado",abs(t_calculado)))
```

```
## [1] "t_critico 1.965 ,t_caculado 4.25"
```

```
if (abs(t_calculado)>t_critico) 'Rechazamos' else 'No podemos Rechazar'
```

```
## [1] "Rechazamos"
```

Osea no existe evidencia en la muestra de que $\beta_1 = 1$.

Respuesta 2

2. Un investigador quiere analizar el efecto que tiene los años de educación "*educ*" en la cantidad de hijos nacidos vivos "*children*".

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

a) Basado en el fichero FERTIL2 estime el modelo propuesto y presente los resultados en forma de ecuación.

b) Asumiendo que se cumplen los supuestos del MRLC verifique la significancia individual de la variable "*educ*" mediante intervalos de confianza y un nivel de significancia del 5%.

c) Repita el análisis anterior mediante el p-valor.

d) ¿Existe evidencia para afirmar que β_1=3?, analice mediante zonas de rechazo e intervalos de confianza.

e) ¿Existe evidencia para afirmar que a mayor educación menor será la cantidad de hijos?

Respuesta a)

Primero cargamos los datos

```
rm(list=ls())
df=read_dta("FERTIL2.dta")
```

Estimamos los parámetros de la regresión.

```
regresion=lm(children~educ,data=df) ## Este comando solo nos muestra
                                    ## los parámetros estimados.
regresion
```

```
##
## Call:
## lm(formula = children ~ educ, data = df)
##
## Coefficients:
## (Intercept)         educ
##      3.4955      -0.2097
```

```
regresion %>% summary() ## Este nos muestra mas datos que nos pueden interesar
```

```
##
## Call:
## lm(formula = children ~ educ, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.495 -1.496 -0.399  1.182  9.505
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.49554    0.05612   62.28   <2e-16 ***
## educ        -0.20965    0.00796  -26.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.064 on 4359 degrees of freedom
## Multiple R-squared:  0.1373, Adjusted R-squared:  0.1371
## F-statistic: 693.7 on 1 and 4359 DF,  p-value: < 2.2e-16
```

Finalmente armamos la ecuación.

$$\widehat{Children} = 3.495 - 0.209 Educ$$

Respuesta b)

Necesitamos ahora los intervalos de confianza, usamos una función de R.

Rechazamos si el valor que buscamos no se encuentra en el intervalo.

```
significancia=0.05
confint(regresion,"educ",1-significancia)
```

```
##           2.5 %     97.5 %
## educ -0.2252564 -0.1940445
```

```
## Llegamos al mismo valor si lo calculamos a mano
beta_estimado=coef(regresion)[2] ## BETA ESTIMADO
sd=sqrt(diag(vcov(regresion)))[2] ## ERROR ESTANDAR
n=length(regresion$residuals) ## TAMAÑO DE LA MUESTRA
t_critico=qt(1-significancia/2,n-2) ## T estadistico


paste(significancia*100/2,"%: ",              # Limite
      round(beta_estimado-t_critico*sd,7),    # Inferior
      100-significancia*100/2,"%: ",          # Limite
      round(beta_estimado+sd*t_critico,7))    # Superior
```

```
## [1] "2.5 %:  -0.2252564 97.5 %:  -0.1940445"
```

Rechazamos dado que no se encuentra el cero, por lo tanto nuestra variable es significativa. Osea que existe evidencia de que nos ayuda a explicar la cantidad de hijos.

Respuesta c)

La primera forma de obtenerlo a través de la regresión.

```
summary(regresion)$coefficients[2,4] ## Esto nos permite rescatar el p-value desde
```

```
## [1] 5.413817e-142
```

Y otra forma es calcularlo con una función de R.

```
beta_estimado=coef(regresion)[2] ## Obtenemos el beta de la regresion
sd=sqrt(diag(vcov(regresion)))[2] ## El error estandar
estadistico_t=beta_estimado/sd ## El estadistico t
tamano_muestra=length(df$educ)
pt(estadistico_t,tamano_muestra-2)*2 ## El p-value
```

```
##          educ
## 5.413817e-142
```

Como el p-value es menor a 0.05 volvemos a confirmar que la nula se rechaza.

Respuesta d)

Ocupamos la funcion LinearHypothesis del paquete car para hacer test de dos colar y testeamos si $\beta_{educ} = 3$. Si el p-value es menor a 0.05 podemos rechazar la nula.

```
# La funcion
beta_test=3
linearHypothesis(regresion,paste("educ=",beta_test))
```

```
## Linear hypothesis test
##
## Hypothesis:
## educ = 3
##
## Model 1: restricted model
## Model 2: children ~ educ
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   4360 711264
## 2   4359  18572  1    692692 162583 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Y ocupamos la función intervalos de confianza creada.

```
significancia=0.05
confint(regresion,"educ",1-significancia)
```

```
##           2.5 %     97.5 %
## educ -0.2252564 -0.1940445
```

Que pasa si reemplazamos los limites del intervalo en la función linearHypothesis?

```
beta_test=-0.225256
linearHypothesis(regresion,paste("educ=",beta_test))
```

```
## Linear hypothesis test
##
## Hypothesis:
## educ = - 0.225256
##
## Model 1: restricted model
## Model 2: children ~ educ
##
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1   4360 18588
## 2   4359 18572  1    16.375 3.8434 0.05001 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Lo mismo dara si ocupamos el limite superior
```

Respuesta e)

Yo quiero encontrar evidencia para ver si el β_{educ} es negativo, osea que mientras mas educación menos hijos. Entonces para encontrar evidencia debo rechazar que es positivo.

$$H_0 : \beta_{educ} \geq 0 \, o \, \beta_{educ} = 0 \text{ vs } H_1 : \beta_{educ} < 0$$

Recordemos que la regla para cuando tenemos en la nula $\beta \geq 0$, es que el t calculado tiene que ser menor que -1 por el t critico.

$$t_{calculado} \leq -t^c$$

En ese caso podemos rechazar.

```
# La funcion
beta_test=0
t_calculado=(beta_estimado-beta_test)/sd
t_critico=qt(1-significancia/2,tamano_muestra-2)
if (t_calculado<-t_critico) 'Rechazamos' else 'No podemos Rechazar'
```

```
## [1] "Rechazamos"
```

Osea existe evidencia para afirmar que la educación afecta de forma negativa en la cantidad de hijos.

