# CIDE
# Maestría en Economía
# Econometría II

Respuestas a la tarea 2

Profesor: Irvin Rojas

Fecha de entrega: 5 de octubre a las 7:00.

## Instrucciones

La tarea debe entregarse de manera individual, pero se recomienda ampliamente colaborar en grupos de estudio. Las secciones teóricas deben estar desarrolladas en un procesador de textos y enviadas en formato .docx o .pdf. Las secciones prácticas deberán contener archivos de código replicable y archivos de salida en R (o similares, en caso de usar otro software) para considerarse completas. Las tareas deben entregarse antes de la fecha límite a través de Teams. Puede crear una carpeta comprimida que contenga todos sus archivos y subir esta carpeta en Teams. Recuerde que en Teams debe asegurarse de que los archivos se han subido correctamente.

## Pregunta 1

1. Retome la base de la base *motral2012.csv* usada en la Tarea 1. Estimará un modelo Tobit para explicar los factores que afectan la oferta laboral femenina. En esta la base de datos la variable **hrsocup** registra las horas trabajadas a la semana.

a. [2 punto] ¿Qué proporción de la muestra femenina reporta horas trabajadas iguales a cero?

*Si hacemos una dummy de horas positivas, al sacarle la media obtenemos la proporción:*

```
data.salarios<-read_csv("./motral2012.csv",
                        locale = locale(encoding = "latin1"))

#1a % de mujeres con horas igua a cero
data.salarios <- data.salarios %>%
  filter(sex==2) %>%
  mutate(zerohrs=ifelse(hrsocup==0,1,0))

#La media de la dummy zerohrs da el porcentaje de mujeres con horas cero
stat.desc(data.salarios$zerohrs)
##      nbr.val     nbr.null       nbr.na          min          max        range
## 2.625000e+03 1.699000e+03 0.000000e+00 0.000000e+00 1.000000e+00 1.000000e+00
##          sum       median         mean      SE.mean CI.mean.0.95          var
## 9.260000e+02 0.000000e+00 3.527619e-01 9.328052e-03 1.829108e-02 2.284080e-01
##      std.dev     coef.var
## 4.779204e-01 1.354796e+00
```

b. [3 puntos] Se desea estimar el efecto de los años de educación (**anios_esc**) sobre la oferta laboral femenina controlando por el estado marital (**casada**), la edad (**eda**) y el número de hijos (**n_hij**) como una variable continua. En la base, **e_con** toma el valor de 5 para las personas casadas. Genere la variable dummy **casada** que tome el valor de 1 para las mujeres casadas y cero en otro caso. Estime un modelo de MCO para **hrsocup** mayor que cero, usando solo la población femenina. Reporte errores robustos. ¿Cuál es la interpretación sobre el coeficiente de los años de escolaridad?

*El estimar por MCO, un año más de escolaridad se asocia con 0.17 horas trabajadas más a la semana. Sin embargo, este efecto no es estadísticamente significativo.*

```
#1b Dummy de casada y MCO
data.salarios <- data.salarios %>%
  mutate(casada=ifelse(e_con==5,1,0))

reg1b<-lm(hrsocup ~ anios_esc+casada+eda+n_hij,
      data=filter(data.salarios,hrsocup>0))
coeftest(reg1b,
     vcov = vcovHC(reg1b, "HC1"))[1:4,]
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 36.70129720 1.99116828 18.432042 2.742336e-69
## anios_esc    0.17465627 0.10353350  1.686954 9.179628e-02
## casada      -3.52571327 0.89724706 -3.929479 8.855253e-05
## eda          0.06949593 0.04914655  1.414055 1.575295e-01
```

c. [3 puntos] ¿Qué problema existe con el modelo planteado en el punto anterior en términos de la selección? ¿Considera que se trata de un caso de censura o de truncamiento?

*Podemos racionalizar las horas trabajadas en un modelo microeconómico de oferta laboral. Las horas trabajadas observadas son positivas cuando la solución óptima es una cantidad positiva de horas. Sin embargo, si la solución óptima implicara horas negativas, las horas observadas se codificarían como cero. En este caso tenemos datos censurados en cero. Si existe una relación positiva entre educación y horas trabajadas, al estimar un modelo por MCO usando solo los datos con horas positivas estamos sobreestimando la media condicional pues se habrán omitido del análisis aquellas mujeres cuya solución a su problema de optimización eran horas iguales a cero o negativas.*

d. [8 puntos] Estime un modelo Tobit de datos censurados. ¿Qué resuelve el modelo Tobit en este caso? Interprete nuevamente el coeficiente sobre los años de escolaridad.

*La función tobit permite hacer esto muy fácilmente. Noten que left especifica dónde está la censura. La opción gaussian pone explícito uno de los supuestos críticos del modelo tobit visto en clase: errores normales. Además, se asume homocedasticidad.*

```
reg1d <- tobit(hrsocup ~ anios_esc+casada+eda+n_hij,
           left = 0,
           right = Inf,
           dist = "gaussian",
           data = data.salarios)
summary(reg1d)
##
## Call:
## tobit(formula = hrsocup ~ anios_esc + casada + eda + n_hij, left = 0,
##     right = Inf, dist = "gaussian", data = data.salarios)
##
## Observations:
##          Total  Left-censored     Uncensored Right-censored
##           2625            926           1699              0
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.88236    3.19905   0.276  0.78269
## anios_esc     0.85530    0.17509   4.885 1.04e-06 ***
## casada      -10.99515    1.43025  -7.688 1.50e-14 ***
## eda           0.41621    0.07665   5.430 5.64e-08 ***
## n_hij        -1.73840    0.55887  -3.111  0.00187 **
## Log(scale)    3.44512    0.01887 182.608  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Scale: 31.35
##
## Gaussian distribution
## Number of Newton-Raphson Iterations: 3
## Log-likelihood: -9086 on 6 Df
## Wald-statistic: 127.9 on 4 Df, p-value: < 2.22e-16
```

