MICROECONOMETRÍA APLICADA

David Bravo

Javiera Vásquez*

Agosto, 2008

* Agradezco a José Manuel Eguiguren la revisión y comentarios de este apunte. Cualquier comentario o sugerencia enviar correo electrónico a jvasquez@econ.uchile.cl.

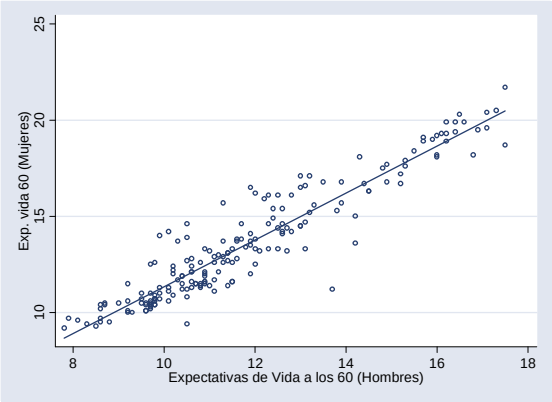
INDICE

Capítulo I. Modelo de regresión lineal

I.1. Introducción

Supongamos que estamos interesados en la relación entre dos variables. En este caso, un gráfico que nos permite examinar esta posibilidad es el diagrama de dispersión (scatterplot). El gráfico 1 muestra una asociación positiva entre las expectativas de vida de los hombres y de las mujeres a los 60 años para un conjunto de 188 países según datos de la Organización Mundial de la Salud. Como las observaciones están bastante cercanas a la recta de regresión, se puede decir que la correlación exhibida es fuerte. De hecho, el coeficiente de correlación es 0,94. Se puede apreciar, además, que la pendiente de la recta es mayor que 1. Esto ocurre porque en casi todos los países, las mujeres tienen expectativas de vida mayores que las de los hombres.

Gráfico 1



Nota:

- Primero se llaman los datos:

```
use who2001, clear
```

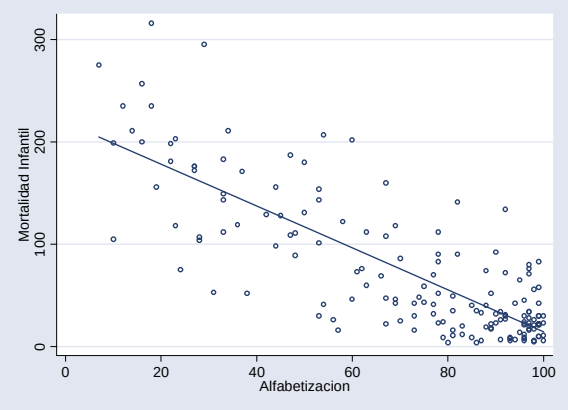
- Luego se usa el comando "graph twoway" con la siguiente instrucción:

```
graph twoway (lfit lex60_f lex60_m)(scatter lex60_f lex60_m, mstyle(p1) ms(oh) ),
xtitle("Expectativas de Vida a los 60 (Hombres)") ytitle("Exp. vida 60 (Mujeres)")
legend(off) name(g1, replace)
```

"lfit" permite agregar una línea de regresión entre las dos variables; "scatter" indica la modalidad del gráfico;

El Gráfico 2 exhibe un patrón de asociación negativo entre dos variables, en este caso, entre la mortalidad infantil y la alfabetización de las mujeres, a partir de datos de 162 países provenientes de UNICEF. El coeficiente de correlación es -0,80.

Gráfico 2

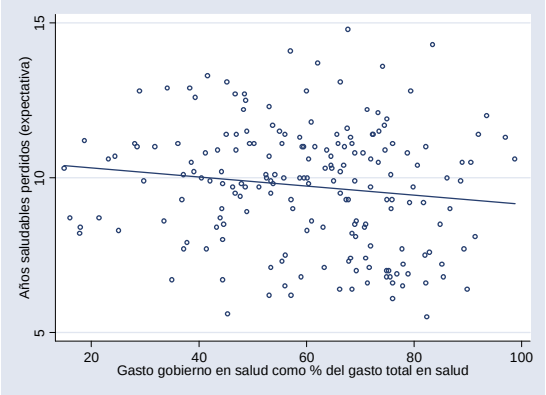


Nota:

```
use unicef, clear
graph twoway (lfit mort5_1999 literacy_f) (scatter mort5_1999 literacy_f,
mstyle(p1) ms(oh)), ytitle(Mortalidad Infantil) xtitle(Alfabetizacion) name(g2,
replace) legend(off)
```

Por otra parte, el Gráfico 3 es un ejemplo de la ausencia de relación evidente (o bien de relación débil) entre dos variables: en este caso se examina la asociación entre el gasto del gobierno en salud como proporción del gasto total en salud y la expectativa de años saludables perdidos. El coeficiente de correlación es de -0,13.

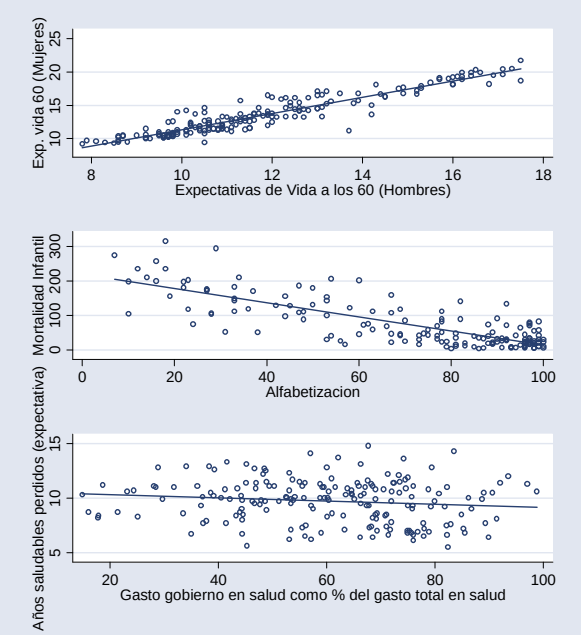
Gráfico 3



Nota:

```
use life, clear
graph twoway (lfit lhyb_m govexp)(scatter lhyb_m govexp, mstyle(p1) ms(oh)),
ytitle("Años saludables perdidos (expectativa)") xtitle("Gasto gobierno en salud
como % del gasto total en salud") name(g3, replace) legend(off)
```

Los tres gráficos anteriores se pueden poner juntos utilizando la siguiente instrucción: `graph combine g1 g2 g3, rows(3) ysize(6)`



Que el patrón de asociación encontrado, más o menos fuerte, no indica una relación de causalidad es algo claro. Por ejemplo, en el gráfico 1, la expectativa de vida de las mujeres no está causada, probablemente, por la expectativa de vida de los hombres. Y aún cuando pueda haber razones teóricas para pensar que hay una influencia causal de la alfabetización sobre la mortalidad infantil, ni el gráfico ni la recta de regresión pueden probar esta relación.