Extra: Gráficos en R.

Ocuparemos dos paquetes el primeros es el base y el segundo es ggplot. Y usaremos la misma base de datos de recién

```r
rm(list=ls())
## Librerias a usar
library(car)
library("haven")
library("tidyverse")
```

```r
## La base
df=read_dta("FERTIL2.dta")
```

```r
## Primero revisamos los datos
df %>% str()
```

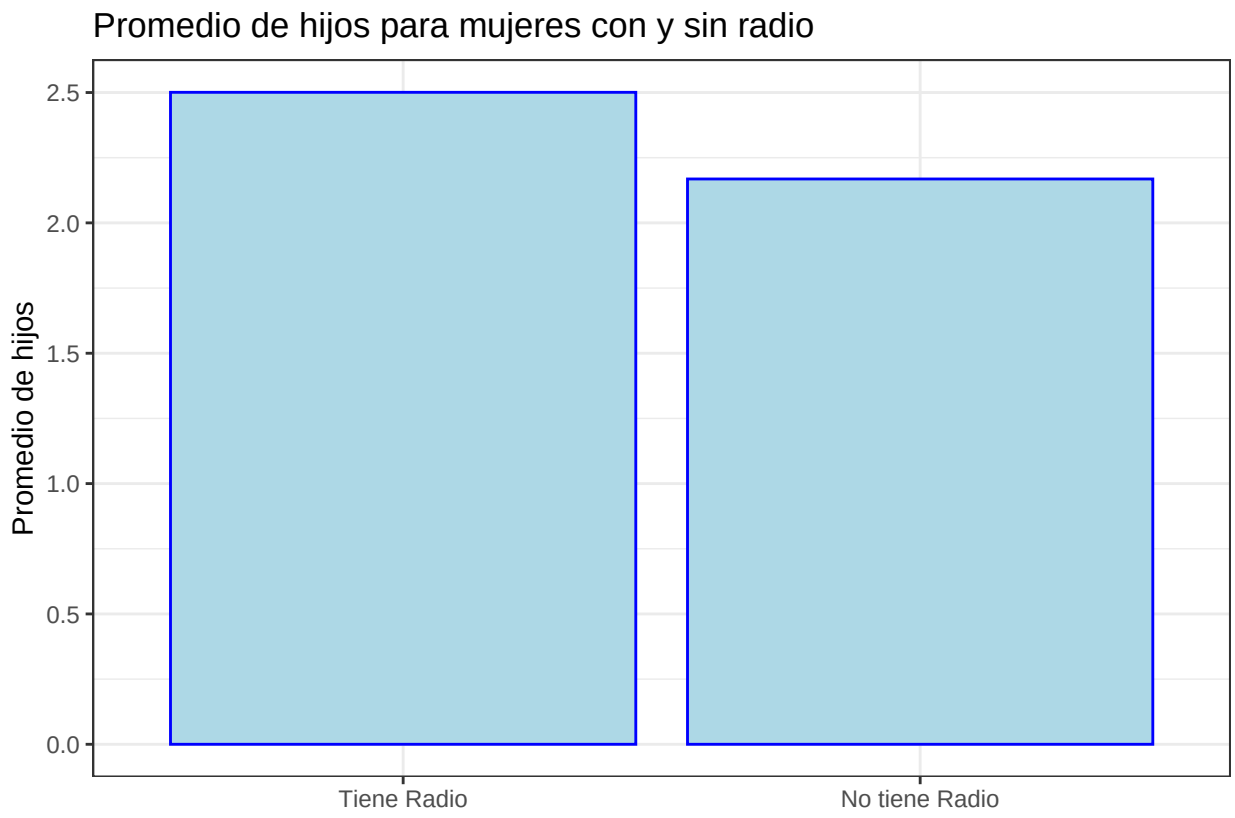
```
## tibble [4,361 x 27] (S3: tbl_df/tbl/data.frame)
##  $ mnthborn: num [1:4361] 5 1 7 11 5 8 7 9 12 9 ...
##   ..- attr(*, "label")= chr "month woman born"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ yearborn: num [1:4361] 64 56 58 45 45 52 51 70 53 39 ...
##   ..- attr(*, "label")= chr "year woman born"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ age     : num [1:4361] 24 32 30 42 43 36 37 18 34 49 ...
##   ..- attr(*, "label")= chr "age in years"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ electric: num [1:4361] 1 1 1 1 1 1 1 1 0 1 ...
##   ..- attr(*, "label")= chr "=1 if has electricity"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ radio   : num [1:4361] 1 1 0 0 1 0 1 1 1 1 ...
##   ..- attr(*, "label")= chr "=1 if has radio"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ tv      : num [1:4361] 1 1 0 1 1 0 1 1 0 0 ...
##   ..- attr(*, "label")= chr "=1 if has tv"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ bicycle : num [1:4361] 1 1 0 0 1 0 1 1 0 0 ...
##   ..- attr(*, "label")= chr "=1 if has bicycle"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ educ    : num [1:4361] 12 13 5 4 11 7 16 10 5 4 ...
##   ..- attr(*, "label")= chr "years of education"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ ceb     : num [1:4361] 0 3 1 3 2 1 4 0 1 0 ...
##   ..- attr(*, "label")= chr "children ever born"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ agefbrth: num [1:4361] NA 25 27 17 24 26 20 NA 19 NA ...
##   ..- attr(*, "label")= chr "age at first birth"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ children: num [1:4361] 0 3 1 2 2 1 4 0 1 0 ...
##   ..- attr(*, "label")= chr "number of living children"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ knowmeth: num [1:4361] 1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "=1 if know about birth control"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ usemeth : num [1:4361] 0 1 0 0 1 1 1 1 1 0 ...
##   ..- attr(*, "label")= chr "=1 if ever use birth control"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ monthfm : num [1:4361] NA 11 6 1 3 11 5 NA 7 11 ...
```

```
##   ..- attr(*, "label")= chr "month of first marriage"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ yearfm  : num [1:4361] NA 80 83 61 66 76 78 NA 72 61 ...
##   ..- attr(*, "label")= chr "year of first marriage"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ agefm   : num [1:4361] NA 24 24 15 20 24 26 NA 18 22 ...
##   ..- attr(*, "label")= chr "age at first marriage"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ idlnchld: num [1:4361] 2 3 5 3 2 4 4 4 4 4 ...
##   ..- attr(*, "label")= chr "'ideal' number of children"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ heduc   : num [1:4361] NA 12 7 11 14 9 17 NA 3 1 ...
##   ..- attr(*, "label")= chr "husband's years of education"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ agesq   : num [1:4361] 576 1024 900 1764 1849 ...
##   ..- attr(*, "label")= chr "age^2"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ urban   : num [1:4361] 1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "=1 if live in urban area"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ urb_educ: num [1:4361] 12 13 5 4 11 7 16 10 5 4 ...
##   ..- attr(*, "label")= chr "urban*educ"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ spirit  : num [1:4361] 0 0 1 0 0 0 0 0 0 1 ...
##   ..- attr(*, "label")= chr "=1 if religion == spirit"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ protest : num [1:4361] 0 0 0 0 1 0 0 0 1 0 ...
##   ..- attr(*, "label")= chr "=1 if religion == protestant"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ catholic: num [1:4361] 0 0 0 0 0 0 1 1 0 0 ...
##   ..- attr(*, "label")= chr "=1 if religion == catholic"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ frsthalf: num [1:4361] 1 1 0 0 1 0 0 0 0 0 ...
##   ..- attr(*, "label")= chr "=1 if mnthborn <= 6"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ educ0   : num [1:4361] 0 0 0 0 0 0 0 0 0 0 ...
##   ..- attr(*, "label")= chr "=1 if educ == 0"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ evermarr: num [1:4361] 0 1 1 1 1 1 1 0 1 1 ...
##   ..- attr(*, "label")= chr "=1 if ever married"
##   ..- attr(*, "format.stata")= chr "%9.0g"
```