*El modelo tobit para datos censurados toma en cuenta que hay una masa de ceros en las horas trabajadas para individuos para los que disponemos de sus características en la base de datos. El modelo tobit ajusta la probabilidad de observar esta masa de ceros. El coeficiente estimado será ahora consistente si el modelo está bien especificado, es decir, si el proceso subyacente es lineal en los parámetros y con un error normal homoscedástico (los supuestos de tobit básico). En este caso, un año más de educación se asocia con 0.85 más horas semanales trabajadas, un efecto estadísticamente significativo. Usar MCO subestimaba el efecto de la escolaridad.*

e. [4 puntos] ¿Cuál es el efecto marginal de un incremento de un año de educación en la oferta laboral? ¿Cómo cambia su respuesta si, en lugar de considerar la variable latente, considera la variable censurada?

*El efecto marginal en la variable latente es directamente el coficiente estimado en la parte d., es decir 0.855.*

*El efecto marginal en la media censurada está dado por:*

$$\frac{\partial E(y|x)}{\partial x_j} = \beta_j \Phi(x_i'\beta)$$

*Lo que hice aquí fue calcular este efecto marginal para cada individuo y luego obtener el promedio de los efectos marginales en aquellos individuos con horas positivas.*

```
#Efecto marginal promedio
data.salarios <- data.salarios %>%
  mutate(index1=predict(reg1d,.)) %>%
  mutate(phi=pnorm(index1/reg1d$scale)) %>%
  mutate(mfx_anis_esc=reg1d$coefficients[2]*phi,
     mfx_eda=reg1d$coefficients[4]*phi,
     mfx_n_hij=reg1d$coefficients[5]*phi)

data.salarios %>%
  filter(hrsocup>0) %>%
  summarise(mfx_anis_esc=mean(mfx_anis_esc))
## # A tibble: 1 x 1
##   mfx_anis_esc
```

```
##          <dbl>
## 1        0.612
```

## Pregunta 2

Usando los mismos datos de la base *motral2012.csv* implementará un ejercicio en el mismo espíritu del famoso estudio de Mroz (1987)[1] sobre la oferta laboral femenina. El propósito es estimar la relación entre el salario y el número de horas trabajadas, concentrándonos en la muestra de mujeres.

a. [5 puntos] El primer problema al que nos enfrentamos es que el salario será no observado para las mujeres que no trabajan. Estime un modelo lineal para el log del salario por hora, **ing_x_hrs**, usando las variables **anios_esc**, **eda**, **n_hij** y **casada**, usando la submuestra de mujeres con salario por hora positivo. Use los coeficientes estimados para imputar el ingreso por hora faltante para las mujeres que reportan 0 en las horas trabajadas.

*Imputamos el salario faltante:*

```
data.salarios<-read_csv("./motral2012.csv",
                        locale = locale(encoding = "latin1")) %>%
  filter(sex==2) %>%
  mutate(casada=ifelse(e_con==5,1,0))

#Log de salario ly
data.salarios <- data.salarios %>%
  mutate(ly=ifelse(ing_x_hrs>0,log(ing_x_hrs),NA))

#Modelo lineal
reg2a <- lm(ly~anios_esc+casada+eda+n_hij,
            data=data.salarios)

#Imputación
data.salarios <- data.salarios %>%
  mutate(lyhat = predict(reg2a, .)) %>%
  mutate(ly=ifelse(is.na(ly),lyhat,ly))
```

b. [5 puntos] Use una función[2] para estimar por máxima verosimilitud un *heckit* para las horas trabajadas **hrsocup**. En la ecuación de selección (si la persona trabaja o no) incluya como variable explicativa el salario por hora (imputado para las mujeres que no trabajan), además de **anios_esc**, **eda**, **n_hij** y **casada**. En la ecuación de horas, incluya los mismos regresores, excepto **n_hij**.

*La función heckit permite estimar el modelo de Heckman por máxima verosimilitud de manera muy simple. Hay que especificar method="ml" para que la estimación sea por máxima verosimilitud:*

```
data.salarios <- data.salarios %>%
  mutate(trabaja=ifelse(hrsocup>0,1,0)) %>%
  mutate(trabaja=factor(trabaja,levels=c(0,1)))

reg2b <- heckit(trabaja ~ anios_esc+casada+eda+ly+n_hij,
                hrsocup ~ anios_esc+casada+eda+ly,
                method="ml",
```

[1] Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica*: Journal of the econometric society, 765-799.

[2] Por ejemplo, la función *heckit* del paquete *sampleSelection* en R.

```
            data = data.salarios)
summary(reg2b)
## --------------------------------------------
## Tobit 2 model (sample selection model)
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 12 iterations
## Return code 2: successive function values within tolerance limit
## Log-Likelihood: -8648.707
## 2625 observations (926 censored and 1699 observed)
## 13 free parameters (df = 2612)
## Probit selection equation:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.158822   0.169184  -0.939 0.347944
## anios_esc    0.027962   0.008172   3.422 0.000631 ***
## casada      -0.345200   0.055716  -6.196 6.72e-10 ***
## eda          0.017460   0.003037   5.749 1.00e-08 ***
## ly          -0.042329   0.058446  -0.724 0.468977
## n_hij       -0.055036   0.019405  -2.836 0.004600 **
## Outcome equation:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 67.36837    2.99153  22.520  < 2e-16 ***
## anios_esc    0.78457    0.11992   6.542 7.26e-11 ***
## casada      -0.94025    0.94859  -0.991    0.322
## eda          0.03031    0.04463   0.679    0.497
## ly          -9.82070    0.70437 -13.943  < 2e-16 ***
##    Error terms:
##       Estimate Std. Error t value Pr(>|t|)
## sigma 17.33388    0.73855  23.470   <2e-16 ***
## rho   -0.70890    0.07206  -9.838   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

c. [10 puntos] Estime ahora el *heckit* en dos pasos, *a mano*. Es decir, siga los siguientes pasos: i) estime un probit para la ecuación de selección y obtenga el índice $x_i'\hat{\beta}$; ii) calcule el inverso de la razón de Mills $\lambda_i(x_i'\hat{\beta})$; y iii) estime por MCO la ecuación para las horas trabajadas con la submuestra que tiene horas trabajadas positivas, incluyendo como regresor el inverso de la razón de Mills estimado y el resto de los regresores.

Compare los coeficientes y los errores estándar obtenidos en esta parte con los de la parte b. ¿Por qué son iguales o por qué difieren?

*Estimamos ahora el heckit* a mano*:*

