```
%load_ext rpy2.ipython
```

Tarea #2: Máxima verosimilitud, Truncamiento y Censura

EC-4300 Microeconometría

Estudiantes: B75115 y B76137

Primera parte: Máxima verosimilitud

1. Cargue la base y cree las variables necesarias para empezar a trabajar.

1.a Se instalan los paquetes necesarios.

```
%%R
install.packages(c("readxl","dplyr","ggplot2", "plotly","tidyr","scales","aod"))
```

```
     UsageError: Cell magic `%%R` not found.
```

1.b Se cargan los paquetes necesarios.

```
%%R
library(readxl)
library(dplyr)
library(ggplot2)
library(plotly)
library(tidyr)
library(scales)
library(aod)
```

1.c Se carga la base de datos y se crean las variables necesarias.

```
%%R
Cobb_Douglas<- read_excel("C:/Users/David Mora Salazar/Documents/ECONOMÍA UNIVERSIDAD DE COST
Cobb_Douglas<-as.matrix(Cobb_Douglas[,2:4]) # Convertimos a matriz y quitamos la primera colu
Insumo_1<-Cobb_Douglas[,1]  # Insumo_1 está en la primera columna
Insumo_2<-Cobb_Douglas[,2]  # Insumo_2 en la segunda
Producto<-Cobb_Douglas[,3]  # Producto en la tercera
```

2. Calcule la función de verosimilitud y optimícela. ¿Por qué se está utilizando una función de verosimilitud?,¿Cuál es el propósito de encontrar la máxima verosimilitud de la función? ¿Cuál es el valor de máxima verosimilitud?

2.a Se crea la función de verosimilitud.

```
mv_fprod<-function(alfa,Z1,Z2,Q) {
  n<-length(Z1)  #NÃºmero de observaciones
  e<-Q-alfa[1]*(Z1^alfa[2])*(Z2^alfa[3])   # Forma funcional de los residuos
  logl<- -0.5*n*log(2*pi)-0.5*n*log(alfa[4])-((t(e)%*%e)/(2*alfa[4])) # FunciÃ³n de verosimil
  return(-logl)  # FunciÃ³n da el negativo de la verosimilitud. Necesario para funciÃ³n de op
}
```

2.b Se optmiza la función de verosimilitud.

```
resultado<-optim(c(1,0.5,0.5,0.01),mv_fprod,method="BFGS",hessian=T,Z1=Insumo_1,Z2=Insumo_2,Q
```

2.c Se muestra el hessiano asociado y se averigua el valor de máxima verosimilitud.

```
resultado  #Muestra todos los resultados del hessiano
resultado$value  # Valor resultante para la máxima verosimilitud
paste("La maxima verosimilitud es:", resultado$value)
```

El valor de la máxima verosimilitud es 17.6542220632782.

Se está usando la función de verosimilitud para obtener la medida de la densidad conjunta de un conjunto de datos, porque esta es una función de los parámetros que permite realizar inferencias acerca del valor de los parámetros a partir de un conjunto de observaciones. Además, la función de verosimilitud se acopla para inferir bien con parámetros no lineales del modelo.

El propósito de encontrar la máxima verosimilitud es encontrar un conjunto de parámetros que maximice la densidad de la variable dependiente, es decir, que maximiza la probabilidad de obtener los datos que tenemos.

3. Con respecto a la optimización del punto 2, muestre los resultados para cada uno de los parámetros de una forma ordenada y clara.

3.a Se ordenan los valores estimados de los parámetros en una tabla.

```
resultado$par
cuadro <- matrix(resultado$par,ncol=4,byrow=FALSE)
colnames(cuadro) <- c("alfa[1]","alfa[2]","alfa[3]","alfa[4] o varianza")
cuadro <- as.table(cuadro)
cuadro
```

```
> cuadro <- matrix(resultado$par,ncol=4,byrow=FALSE)
> colnames(cuadro) <- c("alfa[1]","alfa[2]","alfa[3]","alfa[4] o varianza")
> cuadro <- as.table(cuadro)
> cuadro
    alfa[1]   alfa[2]   alfa[3] alfa[4] o varianza
A 0.5154801 0.3256620 0.8607203          0.2184277
>
```

4. Se quiere conocer si la función de producción presenta rendimientos constantes, realice una prueba de Wald para determinarlo. ¿En qué consiste la prueba? ¿En qué caso es recomendable utilizar la prueba de Wald? ¿Cuál es su hipótesis nula? ¿Cuál es la hipótesis alternativa? ¿Se rechaza o no la hipótesis nula?

4.a Se realiza la prueba de Wald.

```
R<-cbind(0,1,1,0) # Vector con coeficientes de la restricciÃ³n:  debe tener tamaÃ±o del vecto
V<-solve(resultado$hessian) # Matriz de informaciÃ³n
prueba_wald<-wald.test(b = resultado$par,Sigma =V,L=R,H0=1)
prueba_wald$result
```

```
> prueba_wald<-wald.test(b = resultado$par,Sigma =V,L=R,H0=1)  # H0 es que Rb = 1
> prueba_wald$result
$chi2
    chi2       df        P
195.6993   1.0000   0.0000

>
```

4.b La prueba consiste en testear la hipótesis nula de que un grupo o canasta de parámetros de la función (en este caso, de producción) sea igual a un valor o valores en específico.

4.c La prueba es recomendable cuando se desea estimar el modelo sin restricciones (Likelihood restringido), porque Wald no necesita imponer la restricción para obtener el estadístico y además, es recomendable cuando se desea realizar cálculos de hipótesis más sencillas o no muy complejas.

4.d La hipótesis nula es que alfa2+alfa3=1, donde alfa2 y alfa3 representan los coeficientes que acompañan a una función de producción Cobb Douglas.

4.e La hipótesis alternativa es que alfa2+alfa3 ≠ 1, donde alfa2 y alfa3 representan los coeficientes que acompañan a una función de producción Cobb Douglas.