¿Cuál es la recta que mejor se ajusta a la nube de puntos en los gráficos anteriores?. El Método de Mínimos Cuadrados Ordinarios (MICO) elige los parámetros relevantes de manera de minimizar la suma de los errores al cuadrado.

Analicemos el ejemplo del Gráfico 2. La recta allí trazada supone lo siguiente:

$$Y_i = \alpha + \beta X_i + e_i$$

donde Y es un indicador de mortalidad infantil (definido como la probabilidad de morir entre el nacimiento y exactamente los 5 años, expresado por cada 1000 nacidos vivos) y la variable X representa la alfabetización de las mujeres (definido como la proporción de las mujeres de 15 y más años que pueden leer y escribir). En esta muestra n=162 (países).

Si se minimiza la suma de los errores al cuadrado:

$$\min \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

se obtienen los siguientes coeficientes estimados:

$$\hat{\alpha}^{MICO} = \overline{Y} - \hat{\beta}\overline{X}$$

$$\hat{\beta}^{MICO} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$

Para obtener el resultado en STATA hacemos lo siguiente:

```
regress mort5_1999 literacy_f

      Source |       SS       df       MS              Number of obs =     162
-------------+------------------------------           F(  1,   160) =  293.76
       Model |  501010.663     1  501010.663           Prob > F      =  0.0000
    Residual |  272883.071   160   1705.5192           R-squared     =  0.6474
-------------+------------------------------           Adj R-squared =  0.6452
       Total |  773893.735   161  4806.79338           Root MSE      =  41.298

------------------------------------------------------------------------------
  mort5_1999 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  literacy_f |  -2.046104   .1193802   -17.14   0.000    -2.281868   -1.810339
       _cons |   219.0556   9.071686    24.15   0.000     201.1399    236.9713
------------------------------------------------------------------------------
```

donde podemos apreciar que:

$$\hat{\alpha} = 219{,}06$$
$$\hat{\beta} = -2{,}046$$

es decir, sin alfabetización alguna, la mortalidad infantil en los países sería de 219 por cada 1000 nacidos vivos; pero, por otra parte, por cada punto porcentual de alfabetización de las madres, el índice de mortalidad infantil desciende en 2 por cada 1000 nacidos vivos.

Así, la recta en el gráfico 2 se puede representar por la siguiente ecuación:

$$\hat{Y}_i = 219{,}06 - 2{,}046 \cdot X_i$$

De esta manera, si la tasa de alfabetización fuera de 100%, la predicción para la mortalidad infantil sería de algo más de 14.

Note que ud. puede usar el comando "display" como una calculadora:

```
. display 219.0556-2.046104*100
14.4452
```

Además de los errores que se pueden cometer escribiendo números, está el hecho que los coeficientes que se exhiben en el output de STATA están redondeados. Una forma más simple de acceder a los coeficientes es utilizando lo que queda almacenado en STATA: "_b[*nombre de la variable*]"

```
. display _b[_cons]+_b[literacy_f]*100
14.44524
```

lo cual difiere ligeramente de lo calculado antes porque STATA es preciso hasta el decimal número 16.

Supongamos que queremos mirar los datos y examinar precisamente los casos en que la alfabetización femenina es de 100%. Ya sabemos que el valor predicho para la mortalidad es de 14.44524. ¿Cuál es el valor observado para la mortalidad?

```
. list mort5_1999 literacy_f if literacy_f==100

     +-------------------------+
     | mort5_1999   literacy_f |
     |-------------------------|
 64. |       23        100     |
 91. |       30        100     |
101. |       11        100     |
157. |        6        100     |
     +-------------------------+
```

Se puede ver que aparecen valores entre 6 y 30, que difieren del 14,4 esperado. Estas diferencias corresponden a los residuos. Es decir, los residuos (e) corresponden a las desviaciones entre el Y observado y el Y predicho por la recta de regresión:

$$e_i = Y_i - \hat{Y}_i = Y_i - 219{,}06 - 2{,}046 \cdot X_i$$

Si se quiere computar el valor predicho para cada país en la base de datos, se pueden usar los coeficientes de la regresión que quedan grabados en STATA. Es decir, para guardar el valor predicho en una variable que llamaremos ygorro, habría que escribir:

```
. generate ygorro=_b[_cons]+_b[literacy_f]*literacy_f
```

Otra forma de obtener el mismo resultado es escribir, después de haber corrido la regresión:

```
. predict ygorro
```

con lo que se crea una variable "ygorro" que contiene los valores predichos de la variable dependiente de la regresión.

Los residuos de la regresión se pueden obtener, entonces, fácilmente:

```
. generate residuo=mort5_1999-ygorro
```

Alternativamente, también se podría haber usado el mismo comando "predict" con la opción "resid", que obtiene lo que necesitamos:

```
. predict residuo, resid
```

La suma de los residuos al cuadrado está denotada en el output de STATA por residual SS y asciende a 272883,071, lo que equivale a:

$$\sum_{i=1}^{N} e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

Se puede demostrar que:

$$\sum_{i=1}^{N} (Y_i - \bar{Y}_i)^2 = \hat{\beta}^2 \sum_{i=1}^{N} (X_i - \bar{X})^2 + \sum_{i=1}^{N} e_i^2$$

La última expresión dice que la suma de los cuadrados de la variable dependiente (Total SS en el output de STATA, e igual a 773893,735) es igual a una fracción determinada por la varianza de la variable independiente (Model SS en el output anterior, e igual a 501010,663) más la suma de los errores al cuadrado (residual SS=272883,071).

Todos estos resultados, denominados ANOVA (Analysis of variance) aparecen en la parte superior izquierda del output de STATA en una regresión. Asimismo, el término "df" denota los grados de libertad y "MS" la suma de cuadrados medios que equivale a la columna SS dividida por df.

El output que aparece en el borde superior derecho contiene el número de observaciones: 162 (países) en el caso que nos ocupa.