```
#Probit
mod.probit <- glm(trabaja ~ anios_esc+casada+eda+ly+n_hij,
              family = binomial(link = "probit"),
              data = data.salarios)

#Predicción del índice y cálculo de IMR
data.salarios <- data.salarios %>%
  mutate(index = predict(mod.probit, .)) %>%
  mutate(imr = dnorm(index)/pnorm(index))

#Segunda etapa
```

```r
reg2c <- lm(hrsocup ~ anios_esc+casada+eda+ly+imr,
        data=filter(data.salarios,trabaja==1))

summary(reg2c)
##
## Call:
## lm(formula = hrsocup ~ anios_esc + casada + eda + ly + imr, data = filter(data.salarios,
##     trabaja == 1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.529  -9.692   1.829   9.189  58.977
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 103.4350    13.5104   7.656  3.2e-14 ***
## anios_esc     0.0453     0.2965   0.153 0.878588
## casada        7.1635     3.0995   2.311 0.020945 *
## eda          -0.2381     0.1059  -2.249 0.024663 *
## ly           -9.7272     0.6028 -16.138  < 2e-16 ***
## imr         -50.7272    14.1222  -3.592 0.000337 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.81 on 1693 degrees of freedom
## Multiple R-squared:  0.1621, Adjusted R-squared:  0.1596
## F-statistic:  65.5 on 5 and 1693 DF,  p-value: < 2.2e-16

#El heckit por MV y en dos etapas no coinciden
```

*El heckit estimado por máxima verosimilitud y en dos etapas a mano no coinciden. Notemos primero que podemos estimar el modelo en dos etapas con la función heckit:*

```r
reg2c_alt <- heckit(trabaja ~ anios_esc+casada+eda+ly+n_hij,
            hrsocup ~ anios_esc+casada+eda+ly,
            method="2step",
            data = data.salarios)

stargazer(reg2c, reg2c_alt, reg2b,
      title="Comparación de heckit en 2 etapas, a mano y con la función heckit", type="text",
      df=FALSE, digits=4)
##
## Comparación de heckit en 2 etapas, a mano y con la función heckit
## ================================================================
##                                 Dependent variable:
##                     ---------------------------------------------
##                                       hrsocup
##                           OLS                  Heckman
##                                               selection
##                           (1)             (2)             (3)
## ----------------------------------------------------------------
## anios_esc               0.0453          0.0453         0.7846***
##                        (0.2965)        (0.6152)        (0.1199)
```

```
##
## casada                7.1635**          7.1635          -0.9403
##                       (3.0995)         (6.3041)         (0.9486)
##
## eda                   -0.2381**         -0.2381          0.0303
##                       (0.1059)         (0.2187)         (0.0446)
##
## ly                   -9.7272***       -9.7273***       -9.8207***
##                       (0.6028)         (1.6212)         (0.7044)
##
## imr                  -50.7272***
##                       (14.1222)
##
## Constant             103.4350***      103.4349***      67.3684***
##                       (13.5104)        (27.4220)        (2.9915)
##
## ------------------------------------------------------------------
## Observations            1,699            2,625            2,625
## R2                     0.1621           0.1621
## Adjusted R2            0.1596           0.1596
## rho                                     -1.2732
## Inverse Mills Ratio                -50.7270* (28.4440)
## Residual Std. Error   14.8107
## F Statistic          65.5003***
## ==================================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

*Tardé bastante en darme cuenta por qué los coeficientes de máxima verosimilitud y del procedimiento en dos etapas no eran casi iguales. El problema es que la variable de número de hijos no ayudaba a discriminar bien entre quienes participaban y quienes no. Entonces, añadí dos variables al proceso de selección: el número de hijos al cuadrado y una variable que identifica si el individuo ha buscado trabajo recientemente:*

```
#heckit por MV, aumentado

reg_mv_aumentado <- heckit(trabaja ~ anios_esc+casada+eda+ly+n_hij+n_hij^2+busqueda,
            hrsocup ~ anios_esc+casada+eda+ly,
            method="ml",
            data = data.salarios)

reg_2etapas_aumentado <- heckit(trabaja ~ anios_esc+casada+eda+ly+n_hij+n_hij^2+busqueda,
                      hrsocup ~ anios_esc+casada+eda+ly,
                      method="2step",
                      data = data.salarios)


#Usé stargazer para poner mis resultados
stargazer(reg_mv_aumentado, reg_2etapas_aumentado,
      title="Comparación de heckit con MV y en 2 etapas", type="text",
      df=FALSE, digits=4)
##
## Comparación de heckit con MV y en 2 etapas
## ================================================
##                          Dependent variable:
```

```
##                         ----------------------------
##                                    hrsocup
##                            (1)           (2)
## ---------------------------------------------------
## anios_esc               1.0489***     1.0493***
##                         (0.1000)       (0.1002)
##
## casada                 -3.5838***     -3.5623***
##                         (0.7798)       (0.7816)
##
## eda                     0.1161***     0.1163***
##                         (0.0390)       (0.0391)
##
## ly                     -9.8400***     -9.8404***
##                         (0.6040)       (0.6055)
##
## Constant               55.6425***     55.7211***
##                         (2.1778)       (2.1842)
##
## ---------------------------------------------------
## Observations              2,625          2,625
## R2                       0.1564
## Adjusted R2              0.1539
## rho                                    -0.2873
## Inverse Mills Ratio                 -4.2798 (3.5489)
## ===================================================
## Note:                  *p<0.1; **p<0.05; ***p<0.01
```