4.f La hipótesis nula se rechaza con un 5% de significancia que afirma que la suma de los coeficientes son iguales a 1.

5. Realice una prueba de razón de verosimilitud. ¿En qué consiste la prueba? ¿En qué caso es recomendable utilizar la prueba de razón de verosimilitud? ¿Cuál es su hipótesis nula? ¿Cuál es la hipótesis alternativa? ¿Se rechaza o no la hipótesis nula?

5.a Se crea la función de verosimilitud restringida.

```
mv_fprod_restr<-function(alfa,Z1,Z2,Q) {

  n<-length(Z1)  #NÃºmero de observaciones
  e<-Q-alfa[1]*(Z1^alfa[2])*(Z2^(1-alfa[2]))   # Se impone restricciÃ³n de que exponentes sum
  logl<- -0.5*n*log(2*pi)-0.5*n*log(alfa[4])-((t(e)%*%e)/(2*alfa[4])) # FunciÃ³n de verosimil
  return(-logl)  # FunciÃ³n da el negativo de la verosimilitud
}
```

5.b Se optmiza la función de verosimilitud restringida.

```
resultado_restr<-optim(c(1,0.5,0.5,0.01),mv_fprod_restr,method="BFGS",hessian=T,Z1=Insumo_1,Z
```

5.c Se deduce la razón de verosimilitud, se calcula el valor de su estadístico y del P-Value.

```
LRestad<- 2*(resultado$value - resultado_restr$value)*-1
pvalueLR<-1-pchisq(LRestad,1)  # Pvalue de Ji2 con 1 gl que se usa para comparar la razón de
LRestad
pvalueLR
valores_rv <- c(LRestad,pvalueLR)
valores_rv
```

5.d Se ordenan los datos en una tabla.

```
cuadro2 <- matrix(valores_rv,ncol=2,byrow=FALSE)
```

```
colnames(cuadro2) <- c("Razon de verosimilitud","Pvalue")
rownames(cuadro2) <- c("")
cuadro2 <- as.table(cuadro2)
cuadro2
```

```
> cuadro2
 Razon de verosimilitud       Pvalue
           5.071084e+01 1.070255e-12
```

5.e La prueba consiste en la lógica de que si la restricción es válida, su imposición no debería conducir a una gran reducción en la función de probabilidad logarítmica, por tanto, se comparan los dos modelos (restringido y sin restricción) para buscar si existe una diferencia significativa en su ajuste y si se necesita o no remover variables predictivas. En el caso en que la diferencia es significativa se puede pensar que el modelo no restringido calza los datos mejor que el modelo restringido.

5.f Es recomendable cuando se desean hacer estimaciones más complejas, o de un mayor tamaño y se necesita imponer la restricción a los parámetros. Y cuando se tiene la información de la likelihood restringida como no restringida.

5.g La hipótesis nula es que la razón de verosimilitud sea igual a 1.

5.h La hipótesis nula es que la razón de verosimilitud sea estrictamente menor a 1.

5.i Se rechaza la hipótesis nula.

Segunda parte: Truncamiento

1. Se pretende estudiar el comportamiento de la variable hours, que representa la cantidad de horas trabajadas, con respecto a distintas variables que forman parte de la base de datos. El método predilecto para trabajar son usualmente los Mínimos Cuadrados Ordinarios. Corra una regresión de MCO donde hours sea la variable dependiente. Tome los siguientes regresores: educación, experiencia, experiencia al cuadrado, edad, hijos menores a 6 años e hijos entre 6 y 18 años. Son en total seis variables independientes.

```
Mroz <- load("C:/Users/David Mora Salazar/Documents/ECONOMÍA UNIVERSIDAD DE COSTA RICA/Microe
Modelo<- lm(formula=hours~educ+exper+expersq+age+kidslt6+kidsge6,data=data)
```

```
summary(Modelo)
```

```
Call:
lm(formula = hours ~ educ + exper + expersq + age + kidslt6 +
    kidsge6, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1520.3  -536.1  -144.4   533.1  3546.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1367.3907   269.5616   5.073 4.96e-07 ***
educ          23.1997    12.2939   1.887   0.0595 .
exper         66.6625     9.9417   6.705 3.97e-11 ***
expersq       -0.6931     0.3247  -2.135   0.0331 *
age          -31.6120     4.2900  -7.369 4.58e-13 ***
kidslt6     -445.8039    58.8156  -7.580 1.03e-13 ***
kidsge6      -34.3349    23.1607  -1.482   0.1386
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 750.6 on 746 degrees of freedom
Multiple R-squared:  0.2638,    Adjusted R-squared:  0.2579
F-statistic: 44.56 on 6 and 746 DF,  p-value: < 2.2e-16
```

2. Con respecto a la regresión anterior, ¿Cuáles coeficientes son significativos?, ¿Cuáles son no significativos?, ¿Cómo se interpreta cada coeficiente?

Utilizando un nivel de significancia del 95%, los coeficientes de las variables: experiencia ,experiencia al cuadrado, edad, y el número de niños menores a 6 años, son significativos. Mientras que los coeficientes de las variables: educación y el número de niños entre 18 y 6 años incluido, no son significativos.

La interpretación de los coeficientes significativos es que sí existe evidencia estadística de una relación entre las variables de esos coeficientes significativos y las horas trabajadas, es decir que cada variable es significativa para explicar las horas trabajadas.