Adicionalmente, contiene una medida de la calidad del ajuste de la regresión. Para ello utiliza el coeficiente de determinación o R^2:

$$R^2 = \frac{SS(Model)}{SS(Total)} = 1 - \frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2} = 1 - \frac{SS(\text{Re}sidual)}{SS(Total)}$$

En nuestro ejemplo R^2 =0,6474, indicando que la tasa de alfabetización femenina explica casi el 65% de la varianza en la mortalidad infantil.

Alternativamente, también se proporciona el R^2 ajustado por grados de libertad, que puede ser de utilidad cuando la regresión tiene pocas observaciones. En el ejemplo que analizamos este indicador es de 0,6452, muy parecido al R^2 no ajustado, porque estamos usando sólo una variable independiente.

$$\overline{R}^2 = 1 - \frac{\sum_{i=1}^{N} e_i^2 / (n-k)}{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 / (n-1)}$$

Otras dos filas del output en el mismo sector se refieren al test F. En el ejemplo, aparece "F(1,160)" y también "Prob>F". El valor F(1,160) corresponde al valor del test para la hipótesis nula de que todos los coeficientes son cero salvo la constante. Este test tiene dos grados de libertad: el número de restricciones (en este caso, corresponde solo a 1, la pendiente es igual a cero); y el número de observaciones menos el número de parámetros estimado (n-2). Por ello, el test tiene grados de libertad (1 y 160).

Se puede mostrar que este test corresponde a:

$$\frac{\frac{R^2}{k-1}}{\frac{1-R^2}{n-k}} \sim F_{(k-1,n-k)}$$

y en el ejemplo, toma un valor de 293,76. Dicho de otro modo, esta prueba de hipótesis nos permite testear si el R^2 calculado con la muestra es significativamente distinto de 0.

Otra medida de ajuste del modelo, alternativa al coeficiente de determinación, es la raíz del Error Cuadrático Medio:

$$\sqrt{ECM} = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{n-k}}$$

Este estadístico se expresa en la misma unidad que la variable dependiente. En el ejemplo del output de STATA, la raíz del Error Cuadrático Medio se denota por "Root MSE" y el valor es de 41,298. Esto puede ser interpretado como que, en promedio, nuestro modelo, al predecir la mortalidad infantil, se equivoca en 41 muertes por 1000 nacidos vivos.

Miremos ahora, con más detención, la sección de resultados de STATA en la que aparecen los coeficientes estimados.

```
------------------------------------------------------------------------------
  mort5_1999 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  literacy_f |  -2.046104   .1193802   -17.14   0.000    -2.281868   -1.810339
```

```
       _cons |   219.0556   9.071686    24.15   0.000     201.1399    236.9713
------------------------------------------------------------------------------
```

En la columna "Std.Err" se reporta el error estándar de los coeficientes estimados. Recuérdese que los coeficientes estimados (intercepto y pendiente, en este caso) provienen de una muestra aleatoria obtenida de una población y, por lo tanto, son estimadores de los valores equivalentes a nivel poblacional (los que son desconocidos). Si fuera posible obtener una segunda muestra aleatoria de la misma población, obtendríamos distintos estimadores. Hay, por lo tanto, una variabilidad inherente en los coeficientes que estimamos a partir de una muestra, y un indicador de esta variabilidad está dado por los errores estándar de los coeficientes, que simplemente equivalen a las desviaciones estándar muestrales de ellos.

Nótese que, en el caso de una regresión bivariada, la varianza del estimador es:

$$V(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

Sin embargo, como no conocemos el parámetro σ^2, la varianza estimada del coeficiente debe utilizar un estimador insesgado de σ^2, es decir: s^2, la suma de los residuos al cuadrado dividido por los grados de libertad.

$$\hat{V}(\hat{\beta}) = \frac{s^2}{\sum_{i=1}^{N}(X_i - \overline{X})^2} = \frac{\frac{\sum_{i=1}^{N} e_i^2}{n-k}}{\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

$$\text{Error Estándar de } \hat{\beta} = EE(\hat{\beta}) = \sqrt{\frac{\frac{\sum_{i=1}^{N} e_i^2}{n-k}}{\sum_{i=1}^{N}(X_i - \bar{X})^2}}$$

Los resultados muestran un error estándar de 0,119 para el coeficiente asociado a la variable alfabetización femenina cuyo valor estimado es de -2,046; el error estándar del intercepto (la constante) es de 9,07 (el valor estimado de la constante fue de 219,06).

Después del error estándar, la columna siguiente de resultados muestra un test t para la hipótesis nula de que cada coeficiente, individualmente, es igual a cero. El estadístico estimado así, tiene una distribución t de student con n-k grados de libertad y corresponde a lo siguiente:

$$\frac{\hat{\beta}}{EE(\hat{\beta})} \sim t_{n-k}$$

Se puede apreciar que este estadístico corresponde al cuociente de las columnas anteriores: la del coeficiente estimado y el error estándar respectivo. Para evaluar la hipótesis nula de que cada coeficiente es distinto de cero, se debe comparar el valor t estimado con el correspondiente en la tabla t de student para un determinado nivel de significancia. Para más de 30 grados de libertad, para un nivel de significancia de 5% (y un test de dos colas), el valor crítico según la tabla t-student es de -1.96 y +1.96. Por lo tanto, si el valor t estimado es mayor que 1.96 o bien menor que -1.96, entonces se rechaza la hipótesis nula.

La información proporcionada por el cuadro de resultados indica que un estadístico t de -17 para el coeficiente de la variable alfabetización femenina. De esta manera, se rechaza la hipótesis nula de que el coeficiente de esta variable es cero. O, dicho de otro modo, se concluye que esta variable es estadísticamente significativa. Lo mismo se concluye para la constante o intercepto del modelo estimado.

Una forma alternativa de examinar esta evidencia es mirar la columna "P>|t|", que contiene el p-value del test. Es decir, esta columna nos indica cuán probable es observar un valor del coeficiente como el obtenido de la regresión si es que el coeficiente verdadero fuera el de la hipótesis nula (es decir, cero). En general, un valor pequeño (típicamente menor a 0,05) nos indica que es improbable observar este valor si es que el valor verdadero poblacional es cero. Para el ejemplo en cuestión, para ambos parámetros el p-value es de 0.