*Ahora sí, notamos que los coeficientes son casi iguales. Por otro lado, los errores estándar también son muy parecidos, pues el heckit con dos etapas incorpora ya la corrección propuesta por Heckman para tomar en cuenta que en la primera etapa el inverso de la razón de Mills es estimado.*

*¿Qué pasa si tratamos de hacer el procedimiento a mano, con estos dos regresores añadidos en el proceso de selección? Veamos:*

```
#Probit
mod.probit_aumentado <- glm(trabaja ~ anios_esc+casada+eda+ly+n_hij+n_hij^2+busqueda,
              family = binomial(link = "probit"),
              data = data.salarios)

#Predicción del índice y cálculo de IMR
data.salarios <- data.salarios %>%
  mutate(index_aumentado = predict(mod.probit_aumentado, .)) %>%
  mutate(imr_aumentado = dnorm(index_aumentado)/pnorm(index_aumentado))

#Segunda etapa
reg_amano_aumentado <- lm(hrsocup ~ anios_esc+casada+eda+ly+imr_aumentado,
        data=filter(data.salarios,trabaja==1))

stargazer(reg_mv_aumentado, reg_2etapas_aumentado,reg_amano_aumentado,
      title="Comparación de heckit con MV y en 2 etapas", type="text",
      df=FALSE, digits=4)
##
## Comparación de heckit con MV y en 2 etapas
```

```
## ================================================================
##                                Dependent variable:
##                            ------------------------------------
##                                        hrsocup
##                                 Heckman              OLS
##                                selection
##                              (1)            (2)           (3)
## ----------------------------------------------------------------
## anios_esc                  1.0489***      1.0493***     1.0493***
##                             (0.1000)       (0.1002)      (0.1000)
##
## casada                     -3.5838***     -3.5623***    -3.5623***
##                             (0.7798)       (0.7816)      (0.7806)
##
## eda                        0.1161***      0.1163***     0.1163***
##                             (0.0390)       (0.0391)      (0.0390)
##
## ly                         -9.8400***     -9.8404***    -9.8404***
##                             (0.6040)       (0.6055)      (0.6040)
##
## imr_aumentado                                           -4.2798
##                                                         (3.6295)
##
## Constant                   55.6425***     55.7211***    55.7211***
##                             (2.1778)       (2.1842)      (2.1809)
##
## ----------------------------------------------------------------
## Observations                2,625          2,625          1,699
## R2                          0.1564         0.1564
## Adjusted R2                 0.1539         0.1539
## rho                                       -0.2873
## Inverse Mills Ratio                   -4.2798 (3.5489)
## Residual Std. Error                                     14.8609
## F Statistic                                            62.7732***
## ================================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

*Ahora sí, la magnitud de los coeficientes es práctiamente la misma entre el modelo estimado por máxima verosimilitud (columna 1), con un procedimiento en dos etapas con las fórmulas apropiadas para la varianza del estimador, ya incluidas en la función heckit (columna 2) y con un procedimiento en dos etapas a mano, donde no tomamos en cuenta que el inverso de la razón de Mills es estimado. Aunque en este ejemplo las diferencias son sutiles, en la práctica, debemos usar las fórmas correctas de la matriz de varianza de los estimadores.*

## Pregunta 3

En esta pregunta mostrará cómo para un modelo en dos partes Poisson la log verosimilitud del problema es la suma de log verosimilitud para un proceso binario y la log verosimilitud de un proceso Poisson truncado en cero. Considere una variable aleatoria $Y$ con observaciones iid que sigue una distribución Poisson con parámetro $\lambda$ tal que

$$f(y, \lambda) = P(Y = y) = \frac{\lambda^y exp(-\lambda)}{y!}$$

a. [4 puntos] Obtenga la distribución Poisson truncada en cero, definida como $P(Y = y|Y > 0)$.

*Sabemos que la distribución truncada en cero es:*

$$P(Y = y|Y > 0) = \frac{f(y, \lambda)}{1 - f(0, \lambda)}$$

*Sustituyendo la forma de la densidad Poisson:*

$$P(Y = y|Y > 0) = \frac{\frac{\lambda^y exp(-\lambda)}{y!}}{1 - exp(-\lambda)} = \frac{\lambda^y}{y!(exp(\lambda) - 1)}$$

b. [4 puntos] Considere además un proceso binomial que modela la probabilidad de que la variable $Y$ tome un valor cero o un valor positivo, como sigue:

$$P(Y = y) = \begin{cases} \pi & y = 0 \\ 1 - \pi & y = 1, 2, 3, \ldots \end{cases}$$

Especialice la ecuación del modelo de dos partes vista en la sesión 10, usando la distribución truncada derivada en a. y el proceso binomial definido para obtener una función de masa de probabilidad no condicional para $Y$, $g(y)$.

*En clase vimos la forma general del modelo en dos partes:*

$$g(y) = \begin{cases} f_1(0) & \text{si } y = 0 \\ \frac{(1 - f_1(0)) f_2(y)}{1 - f_2(0)} & \text{si } y \geq 1 \end{cases}$$

*Sea $\pi$ la probabilidad de observar un conteo igual a cero, especializamos la función vista en clase, incorporando la distribución truncada encontrada en la parte a.:*

$$g(y) = \begin{cases} \pi & \text{si } y = 0 \\ (1 - \pi) \frac{\lambda^y}{y!(exp(\lambda) - 1)} & \text{si } y \geq 1 \end{cases}$$

c. [4 puntos] Obtenga la log verosimilitud para la $i$ésima observación. Se sugiere que continúe sus cálculos con una ecuación en dos partes.

*La log verosimilitud de la $i$ésima observación es:*

$$\updownarrow_i(\pi, \lambda, y_i) = \begin{cases} \ln(\pi) & \text{si } y = 0 \\ \ln\left((1 - \pi) \frac{\lambda^{y_i}}{y!(exp(\lambda) - 1)}\right) & \text{si } y \geq 1 \end{cases}$$

d. [4 puntos] En este problema, parametrizaremos $\lambda_i$ como $\lambda_i = exp(x_i' \beta_2)$, como regularmente lo hemos hecho en una regresión Poisson. Por otro lado, podemos trabajar con una parametrización general de la probabilidad $\pi$, $\pi = F(x_i' \beta_1)$. Escriba la función de log verosimilitud del problema usando la parametrización para $\pi_i$ y para $\lambda_i$ que acabamos de describir. Presente esta función en una sola parte.