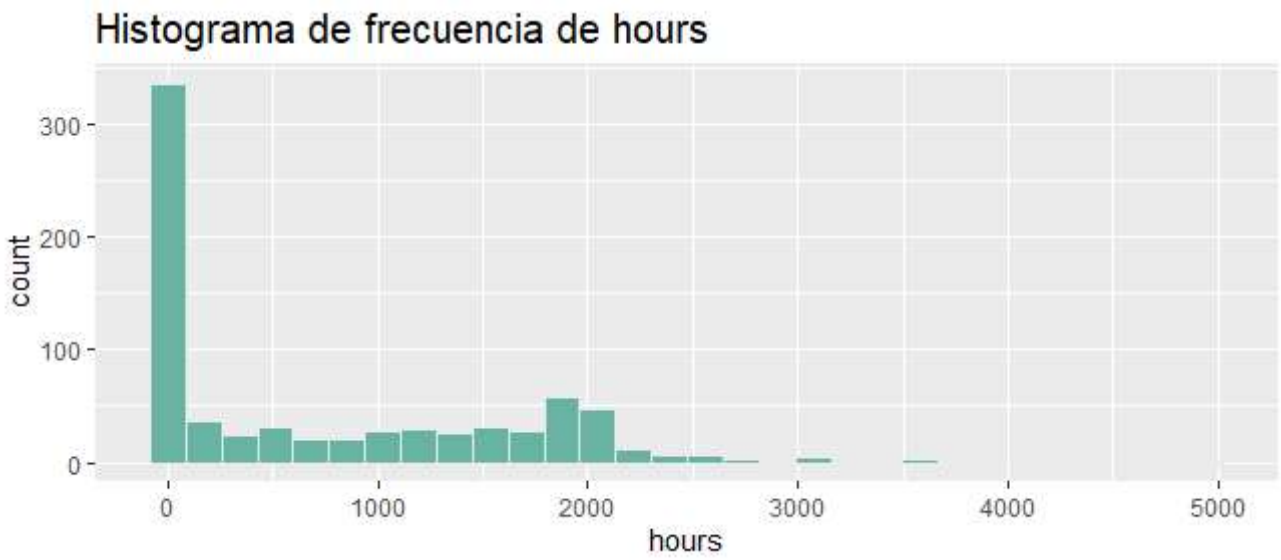
La interpretación de los coeficientes no significativos es que no existe evidencia estadística de una relación entre las variables de esos coeficientes no significativos y las horas trabajadas, es decir que cada variable no es significativa para explicar las horas trabajadas.

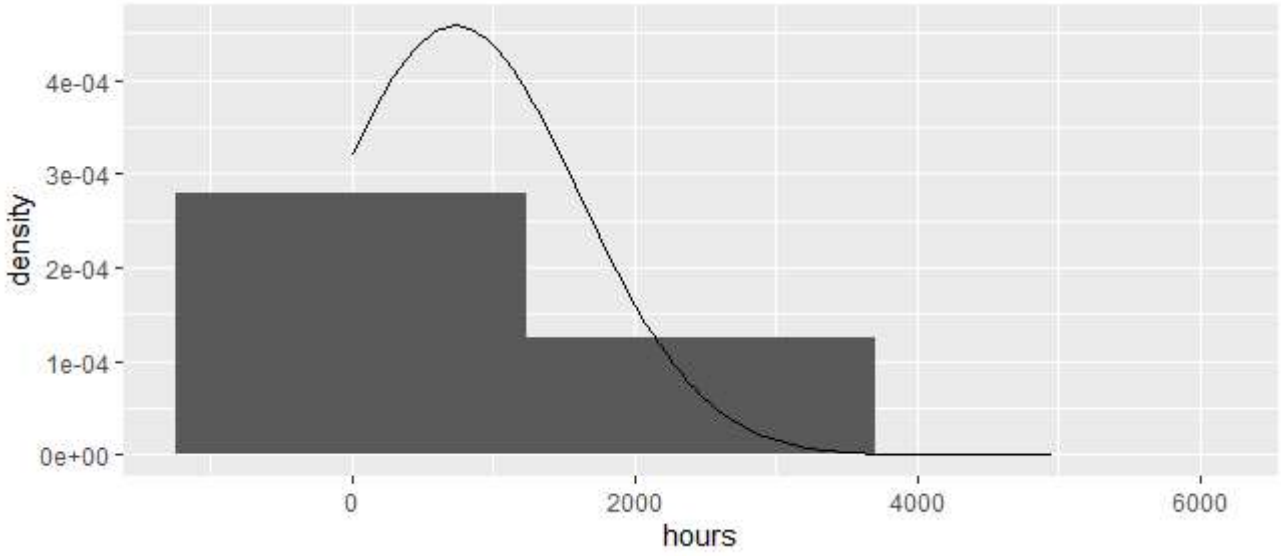
Para cada coeficiente:

1. Educación: No existe evidencia estadística de una relación entre la educación y las horas trabajadas.
2. Experiencia: Si existe evidencia estadística de una relación entre la experiencia y las horas trabajadas. Por cada unidad adicional de experiencia, se puede esperar que las horas trabajadas aumenten un promedio de 66.6625.
3. Experiencia al cuadrado: Si existe evidencia estadística de una relación entre la experiencia al cuadrado y las horas trabajadas, un unidad adicional de experiencia se asocia en promedio a una disminución de 0.69 horas trabajadas. Por lo que refleja una curvatura de la variable experiencia.
4. Edad: Si existe evidencia estadística de una relación entre la edad y las horas trabajadas. Por cada unidad adicional de edad, se puede esperar que las horas trabajadas se reduzcan un promedio de 31.6120.
5. Número de niños menores de 6 años: Si existe evidencia estadística de una relación entre el número de niños menores de 6 años y las horas trabajadas. Por cada unidad adicional de niños menores de 6 años, se puede esperar que las horas trabajadas se reduzcan un promedio de 445.8039.
6. Número de niños entre 6 años y 18 años: No existe evidencia estadística de una relación entre el número de niños entre 6 años y 18 años, y las horas trabajadas.