Gráficos de Barras.

Gráfiquemos un gráficos de barras que contendrá el promedio de hijos para mujeres con y sin radio.

```
df %>% ## Los datos

  select(radio,children) %>% ## Selecciona variables a usar

  na.omit() %>% ## Elimina los nas

  mutate(radio=factor(radio,labels=c('Tiene Radio','No tiene Radio'))) %>%
  ## Hace que la variable raido sea categorica

  group_by(radio) %>% ## Agrupa por los que tienen y no tiene radio.

  summarise(hijos=mean(children)) %>% ## Calcula la media de hijos para c/ grupo.

  ggplot(aes(x=radio,y=hijos)) + ## Marco general donde estara el grafico

  geom_bar(stat='identity',fill='lightblue',color='blue') + ## La figura que se

  ggtitle("Promedio de hijos para mujeres con y sin radio") + ## Titulo

  xlab("") + ## Dejamos la x-axis vacío dado que ya tiene info

  ylab("Promedio de hijos")+

  theme_bw() ## El fondo del grafico
```

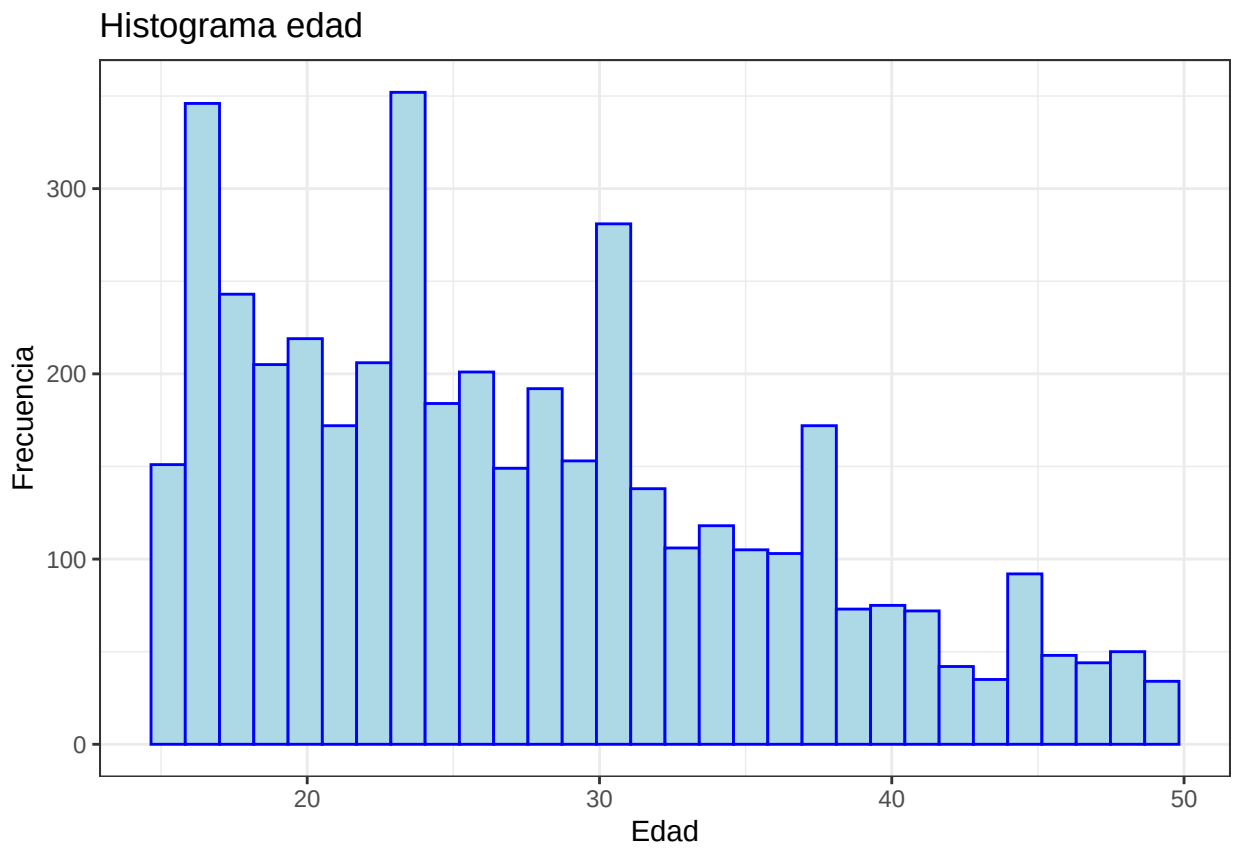


Histogramas

Haremos histogramas para la distribución de la edad de mujeres.

```
df %>% ## Los datos