*Con la parametrización dada, podemos reescribir la log verosimilitud de una observación como:*

$$\updownarrow_i(\pi, \lambda, y_i) = \begin{cases} \ln(F(x_i' \beta_1)) & \text{si } y = 0 \\ \ln\left((1 - F(x_i' \beta_1)) \frac{exp(x_i' \beta_2)^{y_i}}{y!(exp(exp(x_i' \beta_2)) - 1)}\right) & \text{si } y \geq 1 \end{cases}$$

*La log verosimilitud del problema es la probabilidad de observar los datos. Con la parametrización anterior:*

$$\mathcal{L}(\beta_1, \beta_2, y_i) = \ln\left(\prod_{i\in y_i=0} F(x_i'\beta_1) \prod_{i\in y_i>0} (1 - F(x_i'\beta_1)) \prod_{i\in y_i>0} \frac{exp(x_i'\beta_2)^{y_i}}{y!(exp(exp(x_i'\beta_2)) - 1)}\right)$$

*Distribuyendo el logarítmo:*

$$\mathcal{L}(\beta_1, \beta_2, y_i) = \sum_{i\in y_i=0} \ln(F(x_i'\beta_1)) + \sum_{i\in y_i>0} \ln(1 - F(x_i'\beta_1)) + \sum_{i\in y_i>0} x_i'\beta_2 y_i - \sum_{i\in y_i>0} \ln(exp(exp(x_i'\beta_2)) - 1) - \sum_{i\in y_i>0} y!$$

e. [4 puntos] Agrupe los términos para mostrar que $\mathcal{L} = \mathcal{L}_1(\beta_1) + \mathcal{L}_2(\beta_2)$. Así, mostrará que la log verosimilitud del problema se puede descomponer en una log verosimilitud para el modelo binario y otra para el conteo truncado en cero. Por tanto, no perdemos información si estimamos los parámetros de la probabilidad binomial por un lado, y los de la distribución Poisson truncada en cero, por el otro.

*Claramente:*

$$\mathcal{L}(\beta_1, \beta_2, y_i) = \mathcal{L}_\infty(\beta_1, y_i) + \mathcal{L}_\in(\beta_2, y_i)$$

*es decir, la suma de dos log verosimilitudes, una de un proceso binario y otra para el modelo Poisson truncado en cero.*

## Pregunta 4

Partiendo de la variable aleatoria $Y$ con observaciones iid que sigue una distribución Poisson con parámetro $\lambda$ usada en el problema anterior, en este problema caracterizará la estimación de un modelo Poisson inflado en cero.

a. [4 puntos] Especialice la expresión vista en la sesión 10 para obtener la función de masa de probabilidad del modelo Poisson inflado en cero $g(y|\lambda, \pi)$.

*En clase, vimos la expresión general para el modelo inflado en cero:*

$$g(y) = \begin{cases} f_1(0)(1 - f_1(0))f_2(0) & \text{si } y = 0 \\ (1 - f_1(0))f_2(y) & \text{si } y \geq 1 \end{cases}$$

*En el caso particular de un modelo Poisson, sabemos que $f_2(0) = P(Y = 0) = exp(-\lambda)$. Definiendo la probabilidad de observar un conteo cero como $\pi$, la función de masa de probabilidad del modelo inflado en cero es:*

$$g(y) = \begin{cases} \pi + (1 - \pi)exp(-\lambda) & \text{si } y = 0 \\ (1 - \pi)\frac{\lambda^y exp(-\lambda)}{y!} & \text{si } y \geq 1 \end{cases}$$

b. [4 puntos] Provea una expresión para la función de verosimilitud $L(\lambda, \pi) = \prod_{i=1}^N g(y_i|\lambda, \pi)$. Una sugerencia para simplificar sus cálculos es definir una variable $X$ igual al numero de veces que $Y_i$ que toma el valor de cero.

*La función de verosimilitud del problema es:*

$$L(\pi, \lambda, y_i) = \prod_i P(Y_i = y_i)$$

*Con las formas específicas para el modelo Poisson inflado en cero:*

$$L(\pi, \lambda, y_i) = \prod_{i \in y_i = 0} (\pi + (1-\pi)exp(-\lambda)) \prod_{i \in y_i > 0} \left( (1-\pi) \frac{\lambda^{y_i} exp(-\lambda)}{y!} \right)$$

*Haciendo $X$ el número de veces que $y_i$ toma el valor de cero, el primer producto es $(\pi + (1-\pi)exp(-\lambda))$ elevado a la potencia $X$.*

*¿Cuántos términos distintos de cero quedan? Qudan $n - X$. El segundo producto en la verosimilitud es:*

$$\left((1-\pi)exp(-\lambda)\right)^{n-X} \frac{\lambda^{\sum_i y_i}}{\prod_{i \in y_i > 0} y!}$$

*La verosimilitud es por tanto:*

$$L(\pi, \lambda, y_i) = \left(\pi + (1-\pi)exp(-\lambda)\right)^{X} \left((1-\pi)exp(-\lambda)\right)^{n-X} \frac{\lambda^{\sum_i y_i}}{\prod_{i \in y_i > 0} y!}$$

c. [6 puntos] Provea una expresión para la log verosimilitud del problema, $\mathcal{L}(\lambda, \pi)$.

*Dada la verosimilitud planteada en la parte anterior, la log verosimilitud es:*

$$\mathcal{L}(\pi, \lambda, y_i) = X \ln\left(\pi + (1-\pi)exp(-\lambda)\right) + (n-X)\ln(1-\pi) - (n-X)\lambda + n\bar{Y}\ln(\lambda) - \ln\left(\prod_{i \in y_i > 0} y!\right)$$

d. [6 puntos] Obtenga las condiciones de primer orden que caracterizan la solución del problema de máxima verosimilitud, derivando la log verosimilitud con respecto a $\lambda$ y a $\pi$.

*Tenemos dos parámetros, así que tenemos dos condiciones de primer orden. Derivando la log verosimilitud con respecto a $\pi$ obtenemos:*

$$\frac{\partial \mathcal{L}}{\partial \pi} = \frac{X}{\pi + (1-\pi)exp(-\lambda)}(1 - exp(-\lambda)) - \frac{n-X}{1-\pi} = 0$$

*La primer condición (A) es:*

$$\frac{X(1-exp(-\lambda))(1-\pi)}{\pi + (1-\pi)exp(-\lambda)} = n - X \qquad \ldots(A)$$

*Ahora derivando la log verosimilitud con respecto a $\lambda$:*

$$\frac{\partial \mathcal{L}}{\partial \lambda} = -\frac{X}{\pi + (1-\pi)exp(-\lambda)}(1-\pi)exp(-\lambda) - (n-X) + \frac{n\bar{Y}}{\lambda} = 0$$

*La segunda condición (B) es:*

$$\frac{X(1-\pi)exp(-\lambda)}{\pi + (1-\pi)exp(-\lambda)} + (n-X) = \frac{n\bar{Y}}{\lambda} \qquad \ldots(B)$$

*$(\hat{\pi}_{MV}, \hat{\lambda}_{MV})$ son los valores de los parámetros que resulven el sistema dado por (A) y (B).*

## Pregunta 5

Uno de los debates más activos en economía es el relativo a la relación entre años de educación e ingreso. La base de datos *ingresos_iv.dta* contiene una muestra de hombres de entre 24 y 36 años de edad.

a. [2 puntos] Estime una regresión por MCO para explicar el logaritmo del salario (**lwage**) en función de la educación **educ** y los siguientes controles: **exper**, **expersq**, **black**, **south**, **smsa**, **reg661**, **reg662**, **reg663**, **reg664**, **reg665**, **reg666**, **reg667**, **reg668** y **smsa66**. ¿Qué problema encuentra en la estimación de esta relación? ¿El coeficiente sobre **educ** tiene una interpretación causal del efecto de la educación en el salario?

*Estimamos por MCO la relación entre salarios y educación, controlando por un conjunto de regresores:*