3. La variable hours muestra un comportamiento peculiar. ¿Qué está pasando con ella? Investigue la variable, grafíquela en un histograma y comente lo que observa.

```
#3.a Histograma de frecuencias
ggplot(data, aes(x=hours)) + geom_histogram( fill="#69b3a2", color="#e9ecef")+
  ggtitle("Histograma de frecuencia de hours")  +
  theme(
    plot.title = element_text(size=15)
  )
#3.b Histograma de densidad
ggplot(data=data,aes(hours))+geom_histogram(aes(y = ..density..),bins=3)+stat_function(fun =
```





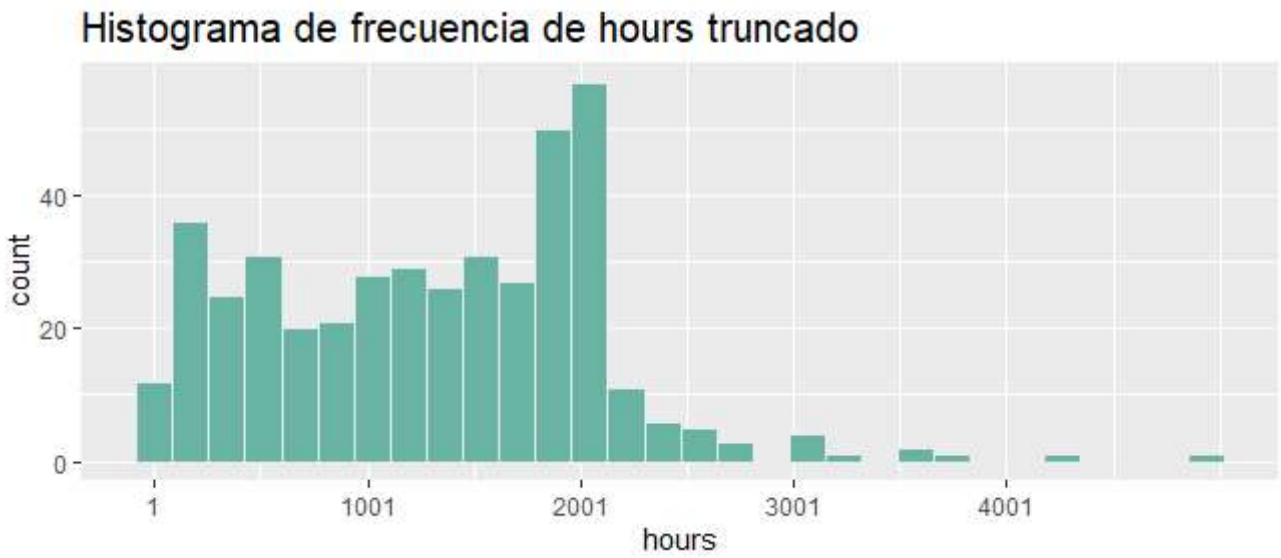
Se observa en el histograma de frecuencia que existe una importante cantidad de observaciones en donde la variable horas trabajadas es igual a cero. En el histograma de densidad, parece que la media se encuentra más cerca de ser menor a 1000 que mayor.

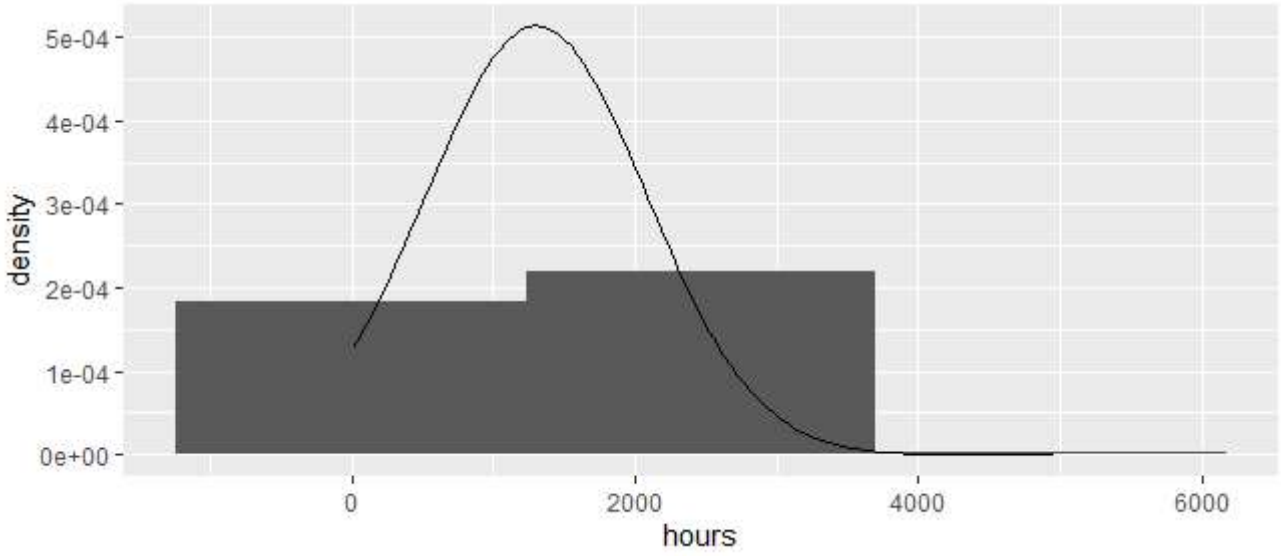
4. La variable hours está creando ruido en la base, esto significa que la gran cantidad de ceros que presenta distorsiona los datos. Realice otro histograma sin dichos ceros. ¿Qué cambios se observan en este nuevo grafico? ¿Es una buena idea trabajar sin este valor mínimo?

```
data_trunc<- data %>% filter(`hours`>0)
ggplot(data_trunc, aes(x=hours)) + geom_histogram( fill="#69b3a2", color="#e9ecef") +
  scale_x_continuous(breaks = seq(1, 5000, by = 1000))+ ggtitle("Histograma de frecuencia de
  theme(
    plot.title = element_text(size=15)
```

```
)

#4.b Histograma de densidad
ggplot(data=data_trunc,aes(hours))+geom_histogram(aes(y = ..density..),bins=3)+stat_function(
```





Con el nuevo histograma de densidad truncado por debajo en cero, parece ser que la media ahora aumenta o se acerca ahora más a las 2000 horas. Parece que la variabilidad también se reduce.