  ggplot(aes(x=age)) + ## El marco general

  geom_histogram(fill='lightblue',color='blue') + ## Histogram

  ggtitle("Histograma edad") + ## Titulo

  xlab('Edad') + ## x-axis

  ylab('Frecuencia') +## y-axis

  theme_bw() ## El fondo del grafico
```

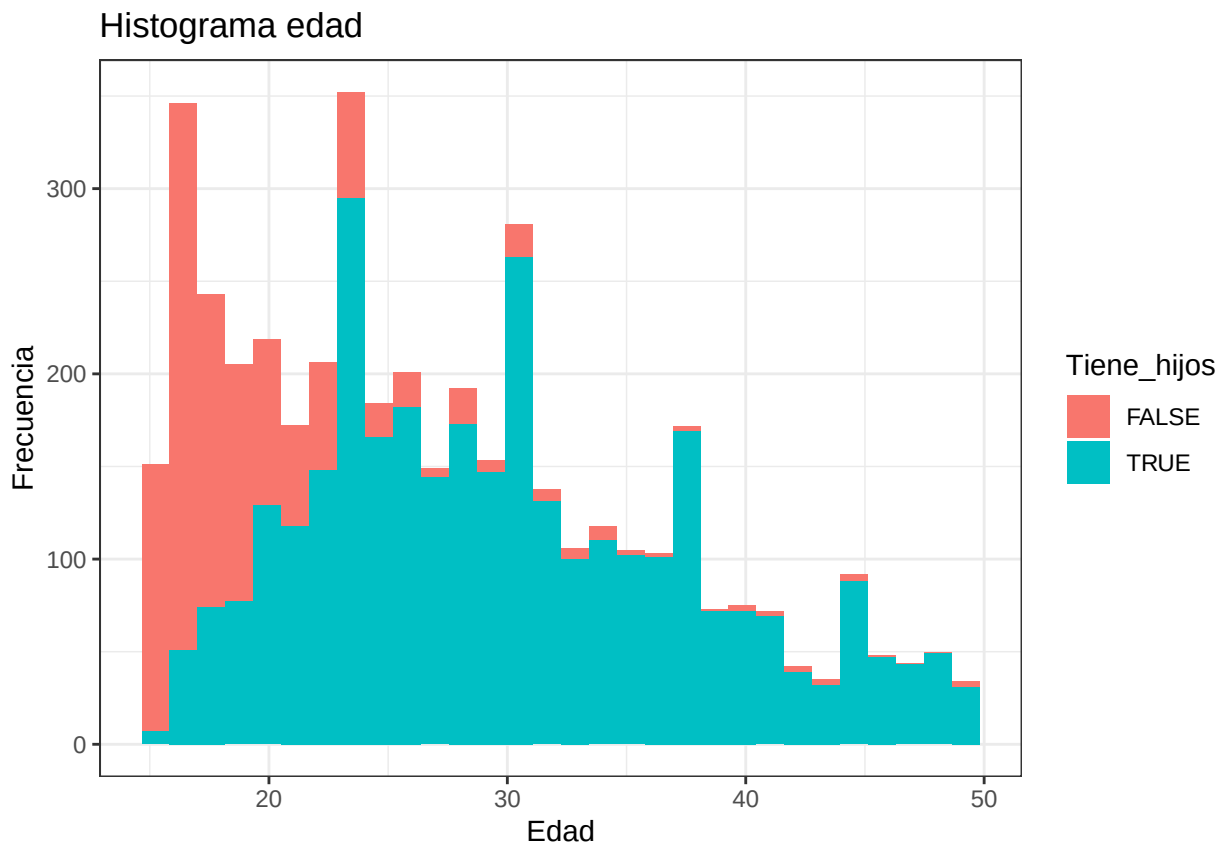
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Vemos ahora la diferencia entre las mujeres que tiene y no tiene hijos.

```
df %>% ## Los datos

  mutate(Tiene_hijos=children>0) %>%  ## Creamos una variable que sera V si tiene
                                      ## hijos y F si no tiene hijos.
  ggplot(aes(x=age,fill=Tiene_hijos)) + ## El marco general

  geom_histogram() + ## Histogram

  ggtitle("Histograma edad") + ## Titulo

  xlab('Edad') + ## x-axis

  ylab('Frecuencia') +## y-axis

  theme_bw() ## El fondo del grafico
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

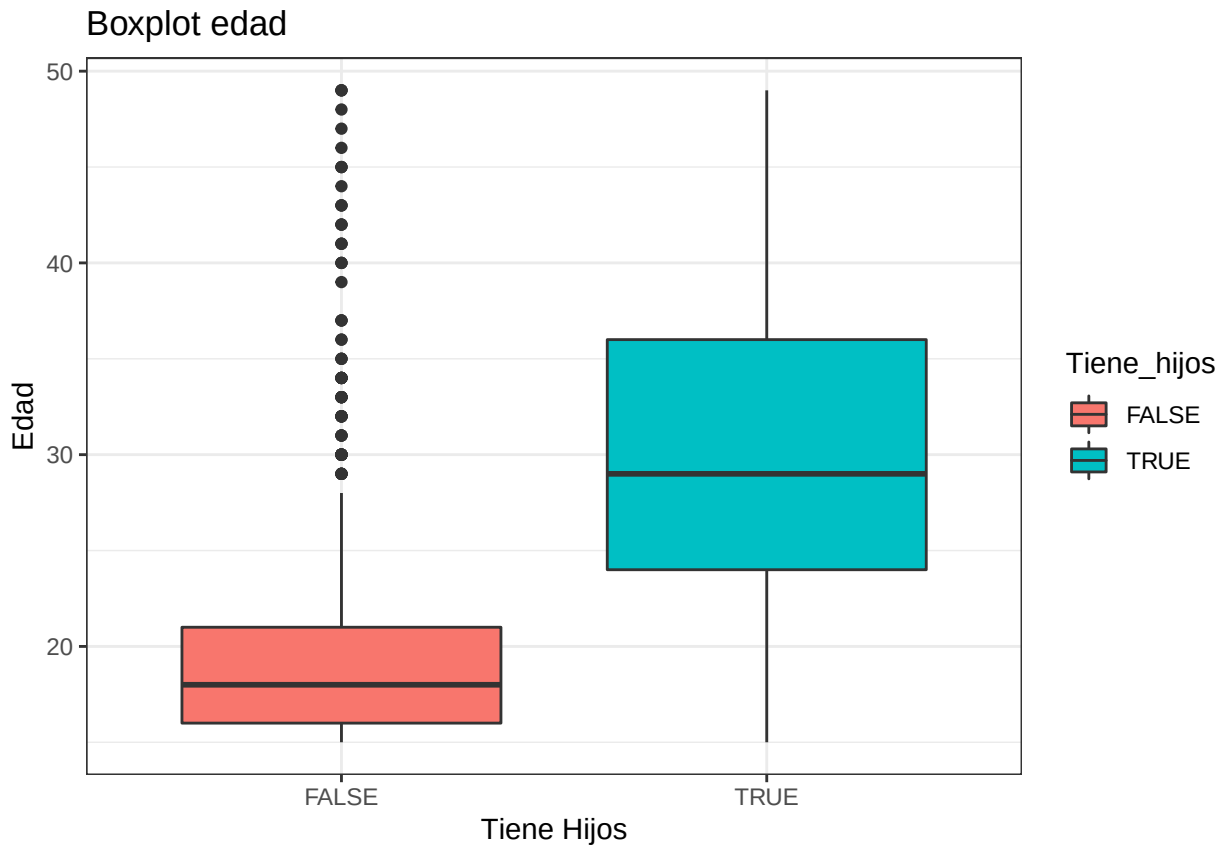


El tema es que no se ve muy bien.

Gráficos de Caja.

El gráfico de caja puede ayudar en parte a ver mejor esas diferencias.