Por último, tan importante como obtener estimaciones puntuales de los coeficientes, es obtener intervalos de confianza para los mismos. Se puede demostrar que un intervalo de confianza del 95% para el coeficiente β toma la siguiente forma:

$$\left[\hat{\beta} - 1.96 \cdot EE(\hat{\beta}), \hat{\beta} + 1.96 \cdot EE(\hat{\beta})\right]$$

donde, como ya se vio, 2.042 corresponde al valor crítico para el 95% de confianza para un test de dos colas, cuando el tamaño muestral es mayor de 30 observaciones.

En el ejemplo que se analiza, el intervalo de confianza es -2,28 a -1,81. Es decir, si se extrajeran muchas muestras de una población, con un 95% de probabilidad el coeficiente de alfabetización femenina verdadero (a nivel poblacional) estaría contenido en el rango de valores entre -2,28 y -1,81.

Debe recordarse que los errores estándar, test t y los intervalos estimados, para ser apropiados, requieren del cumplimiento de los supuestos del modelo lineal clásico: correcta especificación del modelo y linealidad del mismo; término de error de media cero, varianza igual y constante para todas las observaciones y covarianza de cero entre los errores de dos distintas observaciones; además de cero covarianza entre el término de error del modelo y las variables independientes. Sin embargo, no basta lo anterior. En efecto, las propiedades antes enunciadas aseguran que el estimador de Mínimos Cuadrados Ordinarios es el mejor estimador dentro de la categoría de estimadores lineales e insesgados (Teorema de Gauss-Markov). El uso de los intervalos, errores estándar y test t indicados anteriormente requiere de un supuesto adicional que es que el término de error del modelo tenga **distribución normal** (alternativamente, se requeriría que la muestra utilizada sea suficientemente grande como para que se aplique el hecho que, asintóticamente, las distribuciones tienden a ser normales).

I.3.Regresión múltiple

El modelo de regresión múltiple extiende el modelo simple para incorporar k variables independientes.

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \mu_i$$

El comando de STATA es el mismo anterior. Lo importante de recordar es que siempre se coloca en primer lugar la variable dependiente, y luego las variables explicativas. Todos los otros comandos revisados anteriormente se pueden aplicar de igual forma. Los resultados del output de STATA también son los mismos ya analizados. Sólo debe considerarse que la interpretación de los coeficientes estimados corresponde al cambio en la variable dependiente ante un cambio en la variable independiente, manteniendo constantes las otras variables.

Una de los modelos más estimados en los análisis empíricos es aquél que busca explicar los **determinantes de los salarios en el mercado laboral**.

En este caso, con el objeto de tener datos comparables para personas que laboran distintas horas, se utiliza típicamente como variable dependiente el salario por hora trabajada (yph). Dentro de las variables explicativas más utilizadas están distintas medidas de capital humano, siendo la más importante los años de escolaridad del individuo (esc).

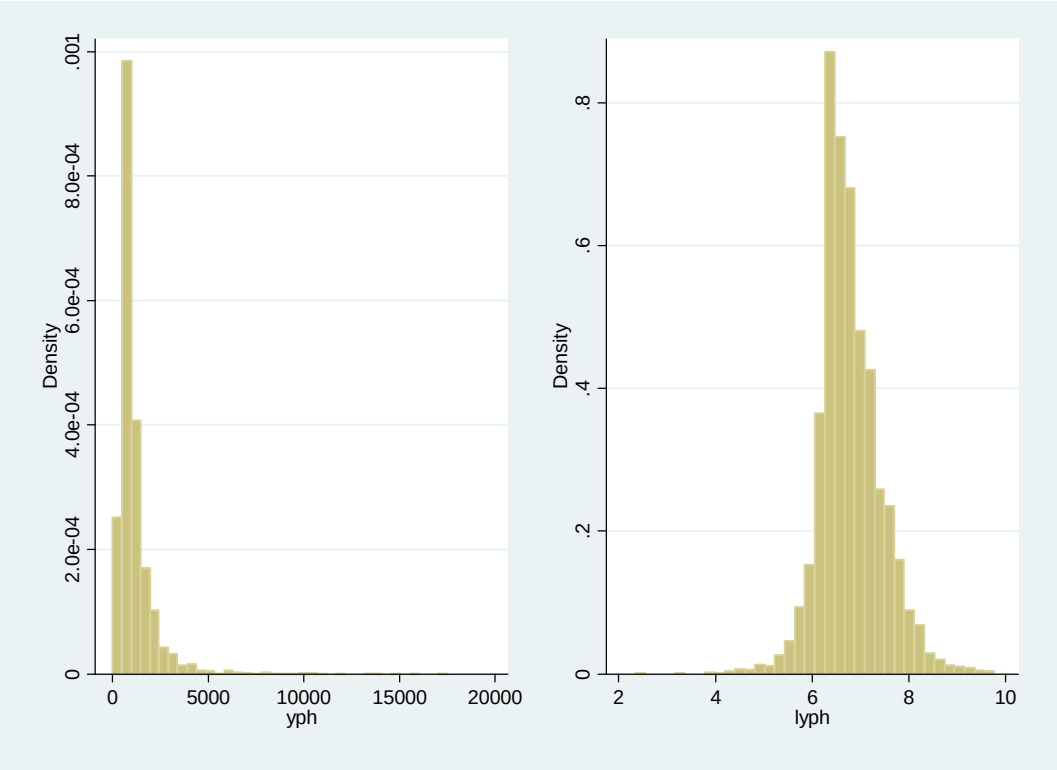
Uno de los parámetros que más interesan es, precisamente, el denominado retorno a la educación, que indica cuánto aumenta porcentualmente el ingreso como resultado de incrementar la escolaridad en un año:

$$\text{Retorno a la educación: } \frac{\Delta\% yph}{\Delta esc}$$

La estimación de este modelo se facilita si como variable dependiente se utiliza el logaritmo del salario por hora y no el salario por hora, puesto que el cambio en el logaritmo del ingreso por hora corresponde aproximadamente al cambio porcentual en el ingreso por hora. De allí que el retorno a la educación pueda obtenerse directamente como el coeficiente de la variable escolaridad en la regresión.