Creemos que si es una buena idea trabajar sin el valor mínimo de cero, pues para efectos de la investigación, se quiere que hayan valores positivos de "hours", pues deseamos ver a un subgrupo de la población que si tenga horas trabajadas para analizar los efectos de otras variables sobre esas horas trabajadas, claro está que se debe adaptar un modelo para datos truncados porque se debe reconocer que la muestra truncada no es una muestra aleatoria de la población.

5. Queda claro que trabajar mediante MCO no es la mejor idea. Según las características de la variable dependiente un truncamiento seria el mejor camino a tomar. Corra exactamente la misma regresión del punto 1, pero en esta aplique un truncamiento en hours >0.

```
Modelo_trunc<-npsf::truncreg(formula=hours~educ+exper+expersq+age+kidslt6+kidsge6,data=data,l
```

```
Truncated regression for cross-sectional data
Limits:
lower limit for left-truncation  = 0
upper limit for right-truncation = Inf
Number of observations (used in regression)              = 428
Number of truncated observations (not used in regression) = 325

Estimation results:

             Estimate Std. Error       z  Pr(>|z|)
(Intercept) 2122.06922  480.81234  4.4135 1.017e-05 ***
educ         -29.65101   21.80783 -1.3596 0.1739409
exper         72.60832   21.23109  3.4199 0.0006264 ***
expersq       -0.94513    0.60787 -1.5548 0.1199870
age          -27.39191    8.10640 -3.3790 0.0007274 ***
kidslt6     -484.85903  153.70929 -3.1544 0.0016083 **
kidsge6     -102.59511   43.49326 -2.3589 0.0183305 *
/sigma       850.77423   43.80149 19.4234 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> View(Modelo_trunc)
```

6. Con respecto a la regresión truncada del punto anterior, ¿Cuáles coeficientes son significativos?, ¿Cuáles son no significativos?, ¿Cambian los coeficientes o su significancia con respecto a la regresión MCO? ¿Por qué se podrían estar dando estos cambios? Comente ampliamente.

Utilizando un nivel de significancia del 95%, los coeficientes de las variables: experiencia, edad, el número de niños menores a 6 años y el número de niños entre 18 y 6 años incluido, son significativos. Mientras que los coeficientes de las variables: educación y experiencia al cuadrado no son significativos.

Tanto los valores de los coeficientes como sus niveles de significancia cambian. Para los valores de los coeficientes, todos cambian, unos con en mayor magnitud que otros, uno inclusive cambia de signo. A nivel de significancia, la variable de número de niños entre 6 años y 18 años pasa a ser significativa y la experiencia al cuadrado pasa a ser no significativa para una significancia de 95%.

Estos cambios se podrían estar dando porque ya no se está usando un modelo de regresión lineal, sino un modelo que se ajuste a los efectos del truncamiento. Recuerde que el modelo MCO es el proceso de estimación más eficiente solo cuando se observa una muestra aleatoria de la población. (Wooldridge(2010)), cuando se incumple el supuesto de que se tiene una muestra aleatoria de n observaciones que sigue el modelo poblacional debido al truncamiento por lo general produce estimadores insesgados hacia cero, según Hausman y Wise (1977).

También ocurre que el indicador de selección cuando un truncamiento proviene de arriba, y los errores estarán correlacionados, incluso de manera condicional sobre la variable independiente (x). Esta es la razón de que MCO sobre la muestra seleccionada no estime de forma consistente los coeficientes. Por los motivos antes expuestos se requiere realizar la regresión tomando en cuenta el término no linear Lambda, de manera que no se generen sesgos por tener una variable lambda omitida.

Además, cuando se tienen datos truncados, se realiza la estimación de máxima verosimilitud, que resulta en un conjunto de coeficientes estimados que maximiza la densidad, por tanto, el método matemático difiere de un modelo lineal MCO. Por tanto, es claro que los coeficientes varían de un modelo a otro.

7. En este caso, como se pretende el análisis de una subpoblación, nos interesan los efectos marginales y no los coeficientes de la regresión truncada. Interprete los coeficientes marginales en la media.

```
#Efectos marginales: sobre la media
#Primero se saca la media de las variables:
mean_educ <- mean(data_trunc$educ)
mean_exper <- mean(data_trunc$exper)
mean_expersq <- mean(data_trunc$expersq)
mean_age <- mean(data_trunc$age)
mean_kidslt6 <- mean(data_trunc$kidslt6)
mean_kidsge6 <- mean(data_trunc$kidsge6)
#Multiplicamos la media de la variable por el coeficiente:
medias <- list("mean_c" = 1, "mean_educ" = mean_educ, "mean_exper" = mean_exper, "mean_expers
coeficientes <- list("c" = 2122.06922,"coef_educ" = -29.65101, "coef_exper" = 72.60832, "coef
mediasxcoeficientes <- list(unlist(medias)*unlist(coeficientes))
suma_mediasxcoeficientes <- 1115.281
sigma <- 850.77423
#Se saca el alpha:
alpha <- (-1115.281)/850.77423
#Se saca phi minúscula y phi de alpha de la tabla normal:
phi_min_alpha <- dnorm(alpha, mean = 0, sd=1)
phi_alpha  <- pnorm(alpha, mean = 0, sd=1)
#Se saca Lambda:
Lambda <- phi_min_alpha/(1-phi_alpha)
```

```
#Se saca delta:
delta <- Lambda*(Lambda-alpha)
uno_menos_delta <- 1-delta
#Se sacan los efectos marginales sobre la media
eff_marg_media <- lapply(coeficientes, "*",uno_menos_delta)
```