```
data.ingresos<-read_csv("./ingresos_iv.csv",
                    locale = locale(encoding = "latin1"))

reg5a <- lm(lwage ~ educ + exper + expersq + black + south + smsa + reg661 +
          reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66,
        data = data.ingresos)
summary(reg5a)
##
## Call:
## lm(formula = lwage ~ educ + exper + expersq + black + south +
##     smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
##     reg667 + reg668 + smsa66, data = data.ingresos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.62326 -0.22141  0.02001  0.23932  1.33340
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.7393765  0.0715282  66.259  < 2e-16 ***
## educ         0.0746933  0.0034983  21.351  < 2e-16 ***
## exper        0.0848320  0.0066242  12.806  < 2e-16 ***
## expersq     -0.0022870  0.0003166  -7.223 6.41e-13 ***
## black       -0.1990123  0.0182483 -10.906  < 2e-16 ***
## south       -0.1479550  0.0259799  -5.695 1.35e-08 ***
## smsa         0.1363845  0.0201005   6.785 1.39e-11 ***
## reg661      -0.1185697  0.0388301  -3.054 0.002281 **
## reg662      -0.0222026  0.0282575  -0.786 0.432092
## reg663       0.0259703  0.0273644   0.949 0.342670
## reg664      -0.0634942  0.0356803  -1.780 0.075254 .
## reg665       0.0094551  0.0361174   0.262 0.793503
## reg666       0.0219476  0.0400984   0.547 0.584182
## reg667      -0.0005887  0.0393793  -0.015 0.988073
## reg668      -0.1750058  0.0463394  -3.777 0.000162 ***
## smsa66       0.0262417  0.0194477   1.349 0.177327
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3723 on 2994 degrees of freedom
## Multiple R-squared:  0.2998, Adjusted R-squared:  0.2963
## F-statistic: 85.48 on 15 and 2994 DF,  p-value: < 2.2e-16
```

*Hay una relación de 7.4% mayor ingreso por cada año de educación adicional. Sin embargo, esta no es una relación causal pues es muy probable que la educación no sea exógena en la ecuación de salarios. Esto puede deberse, por ejemplo, a una variable omitida de habilidad que afecta tanto al número de años de educación alcanzados como al desempeño en el mercado de trabajo.*

b. [2 puntos] Se propone usar una variable dicotómica que indica si el individuo vivía cerca de una universidad cuando tenía cuatro años, como instrumento de los años de educación. ¿Qué condiciones debe cumplir la variable propuesta para funcionar como instrumento válido?

*El instrumento debe cumplir dos condiciones:*

*Exogeneidad: el instrumento no debe pertenecer a la ecuación de salarios. Es decir, el haber crecido cerca de una universidad no debe afectar el salario contemporáneo de forma directa.*

*Relevancia: el instrumento debe estar correlacionado con la variable endógena. En este caso, haber crecido cerca de una universidad debe estar correlacionado con el número de años de educación completados.*

c. [2 puntos] ¿Cómo juzga la propuesta de usar la variable antes descrita como instrumento?

*Este argumento fue usado por Card (1995) para mostrar que los rendimientos a la educación están subestimados por un estimador de MCO. Card muestra que al usar variables instrumentales, el efecto estimado es de 25 a 60% más grande.*

*No hay una respuesta correcta o incorrecta. Quiero leer sus argumentos.*

d. [4 puntos] Estime la relación entre el logaritmo del salario y la educación usando la variable dicotómica de acceso a una universidad como instrumento. Emplee las mismas variables de control que en el modelo de MCO.