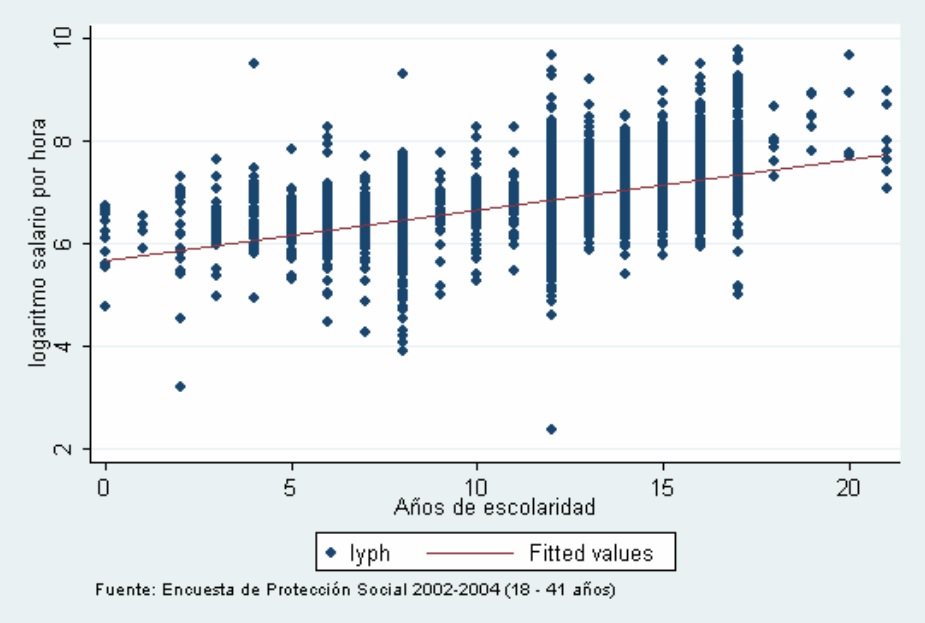
La transformación logarítmica anterior es también conveniente porque la distribución empírica del logaritmo del ingreso es más cercana a una distribución normal que la distribución del ingreso.

Para estimar este modelo contamos con una muestra de 7.312 personas entrevistadas en la Encuesta de Protección Social[1], que en el año 2004 tenían entre 18 y 41 años. Se tomo este universo de personas ya que en la encuesta se pregunta por la historia laboral de las personas desde 1980. De esta forma, las personas mayores de 41 años en el año 2004 reportan una historia laboral censurada, la cual no nos permite obtener una medida apropiada de los años trabajados.



Para esta muestra se tiene la siguiente relación entre ingreso laboral por hora y años de escolaridad:

[1] Para mayores antecedentes de esta encuesta visite www.proteccionsocial.cl.



Nota:

```
use "ingresos_esp(18-41).dta", clear
twoway (scatter lyph esc04), ytitle(logaritmo salario por hora)
xtitle(Años de escolaridad) note(Fuente: Encuesta de Protección
Social 2002-2004 (18 - 41 años)) || lfit lyph esc04
```

Sin embargo, la escolaridad no es la única variable relevante para explicar el salario por hora. En efecto, al realizar una estimación MCO simple del logaritmo del salario por hora y los años de escolaridad, se obtiene el siguiente resultado:

```
. regress lyph esc04 [w=factor]
(analytic weights assumed)
(sum of wgt is   2.9330e+06)

      Source |       SS       df       MS              Number of obs =    4678
-------------+------------------------------           F(  1,  4676) = 1515.56
       Model |  525.129857     1  525.129857            Prob > F      =  0.0000
    Residual |  1620.19669  4676  .346492022            R-squared     =  0.2448
-------------+------------------------------           Adj R-squared =  0.2446
       Total |  2145.32655  4677  .458697146           Root MSE      =  .58864

------------------------------------------------------------------------------
        lyph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       esc04 |   .1064952   .0027355    38.93   0.000     .1011322    .1118581
       _cons |   5.574077   .0334675   166.55   0.000     5.508465    5.639689
------------------------------------------------------------------------------
```

De lo anterior se aprecia que la escolaridad puede explicar sólo un **24,5%** de la varianza total del salario por hora.

Cuando la variable dependiente está medida en logaritmos y la variable explicativa en nivel, el coeficiente estimado representa una semi-elasticidad, como se indicó anteriormente. En este caso, el coeficiente estimado para la variable escolaridad (esc04) mide el impacto en cambio porcentual de la variable dependiente asociado al cambio en una unidad de la variable explicativa. En este caso, el coeficiente tiene un valor de 0.1065; es decir, si la escolaridad aumenta en un año el salario se incrementa en **10,65%**. Lo que representa exactamente el retorno de la educación.

Otra variable importante para explicar el salario por hora es la experiencia laboral de la persona. Debido a falta de medidas precisas de esta variable, usualmente se aproxima la experiencia laboral como la edad menos la escolaridad menos 6. Sin embargo, la EPS entrega información autoreportada de los periodos de tiempo en que la persona ha estado trabajando. Esta información nos permite computar una medida más confiable de experiencia. A continuación se presenta el resultado del modelo que incorpora, además de la escolaridad, la experiencia laboral efectiva (medida en meses):

```
. regress lyph experiencia esc04 [w=factor]
(analytic weights assumed)
(sum of wgt is   2.9330e+06)

      Source |       SS       df       MS              Number of obs =    4678
-------------+------------------------------           F(  2,  4675) =  822.19
       Model |  558.239095     2  279.119548            Prob > F      =  0.0000
    Residual |  1587.08746  4675  .339483948           R-squared     =  0.2602
-------------+------------------------------           Adj R-squared =  0.2599
       Total |  2145.32655  4677  .458697146           Root MSE      =  .58265

------------------------------------------------------------------------------
        lyph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  experiencia |    .001138   .0001152     9.88   0.000     .0009121    .0013639
       esc04 |   .1127432   .0027807    40.55   0.000     .1072918    .1181946
       _cons |   5.352633   .0400028   133.81   0.000     5.274209    5.431057
------------------------------------------------------------------------------
```

Ambas variables explican un 26% (R^2) del comportamiento del salario por hora. Adicionalmente, podemos observar lo siguiente:

- El coeficiente de la variable ***experiencia*** indica que, todo lo demás constante, un mes adicional de experiencia aumenta en un 0,114% el salario por hora; o, alternativamente, un año adicional de experiencia aumenta en un 1.37% el salario por hora (0,114%*12).
- El retorno a la educación estimado cambió con respecto a la estimación anterior. Ahora un año adicional de educación aumenta en un 11,3% el salario por hora. Esto nos muestra las consecuencias de la **omisión de variables relevantes** en la estimación del modelo. Al omitir la variable experiencia (primer modelo) el retorno a la educación estaba siendo subestimado (esto, porque la experiencia tiene una correlación negativa con la escolaridad). En general, la omisión de variables relevantes provoca un sesgo en la estimación MCO, el sesgo tiene la siguiente forma:

$$\hat{\beta}_{\text{esc}} = \beta_{\text{esc}} + \underbrace{\frac{Cov(\text{esc},\text{experiencia})}{V(\text{esc})}\beta_{\text{experiencia}}}_{\text{SESGO}}$$

Tal como se vio en la sección anterior, el output de STATA entrega los estadísticos para ver la significancia individual del modelo, los que nos indican que ambas variables son estadísticamente significativas (p-values igual a cero). Además se puede ver que el modelo es globalmente significativo a través del estadístico F (p-value es cero).