Los efectos marginales en la media se interpretan como:

1. Educación: Con respecto a educación, el efecto marginal de un año de educación se observa como un cambio de -21.36208 en la media del número de horas del trabajadas individuo promedio.
2. Experiencia: Con respecto a experiencia, el efecto marginal de una unidad de experiencia se observa como un cambio de 52.31063 en la media del número de horas trabajadas del individuo promedio.
3. Experiencia al cuadrado: Con respecto a experiencia al cuadrado, el efecto marginal de una unidad de experiencia al cuadrado se observa como un cambio de -.6809173 en la media del número de horas trabajadas del individuo promedio.
4. Edad: Con respecto a la edad, el efecto marginal de un año más de edad se observa como un cambio de -19.73451 en la media del número de horas trabajadas del individuo promedio.
5. Número de hijos menores de 6 años: Con respecto al número de hijos menores de 6 años, el efecto marginal de un niño más menor a 6 años se observa como un cambio de -349.3163 en la media del número de horas trabajadas del individuo promedio.
6. Número de hijos de entre 6 años y 18 años: Con respecto al número de hijos de entre 6 años y 18 años, el efecto marginal de un niño más de entre 6 años y 18 años se observa como un cambio de -73.91465 en la media del número de horas trabajadas del individuo promedio.

8. No hay que dejar de lados los efectos marginales promedios, interprételos también.

```
#En R
summary(Modelo_trunc$marg.effects)
```

Los efectos marginales promedios se interpretan como: es el cambio promedio en la media del numero de infidelidades calculado para la muestra

1. Educación: Con respecto a educación, el efecto marginal de un año de educación se observa como un cambio promedio de -21.14343 en la media del numero de horas del trabajadas calculado para la muestra.
2. Experiencia: Con respecto a experiencia, el efecto marginal de una unidad de experiencia se observa como un cambio promedio de 51.7752 en la media del número de horas trabajadas calculado para la muestra.

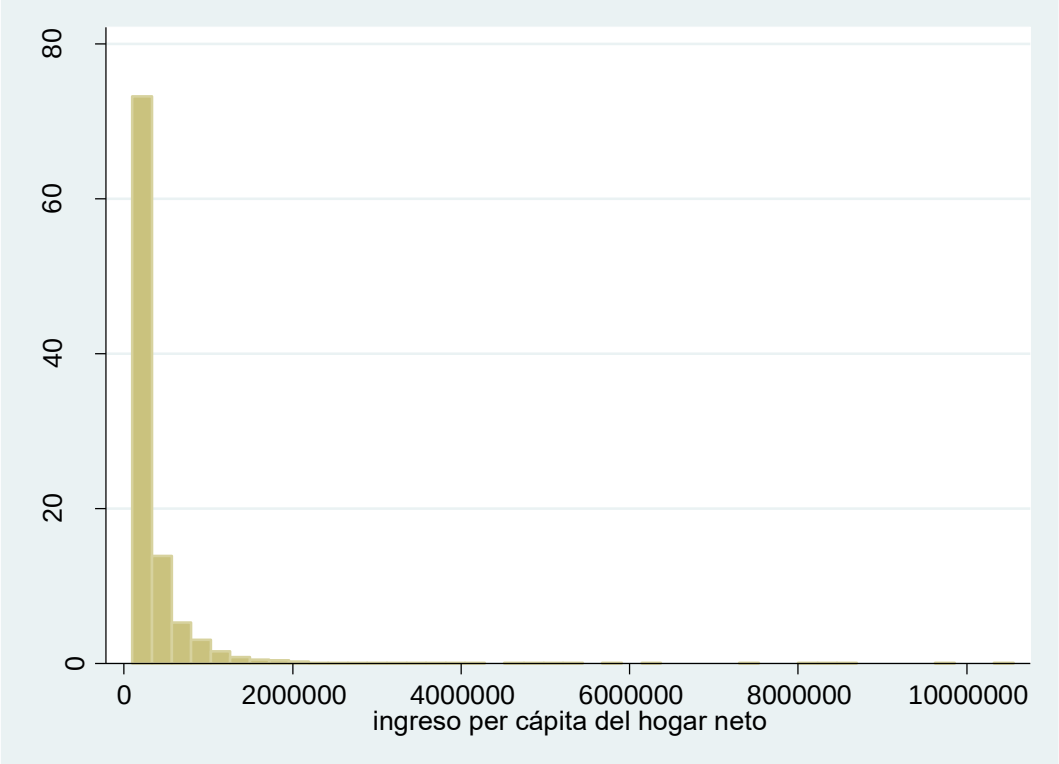
3. Experiencia al cuadrado: Con respecto a experiencia al cuadrado, el efecto marginal de una unidad de experiencia al cuadrado se observa como un cambio promedio de -.6739477 en la media del número de horas trabajadas calculado para la muestra.
4. Edad: Con respecto a la edad, el efecto marginal de un año más de edad se observa como un cambio promedio de -19.53251 en la media del número de horas trabajadas calculado para la muestra.
5. Número de hijos menores de 6 años: Con respecto al número de hijos menores de 6 años, el efecto marginal de un niño más menor a 6 años se observa como un cambio promedio de -345.7408 en la media del número de horas trabajadas calculado para la muestra.
6. Número de hijos de entre 6 años y 18 años: Con respecto al número de hijos de entre 6 años y 18 años, el efecto marginal de un niño más de entre 6 años y 18 años se observa como un cambio promedio de -73.15809 en la media del número de horas trabajadas calculado para la muestra.