```
reg5d <- ivreg(lwage ~ educ + exper + expersq + black + south + smsa + reg661 +
             reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66  |
             nearc4 + exper + expersq + black + south + smsa + reg661 +
             reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66,
           data=data.ingresos)
summary(reg5d)
##
## Call:
## ivreg(formula = lwage ~ educ + exper + expersq + black + south +
##     smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
##     reg667 + reg668 + smsa66 | nearc4 + exper + expersq + black +
##     south + smsa + reg661 + reg662 + reg663 + reg664 + reg665 +
##     reg666 + reg667 + reg668 + smsa66, data = data.ingresos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83164 -0.24075  0.02428  0.25208  1.42760
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.7739651  0.9349470   4.037 5.56e-05 ***
## educ         0.1315038  0.0549637   2.393  0.01679 *
## exper        0.1082711  0.0236586   4.576 4.92e-06 ***
## expersq     -0.0023349  0.0003335  -7.001 3.12e-12 ***
## black       -0.1467757  0.0538999  -2.723  0.00650 **
## south       -0.1446715  0.0272846  -5.302 1.23e-07 ***
## smsa         0.1118083  0.0316620   3.531  0.00042 ***
## reg661      -0.1078142  0.0418137  -2.578  0.00997 **
```

```
## reg662      -0.0070464  0.0329073  -0.214  0.83046
## reg663       0.0404446  0.0317806   1.273  0.20325
## reg664      -0.0579171  0.0376059  -1.540  0.12364
## reg665       0.0384577  0.0469387   0.819  0.41267
## reg666       0.0550887  0.0526597   1.046  0.29559
## reg667       0.0267580  0.0488287   0.548  0.58373
## reg668      -0.1908912  0.0507113  -3.764  0.00017 ***
## smsa66       0.0185311  0.0216086   0.858  0.39119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3883 on 2994 degrees of freedom
## Multiple R-Squared: 0.2382,  Adjusted R-squared: 0.2343
## Wald test: 51.01 on 15 and 2994 DF,  p-value: < 2.2e-16
```

e. [2 puntos] Interprete la primera etapa en términos del coeficiente sobre el instrumento y la magnitud y significancia del estadístico $F$.

*En la primera etapa, haber vivido cerca de una universidad incrementa en 0.32 los años de escolaridad acumulados. Este efecto estadísticamente significativo al 1% El estadístico F es de una magnitud muy por encima de 10, la regla de dedo comúnmente empleada para juzgar la presencia de instrumentos débiles.*

```
reg5e <- lm(educ ~ nearc4 + exper + expersq + black + south + smsa + reg661 +
          reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66,
        data=data.ingresos)
summary(reg5e)
##
## Call:
## lm(formula = educ ~ nearc4 + exper + expersq + black + south +
##     smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
##     reg667 + reg668 + smsa66, data = data.ingresos)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.545 -1.370 -0.091  1.278  6.239
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.8485239  0.2111222  79.805  < 2e-16 ***
## nearc4       0.3198989  0.0878638   3.641 0.000276 ***
## exper       -0.4125334  0.0336996 -12.241  < 2e-16 ***
## expersq      0.0008686  0.0016504   0.526 0.598728
## black       -0.9355287  0.0937348  -9.981  < 2e-16 ***
## south       -0.0516126  0.1354284  -0.381 0.703152
## smsa         0.4021825  0.1048112   3.837 0.000127 ***
## reg661      -0.2102710  0.2024568  -1.039 0.299076
## reg662      -0.2889073  0.1473395  -1.961 0.049992 *
## reg663      -0.2382099  0.1426357  -1.670 0.095012 .
## reg664      -0.0930890  0.1859827  -0.501 0.616742
## reg665      -0.4828875  0.1881872  -2.566 0.010336 *
## reg666      -0.5130857  0.2096352  -2.448 0.014442 *
## reg667      -0.4270887  0.2056208  -2.077 0.037880 *
## reg668       0.3136204  0.2416739   1.298 0.194490
```

```
## smsa66       0.0254805  0.1057692   0.241 0.809644
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.941 on 2994 degrees of freedom
## Multiple R-squared:  0.4771, Adjusted R-squared:  0.4745
## F-statistic: 182.1 on 15 and 2994 DF,  p-value: < 2.2e-16
```

f. [2 puntos] Interprete el coeficiente sobre la variable de educación en la segunda etapa. Compare la magnitud del efecto estimado con el resultado de MCO.

*El coeficiente estimado sobre los años de educación indica que un año adicional de escolaridad incrementa en 13.15% el salario. Este efecto es casi el doble del estimado por MCO y estadísticamente significativo al 5%.*