Test de hipótesis

Para testear otras hipótesis lineales, por ejemplo, de combinación de parámetros, podemos utilizar el comando **`lincom`**, inmediatamente después de haber realizado la regresión.

Por ejemplo, si nos interesa ver cuántos meses de experiencia compensan un año de escolaridad, podemos ver por la magnitud de los coeficientes que aproximadamente 100 meses de experiencia equivalen a un año de escolaridad. Podríamos querer testear si esta hipótesis es estadísticamente significativa:

$$H_0 : \beta_{esc} - 100 \cdot \beta_{\exp eriencia} = 0$$

```
. lincom esc04-100*experiencia
 ( 1)  esc04 - 100 experiencia = 0
------------------------------------------------------------------------------
        lyph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0010572   .0112222    -0.09   0.925    -.0230581    .0209437
------------------------------------------------------------------------------
```

Como se puede apreciar, no se puede rechazar la hipótesis nula de que 100 meses de experiencia equivalen a un año de escolaridad en el impacto sobre el salario por hora.

Por otra parte, para testear hipótesis conjuntas se debe utilizar el comando **test**, inmediatamente después de haber realizado la regresión. Por ejemplo, para testear conjuntamente que el retorno a la educación es de 11% y que, además, el retorno a la experiencia es igual a 0,11% hacemos:

$$H_0: \quad \beta_{\text{esc}} = 0.11$$
$$100 \cdot \beta_{\text{experiencia}} = 0.11$$

```
. test (esc04=0.11) (100*experiencia=0.11)

( 1)  esc04 = .11
( 2)  100 experiencia = .11

       F(  2,  4675) =    0.49
            Prob > F =    0.6111
```

I.5.Predicción

Una vez estimado el modelo podemos utilizar los coeficientes para predecir la variable de interés y hacer recomendaciones de política. Se pueden hacer dos tipos de predicciones: para el valor puntual de la variable dependiente y para el valor esperado de la variable dependiente. En ambos casos la predicción propiamente tal será la misma, siendo la única diferencia el error de predicción que se comete en cada una de ellas. Obviamente, al tratar de predecir un valor puntual de la variable dependiente el error que se comente es mayor, por la naturaleza aleatoria intrínseca de nuestra variable de interés; así, la varianza del error de predicción también será mayor, y los intervalos de confianza más amplios.

Lo pasos son los siguientes:

1- Obtener la estimación MCO de los parámetros de modelo $\longrightarrow \hat{\beta}$

2- Decidir para qué valores de las variables explicativas queremos predecir. Por ejemplo, supongamos que queremos ver cuál es logaritmo del salario para una persona con 12 años de escolaridad y 120 meses de experiencia. O bien podríamos ir tomando distintas combinaciones de estas variables y obtener la predicción; incluso se puede tomar otra muestra de individuos y predecir su salario si tenemos las observaciones de experiencia y escolaridad.

$$\text{Valor puntual} : \hat{Y}_0 = X_0\hat{\beta}$$
$$\text{Valor esperado} : E[\hat{Y}_0|X_0] = X_0\hat{\beta}$$

I.5.1. Predicción de un valor puntual:

Error de predicción:

$$\hat{e}^0 = y^0 - \hat{y}^0$$

$$\hat{e}^0 = x^0\beta + u - x^0\hat{\beta}$$

$$\hat{e}^0 = x^0 \underbrace{(\beta - \hat{\beta})}_{\text{error de estimación}} + \underbrace{u}_{\text{término error típico del modelo}} \longrightarrow E[\hat{e}^0] = 0$$

Varianza del error de predicción:

$$\hat{V}[\hat{e}^0] = \hat{\sigma}^2(1 + x^0(X'X)^{-1}x^{0\prime})$$

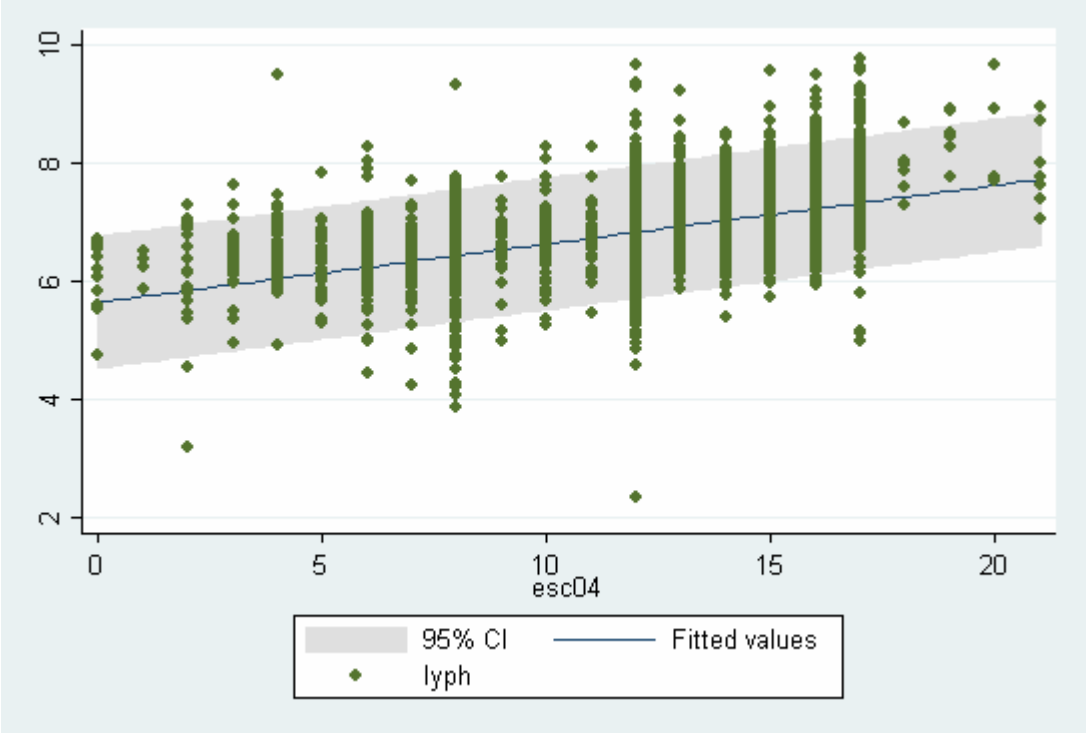
Intervalo de confianza de la predicción:

$$\Pr\left[\hat{y}^0 - t_{n-k,1-\alpha/2}\sqrt{\hat{V}[\hat{e}^0]} \le y^0 \le \hat{y}^0 + t_{n-k,1-\alpha/2}\sqrt{\hat{V}[\hat{e}^0]}\right] = 1-\alpha$$