0 s se ejecutó 21:43

Tercera parte

1

Investigando la variable de ingreso per cápita del hogar neto (ipcn) vemos en el siguiente gráfico la frecuencia de diferentes ingresos medida como porcentaje. Se nota que los datos están acumulados a la izquierda, pero según la base de datos estos parten del valor 99.238. Existe una censura dado al tope sobre esta variable, no permite ver los datos menores al dicho por lo que una regresión tomando esta variable como dependiente por método de MCO presentará estimadores sesgados e inconsistentes, no presentará una verdadera relación entre las variables, por lo tanto se podría utilizar el modelo de regresión Tobit para esta muestra. El valor de censura de 99.238 muestra que no los valores anteriores a este se toman como este valor, por lo que no se puede inferir correctamente la relación con las demás variables, por lo que más adelante se tomará esto en cuenta para hacer regresión censurada.



```
. tab ipcn if ipcn<100000
```

ingreso per cápita del hogar neto	Freq.	Percent	Cum.
99238	5,044	98.90	98.90
99260.79	3	0.06	98.96
99313.31	3	0.06	99.02
99403.57	4	0.08	99.10
99417.89	3	0.06	99.16
99452.38	5	0.10	99.25
99466.5	4	0.08	99.33
99500	4	0.08	99.41
99598.75	4	0.08	99.49
99612.5	2	0.04	99.53
99666.67	3	0.06	99.59
99791.67	2	0.04	99.63
99822.75	2	0.04	99.67
99846.6	2	0.04	99.71
99875	2	0.04	99.75
99893.12	5	0.10	99.84
99900	2	0.04	99.88
99917.33	3	0.06	99.94
99944.44	3	0.06	100.00
Total	5,100	100.00	

2 Manejo de datos

2.1

La variable años de escolaridad (escolari) posee un rango desde 0 (Sin escolaridad, preescolar o enseñanza especial) hasta 21 años de escolaridad. También tiene codificado como 99 años para referirse a ignorado. Se corrige esta variable eliminando aquellas observaciones iguales a 99, no presenta problema de missing values.

```
. count if missing(escolari)
  0
```

```
. tab escolari
```

años de escolaridad	Freq.	Percent	Cum.
sin escolaridad, preescolar o enseñanza	4,648	13.34	13.34
un año	1,026	2.94	16.28
dos años	1,124	3.23	19.51
tres años	1,545	4.43	23.94
cuatro años	1,075	3.09	27.03
cinco años	1,300	3.73	30.76
seis años	7,330	21.04	51.80
siete años	1,698	4.87	56.67
ocho años	1,876	5.38	62.06
nueve años	2,256	6.47	68.53
diez años	1,063	3.05	71.58
once años	4,026	11.55	83.14
doce años	1,044	3.00	86.13
trece años	697	2.00	88.13
catorce años	1,128	3.24	91.37
quince años	1,387	3.98	95.35
dieciseis años	722	2.07	97.42
diecisiete años	678	1.95	99.37
dieciocho años	29	0.08	99.45
diecinueve años	26	0.07	99.53
veinte años	16	0.05	99.57
veintiuno años	75	0.22	99.79
ignorado	74	0.21	100.00
Total	34,843	100.00	

2.2

Primero, vemos que las variables zona y sexo (a4) tienen la siguiente codificación

```
. label list zona a4
zona:
           1 urbana
           2 rural
a4:
           1 hombre
           2 mujer
```

Se corrige entonces con un recode las variables y se redefine su etiqueta. Con el comando codebook se puede verificar la modificación.

```
. recode zona a4 (2=0)
(zona: 10736 changes made)
(a4: 17850 changes made)


. label define zona 0 "rural", modify

. label define a4 0 "mujer", modify
```

2.3

Se eliminan las observaciones para los individuos menores de edad (a5)

```
. drop if a5<18
(9,360 observations deleted)
```

3

Se corre la regresión para el ingreso per cápita del hogar tomando como variables independientes años de escolaridad, zona, sexo, edad y estado conyugal. Sobre esta última se utiliza la oopción i.a6 para crear una dummy y ver los efectos de cada categoría del estado.

```
. reg ipcn escolari zona a4 a5 i.a6
```

Source	SS	df	MS	Number of obs	=	25,348
				F(9, 25338)	=	768.01
Model	1.0275e+15	9	1.1416e+14	Prob > F	=	0.0000
Residual	3.7664e+15	25,338	1.4865e+11	R-squared	=	0.2143
				Adj R-squared	=	0.2140
Total	4.7939e+15	25,347	1.8913e+11	Root MSE	=	3.9e+05

ipcn	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
escolari	46923.32	622.4229	75.39	0.000	45703.34	48143.3
zona	46195.96	5462.535	8.46	0.000	35489.08	56902.85
a4	41166.27	4935.276	8.34	0.000	31492.84	50839.7
a5	4148.324	176.5891	23.49	0.000	3802.199	4494.449
a6						
casado(a)	26416.21	7468.642	3.54	0.000	11777.24	41055.18
divorciado(a)	53302.19	13009.24	4.10	0.000	27803.34	78801.05
separado(a)	-7754.091	10452.39	-0.74	0.458	-28241.38	12733.2
viudo(a)	30038	14085.55	2.13	0.033	2429.503	57646.49
soltero(a)	6509.425	7626.444	0.85	0.393	-8438.845	21457.69
_cons	-304957.3	11326.75	-26.92	0.000	-327158.4	-282756.2

4

Vemos que en esta regresión los coeficientes son significativos con excepción de a6 que indica el estado conyugal. Los coeficientes para este caso indican el cambio en el ingreso per cápita neto del hogar conforme cambia alguna variable independiente. Por ejemplo, un año de escolaridad afecta positivamente el ipcn aumentándolo en 46.923,32 por cada año. Si la persona es de una zona urbana entonces representa un aumento de 46.195,96. Si se trata de un hombre el aumento sobre ipcn es de 41.166,27 y con cada año de edad lo aumenta en 4.148,324 (partiendo de edades superiores de 18 tanto en esta regresión como en las siguientes). El problema con estos coeficientes es que son sesgados e inconsistentes al tratarse de una muestra censura para ipcn por lo que no refleja una relación verdadera de las variables.