```
stargazer(reg5a, reg5d,
      title="Comparación de estimadores de MCO y VI", type="text",
      df=FALSE, digits=4)
##
## Comparación de estimadores de MCO y VI
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                lwage
##                           OLS       instrumental
##                                       variable
##                           (1)            (2)
## ------------------------------------------------
## educ                   0.0747***      0.1315**
##                         (0.0035)      (0.0550)
##
## exper                  0.0848***      0.1083***
##                         (0.0066)      (0.0237)
##
## expersq                -0.0023***    -0.0023***
##                         (0.0003)      (0.0003)
##
## black                  -0.1990***    -0.1468***
##                         (0.0182)      (0.0539)
##
## south                  -0.1480***    -0.1447***
##                         (0.0260)      (0.0273)
##
## smsa                   0.1364***      0.1118***
##                         (0.0201)      (0.0317)
##
## reg661                 -0.1186***    -0.1078***
##                         (0.0388)      (0.0418)
##
## reg662                  -0.0222        -0.0070
##                         (0.0283)      (0.0329)
##
## reg663                   0.0260         0.0404
##                         (0.0274)      (0.0318)
```

```
##
## reg664                   -0.0635*       -0.0579
##                          (0.0357)       (0.0376)
##
## reg665                    0.0095         0.0385
##                          (0.0361)       (0.0469)
##
## reg666                    0.0219         0.0551
##                          (0.0401)       (0.0527)
##
## reg667                   -0.0006         0.0268
##                          (0.0394)       (0.0488)
##
## reg668                  -0.1750***     -0.1909***
##                          (0.0463)       (0.0507)
##
## smsa66                    0.0262         0.0185
##                          (0.0194)       (0.0216)
##
## Constant                4.7394***      3.7740***
##                          (0.0715)       (0.9349)
##
## ------------------------------------------------
## Observations              3,010          3,010
## R2                        0.2998         0.2382
## Adjusted R2               0.2963         0.2343
## Residual Std. Error       0.3723         0.3883
## F Statistic             85.4763***
## ================================================
## Note:                  *p<0.1; **p<0.05; ***p<0.01
```

g. [4 puntos] Realice ahora el siguiente procedimiento. Primero, estime la primera etapa usando una regresión por MCO. Obtenga los valores ajustados de educación y llámelos **educ_hat**. Luego, estime la segunda etapa empleando **educ_hat** como variable independiente, además del resto de variables de control. ¿Cómo cambian sus resultados en comparación con la parte d.?

*La magnitud de los coeficientes estimados es la misma. Esto es lo que esperábamos pues sabemos que el estimador de MC2E puede entenderse como un procedimiento donde primero se estiman los valores ajustados de la variable endógena usando el instrumento y las variables de control y luego se usan estos valores ajustados en la ecuación estructural. En cambio, los errores estándar son algo distintos.*

```
data.ingresos <- data.ingresos %>%
  mutate(educ_hat = predict(reg5e, .))

reg5g <- lm(lwage ~ educ_hat + exper + expersq + black + south + smsa + reg661 +
          reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66,
        data=data.ingresos)

#Comparamos
stargazer(reg5d, reg5g,
      title="Comparación de VI con la función ivreg y el estimador a mano", type="text",
      df=FALSE, digits=4)
##
## Comparación de VI con la función ivreg y el estimador a mano
```

```
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                  lwage
##                        instrumental       OLS
##                          variable
##                             (1)             (2)
## ------------------------------------------------
## educ                     0.1315**
##                          (0.0550)
##
## educ_hat                                  0.1315**
##                                           (0.0565)
##
## exper                    0.1083***       0.1083***
##                          (0.0237)         (0.0243)
##
## expersq                 -0.0023***      -0.0023***
##                          (0.0003)         (0.0003)
##
## black                   -0.1468***      -0.1468***
##                          (0.0539)         (0.0554)
##
## south                   -0.1447***      -0.1447***
##                          (0.0273)         (0.0281)
##
## smsa                     0.1118***       0.1118***
##                          (0.0317)         (0.0326)
##
## reg661                  -0.1078***      -0.1078**
##                          (0.0418)         (0.0430)
##
## reg662                    -0.0070         -0.0070
##                          (0.0329)         (0.0338)
##
## reg663                    0.0404           0.0404
##                          (0.0318)         (0.0327)
##
## reg664                    -0.0579         -0.0579
##                          (0.0376)         (0.0387)
##
## reg665                    0.0385           0.0385
##                          (0.0469)         (0.0483)
##
## reg666                    0.0551           0.0551
##                          (0.0527)         (0.0541)
##
## reg667                    0.0268           0.0268
##                          (0.0488)         (0.0502)
##
## reg668                  -0.1909***      -0.1909***
##                          (0.0507)         (0.0521)
##
```

```
## smsa66                   0.0185          0.0185
##                         (0.0216)        (0.0222)
##
## Constant                3.7740***       3.7740***
##                         (0.9349)        (0.9613)
##
## -----------------------------------------------
## Observations              3,010           3,010
## R2                       0.2382          0.1947
## Adjusted R2              0.2343          0.1907
## Residual Std. Error      0.3883          0.3993
## F Statistic                            48.2537***
## ===============================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

h. [2 puntos] ¿A qué se deben las discrepancias que encuentra? ¿Cuál de las dos estrategias prefiere para estimar el modelo de variables instrumentales?

*De manera análoga a lo que pasó en la estimación del heckit a mano, el problema es que nuestro procedimiento de MC2E a mano no toma en cuenta que en la ecuación estructural estamos usando valores ajustados de la variable endógena. Las funciones en la mayoría de los paquetes utilizados en econometría calculan los errores estándar de manera correcta.*