```
df %>% ## Los datos

  mutate(Tiene_hijos=children>0) %>%  ## Creamos una variable que sera V si tiene
                                      ## hijos y F si no tiene hijos.

  ggplot(aes(x=Tiene_hijos,y=age,fill=Tiene_hijos)) + ## El marco general

  geom_boxplot() + ## Histogram

  ggtitle("Boxplot edad") + ## Boxplot

  xlab('Tiene Hijos') + ## x-axis

  ylab('Edad') + ## y-axis

  theme_bw() ## El fondo del grafico
```

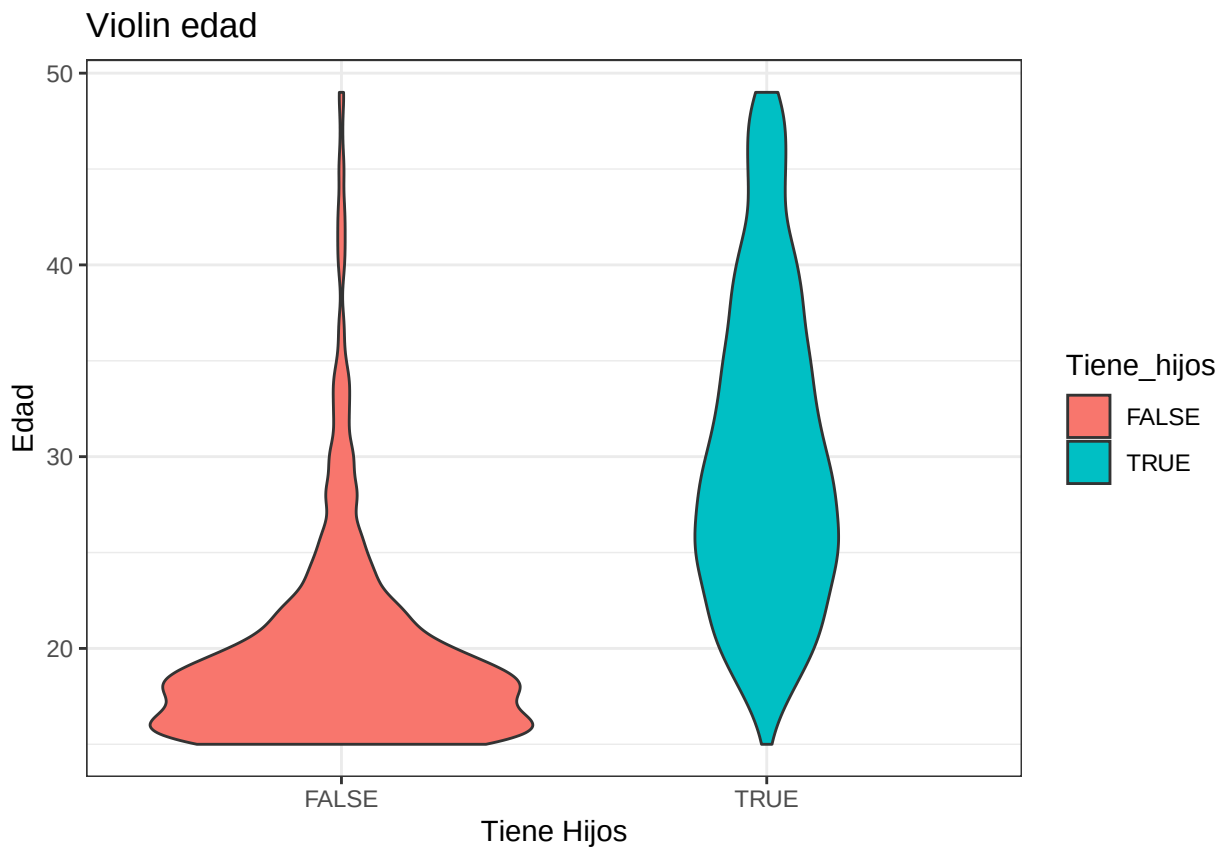


El tema es que no tenemos claro que esta pasando en las orillas de la distribución. Para eso una mejor idea es el gráfico de violín.

Gráficos de Violín.

En este gráfico queda mucho mas claro donde se concentran los datos.

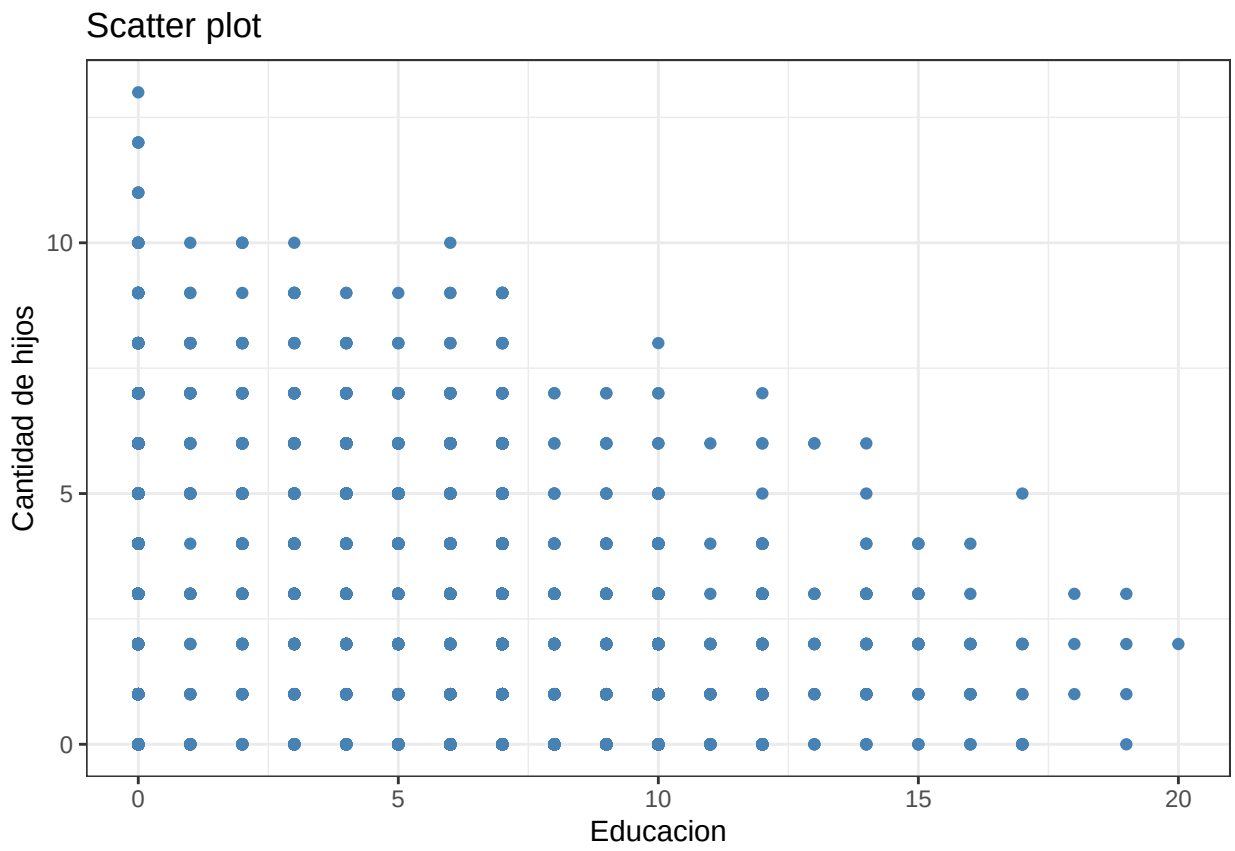
```
df %>% ## Los datos

  mutate(Tiene_hijos=children>0) %>%  ## Creamos una variable que sera V si tiene
                                      ## hijos y F si no tiene hijos.