Para realizar un gráfico entre dos variables e incluir, además de la recta de regresión lineal (predicción), un intervalo de confianza, se debe ejecutar el siguiente comando:

```
twoway (lfitci lyph esc04, ciplot(rarea) stdf) || scatter lyph esc04
```

(**stdf** indica que el intervalo se debe hacer considerando que la predicción es de un valor puntual).



Se puede ver que por defecto STATA asume que se trata de un intervalo de 95% de confianza (CI=confidence interval). Se puede usar otros intervalos utilizando las opciones del comando.

I.5.2. Predicción del valor esperado:

Error de predicción:

$$\tilde{e}^0 = E[y^0 \mid x^0] - \hat{y}^0$$

$$\downarrow$$

$$\tilde{e}^0 = x^0\beta - x^0\hat{\beta}$$

$$\downarrow$$

$$\tilde{e}^0 = x^0 \underbrace{(\beta - \hat{\beta})}_{\text{error de estimación}} \longrightarrow E[\tilde{e}^0] = 0$$

Varianza del error de predicción:

$$\hat{V}[\tilde{e}^0] = \hat{\sigma}^2 x^0 (X'X)^{-1} x^{0\prime}$$

Intervalo de confianza de la predicción:

$$\Pr\left[\hat{E}[y^0 \mid x^0] - t_{n-k,1-\alpha/2}\sqrt{\hat{V}[\tilde{e}^0]} \le E[y^0 \mid x^0] \le \hat{E}[y^0 \mid x^0] + t_{n-k,1-\alpha/2}\sqrt{\hat{V}[\tilde{e}^0]}\right] = 1-\alpha$$

En este caso el error de predicción es menor, la varianza menor, y el intervalo de confianza de la predicción es más pequeño.

Para realizar un gráfico entre dos variables e incluir, además de la recta de regresión lineal (predicción), un intervalo de confianza, se debe ejecutar el siguiente comando:

```
twoway (lfitci lyph esc04, ciplot(rarea) stdp) || scatter lyph esc04
```

(**stdp** indica que el intervalo se debe hacer considerando que la predicción es del valor esperado).

I.6. Test de Normalidad

Tal como se mencionó al término de la segunda sección, el estimador MCO requiere del cumplimiento de varios supuestos, para que este estimador sea el mejor (más eficiente) estimador dentro de los estimadores lineales e insesgados:

- El modelo debe ser lineal
- Los errores se deben distribuir en forma independiente y con idéntica distribución $\mu_i \sim iid(0,\sigma^2)$.
- Las variables explicativas no deben ser colineales
- No debe existir relación entre las variables explicativas y el término de error: $Cov(\mu_i,X_i)=0$.
- Modelo debe estar bien especificado.

En la siguiente clase recordaremos cuáles son las consecuencias de que algunos de estos supuestos no se cumplan.

Para que el estimador MCO cumpla con la propiedad MELI, no se requiere que el error tenga una distribución normal. Sin embargo, para la realización de tests de hipótesis y la construcción de los intervalos de confianza, este supuesto es fundamental, ya que si los errores no son normales, las distribuciones de los estadísticos vistos no son t y F, sino desconocidas.

Una vez estimado el modelo, podemos obtener los errores de estimación, equivalentes a la diferencia entre el valor observado de la variable dependiente y su valor estimado, utilizando el siguiente comando post estimación:

```
predict errores, resid
```

con el cual se genera una nueva variable *errores*, que contiene los errores del modelo recién estimado.

Recordemos que los parámetros que definen la normalidad de una distribución son el coeficiente de asimetría (skewness) y el coeficiente de kurtosis, los que deben tomar los valores 0 y 3 para que la distribución sea normal. Veamos qué valores para estos coeficientes tiene la variable *errores* recién creada:

```
. tabstat errores [w=factor], stats(kurtosis skewness)
(analytic weights assumed)

    variable |  kurtosis  skewness
-------------+--------------------
     errores |   5.96397  .1638036
---------------------------------
```

Podemos apreciar que el coeficiente de asimetría es distinto de cero, y el de kurtosis distinto de 3, pero no sabemos si esta diferencia es estadísticamente significativa. Para indagar en este punto debemos realizar un test de hipótesis. En STATA el comando **sktest** realiza tres acciones: i) un test de hipótesis para la hipótesis nula de que el coeficiente de asimetría es cero; ii) un test de hipótesis de que el coeficiente de kurtosis es igual a tres; y iii) ambas hipótesis en forma conjunta. Los resultados para los errores de este modelo se presentan a continuación:

```
.  sktest errores, noad

                    Skewness/Kurtosis tests for Normality
                                                    ------- joint ------
    Variable |  Pr(Skewness)   Pr(Kurtosis)      chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
     errores |      0.113         0.000          344.39       0.0000
```

Como se puede apreciar, no se puede rechazar la hipótesis nula de simetría, pero si de que la kurtosis es igual a 3. En conjunto, se rechaza la hipótesis nula de normalidad de los errores.

I.7. Boostrap para la obtención de intervalos de confianza

Cuando los errores del modelo no son normales, los estadísticos para los test de hipótesis simple y conjunto no siguen las distribuciones conocidas t y F, respectivamente. En este caso, la distribución de los errores es desconocida. Podemos en este caso utilizar la metodología de simulación bootstrap para computar los intervalos de confianza.