5

Ahora tomando en cuenta la censura se corre una regresión del modelo Tobit con las mismas variables de la regresión anterior. El valor de censura se especifica en 99.238, este corresponde a una censura desde abajo.

```
. tobit ipcn escolari zona a4 a5 i.a6, ll(99238)

Tobit regression                                  Number of obs     =     25,348
                                                  LR chi2(9)        =    6960.42
                                                  Prob > chi2       =     0.0000
Log likelihood = -296260.69                       Pseudo R2         =     0.0116
```

ipcn	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
escolari	57721.82	741.1266	77.88	0.000	56269.17	59174.47
zona	91603.91	6571.089	13.94	0.000	78724.2	104483.6
a4	60006.14	5851.207	10.26	0.000	48537.44	71474.84
a5	4408.017	210.9019	20.90	0.000	3994.637	4821.397
a6						
casado(a)	51032.01	8935.878	5.71	0.000	33517.18	68546.85
divorciado(a)	73630.55	15349.22	4.80	0.000	43545.2	103715.9
separado(a)	-10633.53	12619.6	-0.84	0.399	-35368.67	14101.6
viudo(a)	60187.37	16871.11	3.57	0.000	27119.03	93255.72
soltero(a)	16590.59	9131.104	1.82	0.069	-1306.902	34488.08
_cons	-534374.4	13718.17	-38.95	0.000	-561262.8	-507486
/sigma	440466.8	2218.209			436119	444814.7

```
  5,044  left-censored observations at ipcn <= 99238
 20,304     uncensored observations
      0 right-censored observations
```

6

Como la estimación se pretende para una subpoblación entonces se tiene interés en estimar los efectos marginales. La siguiente regresión muestra los efectos marginales en la media. Vemos aquí que un año adicional de escolaridad representa un cambio de 26.546,93 en la media del ingreso per cápita neto del hogar promedio. Vivir en la zona urbana lleva un cambio de 42.129,7. Ser hombre incrementa en la media del ingreso per cápita neto del hogar promedio un 27.597,52. Un año adicional se entiende como un cambio de 2.027,298 en la media del ipcn promedio y por último, hay un efecto marginal positivo de 23.368,84 para la persona casada, de 34.216,89 para la divorciada, uno de -4.677,119 si se trata de alguien separado, 27.726,16 para el viudo y de 7428,38 para la persona soltera.

```
. margins, predict(e(99238,.)) dydx(_all) atmeans

Conditional marginal effects                    Number of obs     =     25,348
Model VCE    : OIM

Expression   : E(ipcn|ipcn>99238), predict(e(99238,.))
dy/dx w.r.t. : escolari zona a4 a5 2.a6 3.a6 4.a6 5.a6 6.a6
at           : escolari        =    8.597838 (mean)
               zona            =    .7017516 (mean)
               a4              =    .4790911 (mean)
               a5              =    43.59397 (mean)
               1.a6            =    .1699542 (mean)
               2.a6            =    .3708774 (mean)
               3.a6            =    .0466703 (mean)
               4.a6            =    .0812293 (mean)
               5.a6            =     .044974 (mean)
               6.a6            =    .2862948 (mean)
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
escolari	26546.93	349.9143	75.87	0.000	25861.11	27232.75
zona	42129.7	3020.597	13.95	0.000	36209.43	48049.96
a4	27597.52	2691.504	10.25	0.000	22322.27	32872.77
a5	2027.298	97.30753	20.83	0.000	1836.579	2218.017
a6						
casado(a)	23368.84	4047.484	5.77	0.000	15435.92	31301.76
divorciado(a)	34216.89	7322.098	4.67	0.000	19865.84	48567.94
separado(a)	-4677.119	5537.575	-0.84	0.398	-15530.57	6176.328
viudo(a)	27726.16	7957.632	3.48	0.000	12129.49	43322.84
soltero(a)	7428.38	4080.345	1.82	0.069	-568.9488	15425.71

Note: dy/dx for factor levels is the discrete change from the base level.

7

Finalmente, se calculan los efectos marginales promedios abajo. En este caso se interpreta como un cambio promedio, por ejemplo, para la educación, a cada año adicional de educación se tiene un aumento de 27.260,6 sobre el promedio de ingresos per cápita de hogares. Luego, el estar en zona urbana y ser hombre representa en el promedio de ipcn un cambio de 43.262,28 y 28.339,43 respectivamente. El cambio en la edad se muestra como un cambio de 2.081,798 en el promedio de ipcn y el estado conyugal muestra un cambio de 2.3990,86 para casados, 35.081,09 para divorciados, de -4.818,315 en separados, de 28.448,97 para viudos y de 7.641,125 en solteros, esto sobre el promedio de ipcn.

```
. margins, predict(e(99238,.)) dydx(_all)

Average marginal effects                        Number of obs     =     25,348
Model VCE    : OIM

Expression   : E(ipcn|ipcn>99238), predict(e(99238,.))
dy/dx w.r.t. : escolari zona a4 a5 2.a6 3.a6 4.a6 5.a6 6.a6
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
escolari	27260.6	367.1819	74.24	0.000	26540.94	27980.26
zona	43262.28	3107.11	13.92	0.000	37172.46	49352.1
a4	28339.43	2765.15	10.25	0.000	22919.84	33759.02
a5	2081.798	99.96208	20.83	0.000	1885.876	2277.72
a6						
casado(a)	23990.86	4157.673	5.77	0.000	15841.97	32139.75
divorciado(a)	35081.09	7486.991	4.69	0.000	20406.86	49755.33
separado(a)	-4818.315	5705.867	-0.84	0.398	-16001.61	6364.979
viudo(a)	28448.97	8148.272	3.49	0.000	12478.65	44419.29
soltero(a)	7641.125	4197.176	1.82	0.069	-585.1895	15867.44

```
Note: dy/dx for factor levels is the discrete change from the base level.
```