  ggplot(aes(x=Tiene_hijos,y=age,fill=Tiene_hijos)) + ## El marco general

  geom_violin() + ## Violin

  ggtitle("Violin edad") + ## Titulo

  xlab('Tiene Hijos') + ## x-axis

  ylab('Edad') +## y-axis

  theme_bw() ## El fondo del grafico
```



Scatter plot.

Veamos como se ve la relación que modelamos en la pregunta dos. Este es el gráfico clasico para evaluar dos variables continuas.

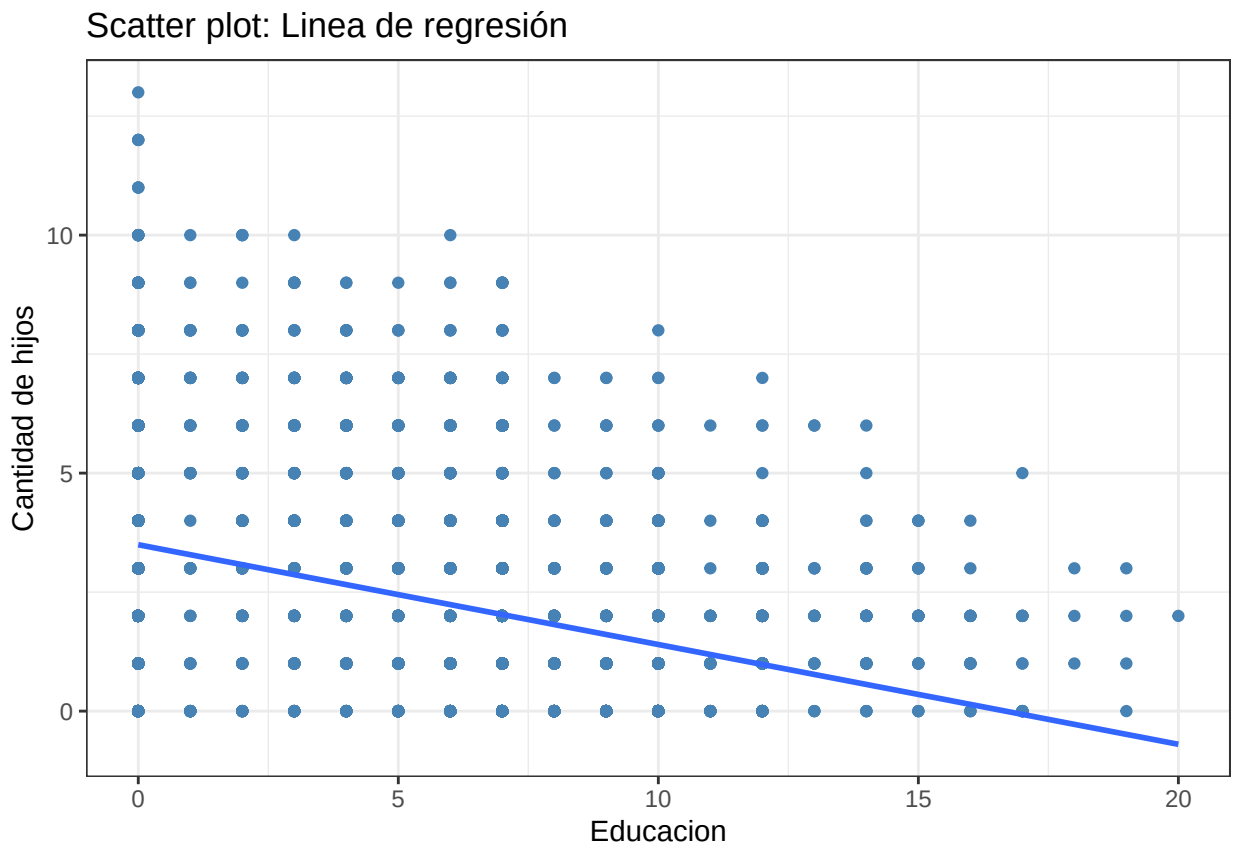
```
df %>% ## Datos

  ggplot(aes(x=educ,y=children))+ ## Marco general

  geom_point(color='steelblue') + ## Puntos

  xlab("Educacion") + ## x-axis

  ylab("Cantidad de hijos") +

  ggtitle("Scatter plot") +

  theme_bw() ## El fondo del grafico
```



Scatter plot : Incluyendo la regresión.

```
df %>% ## Datos

  ggplot(aes(x=educ,y=children))+ ## Marco general

  geom_point(color='steelblue') + ## Puntos

  geom_smooth(method='lm',se=FALSE)+ ## Regresion

  xlab("Educacion") + ## x-axis

  ylab("Cantidad de hijos") + ## y-axis

  ggtitle("Scatter plot: Linea de regresión") +

  theme_bw() ## El fondo del grafico
```

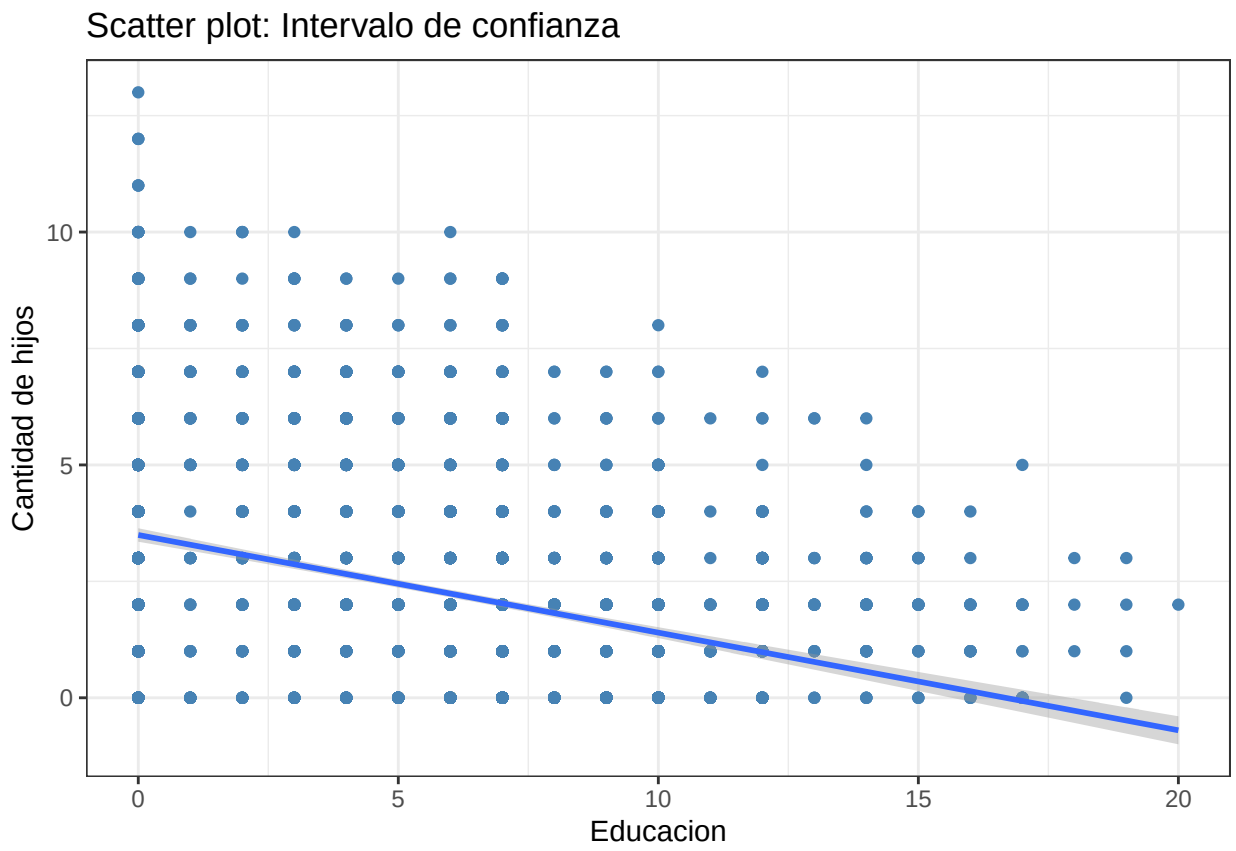
```
## `geom_smooth()` using formula 'y ~ x'
```



Scatter plot: Incluyendo intervalos de confianza.

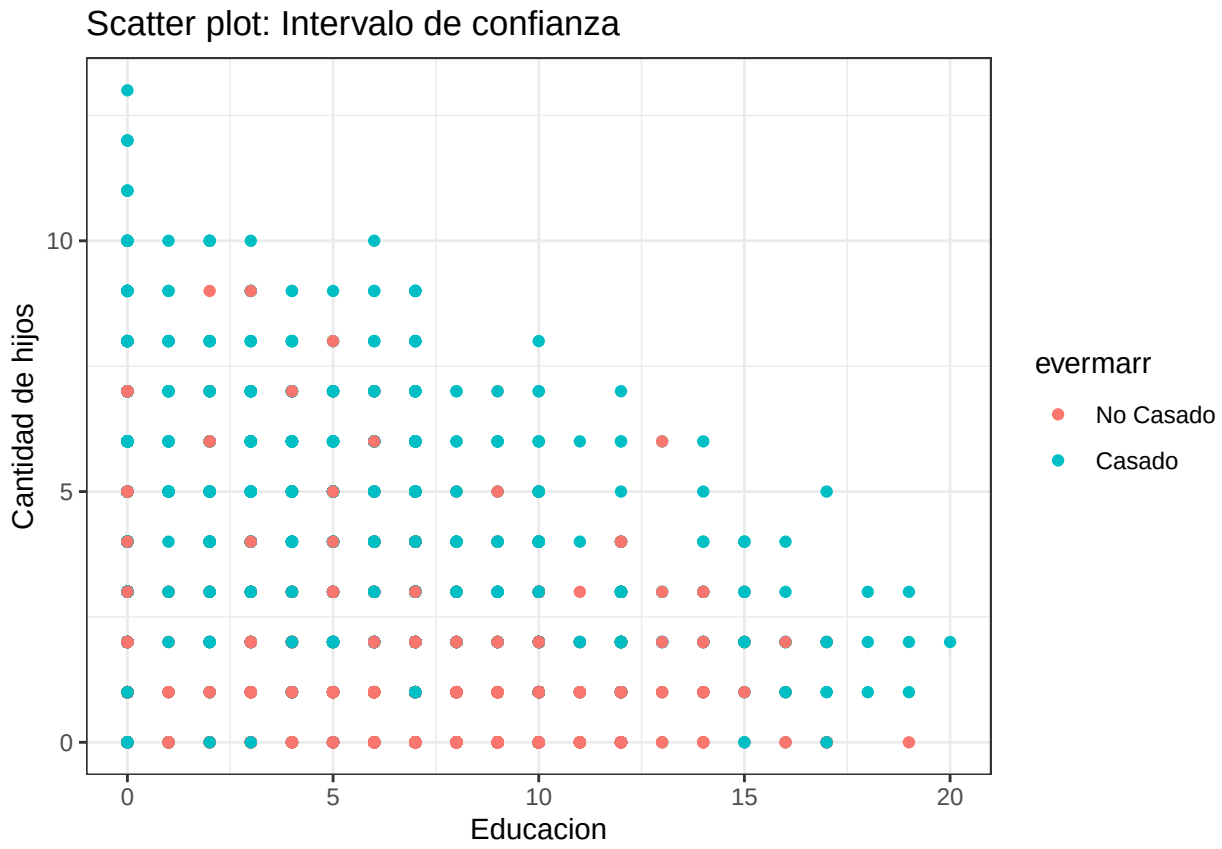
```
df %>% ## Datos

  ggplot(aes(x=educ,y=children))+ ## Marco general

  geom_point(color='steelblue') + ## Puntos

  geom_smooth(method='lm',se=TRUE,level=.99)+ ## Regresion

  xlab("Educacion") + ## x-axis

  ylab("Cantidad de hijos") + ## y-axis

  ggtitle("Scatter plot: Intervalo de confianza") +

  theme_bw() ## El fondo del grafico
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Scatter plot: Una tercera variable.

```
df %>% ## Datos

  mutate(evermarr=factor(evermarr,
                         labels=c('No Casado','Casado'))) %>%
  ## Convertimos a categoría la variable tv

  ggplot(aes(x=educ,y=children,color=evermarr))+ ## Marco general

  geom_point() + ## Puntos

  #geom_smooth(method='lm',se=TRUE,level=.95)+ ## Regresion

  xlab("Educacion") + ## x-axis

  ylab("Cantidad de hijos") + ## y-axis

  ggtitle("Scatter plot: Intervalo de confianza") +

  theme_bw() ## El fondo del grafico
```


Scatter plot: Intervalo de confianza
Cantidad de hijos
10
5
0
0
5
10
15
20
Educacion
evermarr
No Casado
Casado