El comando en STATA para realizar esto es:

```
bootstrap, reps(500) bca: reg lyph esc04 experiencia
```

Con este comando le indicamos que haga la regresión de interés 500 veces. El resultado de ejecutar este comando es:

```
Linear regression                               Number of obs      =      4678
                                                Replications       =       500
                                                Wald chi2(2)       =   1234.11
                                                Prob > chi2        =    0.0000
                                                R-squared          =    0.2486
                                                Adj R-squared      =    0.2483
                                                Root MSE           =    0.5593

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
        lyph |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       esc04 |   .1047146   .0029875    35.05   0.000     .0988591    .1105701
 experiencia |   .0011328   .0001174     9.65   0.000     .0009027    .0013629
       _cons |   5.429254   .0409256   132.66   0.000     5.349041    5.509467
------------------------------------------------------------------------------
```

Luego ejecutando el siguiente comando, obtenemos los intervalos de confianza en que nos debemos fijar:

```
estat bootstrap, all
```

```
Linear regression                               Number of obs      =      4678
                                                Replications       =       500

------------------------------------------------------------------------------
             |    Observed               Bootstrap
        lyph |       Coef.       Bias    Std. Err.  [95% Conf. Interval]
-------------+----------------------------------------------------------------
       esc04 |    .1047146    .000138   .00298754    .0988591   .1105701   (N)
             |                                      .0994602   .1105084   (P)
             |                                      .0987785   .1101035  (BC)
             |                                      .0987785   .1101035 (BCa)
 experiencia |   .00113281  -2.40e-06   .00011739    .0009027   .0013629   (N)
             |                                      .0009084   .0013524   (P)
             |                                      .0009099   .0013591  (BC)
             |                                      .0009088   .0013591 (BCa)
       _cons |   5.4292541  -.0008044   .04092565    5.349041   5.509467   (N)
             |                                      5.352264   5.511744   (P)
             |                                      5.356518   5.513448  (BC)
             |                                      5.356518   5.514764 (BCa)
------------------------------------------------------------------------------
(N)    normal confidence interval
(P)    percentile confidence interval
(BC)   bias-corrected confidence interval
(BCa)  bias-corrected and accelerated confidence interval
```

Capítulo II. Modelo de regresión lineal: especificación y problemas

II.1. Introducción

En el Capítulo I se revisó el estimador de Mínimos Cuadrados Ordinarios (MCO) en el contexto de un modelo de regresión simple (solo una variable explicativa) y un modelo de regresión múltiple (más de una variable explicativa). Si los supuestos del estimador MCO se cumplen, este es el mejor estimador lineal insesgado. También se abordaron los test de hipótesis lineal simple y conjunto, y los intervalos de confianza tanto de los parámetros como de las predicciones que se pueden realizar a través de la estimación. En ambos casos, tanto para realizar inferencias como para computar los intervalos de confianza, el supuesto de normalidad del término de error es fundamental. Si este supuesto no se cumple, la inferencia realizada no es válida. En este caso, se deben utilizar métodos de simulaciones para obtener los intervalos de confianza correctos y realizar la inferencia en forma apropiada.

Otro de los supuestos claves de estimador MCO es que el modelo debe estar correctamente especificado. Esto significa que debemos hacer todos los esfuerzos (considerando la disponibilidad de datos) para incorporar todas las variables relevantes para explicar el comportamiento de la variable de interés (variable dependiente), y de la mejor forma posible. Algunas de las variables claves para explicar el comportamiento de la variable dependiente pueden ser discretas, no continuas; estas generalmente son variables de carácter cualitativo: género, zona geográfica, estatus laboral, etc. Es importante incorporar la información que aportan estas variables en forma correcta en la especificación para obtener una estimación adecuada de los impactos.

Otro problema de especificación es la omisión de variables relevantes, cuando una variable es omitida esta forma parte del término de error. Si la variable omitida tiene

correlación con una o más de las variables explicativas del modelo, la estimación MCO será sesgada. No se cumple uno de los supuestos claves: $COV(\mu_i, X_i) = 0$.

Por otra parte, con el objetivo de evitar el problema de omisión de variables, se pueden incluir variables irrelevantes. En este caso no se genera sesgo en la estimación MCO, pero se pierde eficiencia (el estimador tiene mayor varianza, es menos preciso). Un tipo de variable omitida son aquellas que ayudan a explicar un comportamiento no lineal de la variable dependiente, en estos casos las variables omitidas son potencias de las mismas variables explicativas ya incluidas en el modelo.

Otros supuesto, para la correcta especificación del modelo, es que las variables explicativas no sean colineales entre ellas. Es decir, se deben incluir variables explicativas que no sean muy parecidas o que no expliquen de igual forma el comportamiento de la variable dependiente. Cuando las variables explicativas son muy parecidas, se habla del problema de multicolinealidad. Este problema, se detecta por "síntomas" que se observan en la estimación. No genera sesgo en la estimación, pero el problema es que la estimación es muy "volátil", poco robusta.

El supuesto de homocedasticidad del término de error, es un supuesto que raramente se cumple cuando se trabaja con datos de corte transversal. La ruptura de este supuesto no genera problema de sesgo, pero si de ineficiencia. Veremos cómo detectar y abordar el problema de heterocedasticidad (varianza del error no es constante).

Por último, una vez incorporadas todas las variables relevantes de la mejor forma, en forma binaria o considerando no linealidades, y habiendo detectado y abordado los problemas de multicolinealidad o heterocedasticidad presentes, es posible tener más de un modelo que explique el comportamiento de la variable de interés y que cumple con todos los requisitos de especificación. Entonces, ¿con cuál de los modelos

quedarse?. Existen test de modelos anidados y no anidados que lo ayudarán a tomar la decisión en estos casos.

De esta forma, en este segundo capítulo se revisarán los siguientes temas:

- Consecuencias de la omisión de variables relevantes
- Consecuencias de la inclusión de variables irrelevantes
- Multicolinealidad
- Variables cualitativas o categóricas como regresores: variables ficticias (dummies, binarias, etc.)
- Inclusión de no linealidades
- Heterocedasticidad
- Criterios de selección de modelos anidados y no anidados.

II.2. Aplicación: determinantes de los salarios en el mercado laboral

Una de los modelos más estimados en los análisis empíricos es aquél que busca explicar los determinantes de los salarios en el mercado laboral.

En este caso, con el objeto de tener datos comparables para personas que laboran distintas horas, se utiliza típicamente como variable dependiente el salario por hora trabajada (yph). Dentro de las variables explicativas más utilizadas están distintas medidas de capital humano, siendo la más importante los años de escolaridad del individuo (esc).

Uno de los parámetros que más interesan es, precisamente, el denominado retorno a la educación, que indica cuánto aumenta porcentualmente el ingreso como resultado de incrementar la escolaridad en un año:

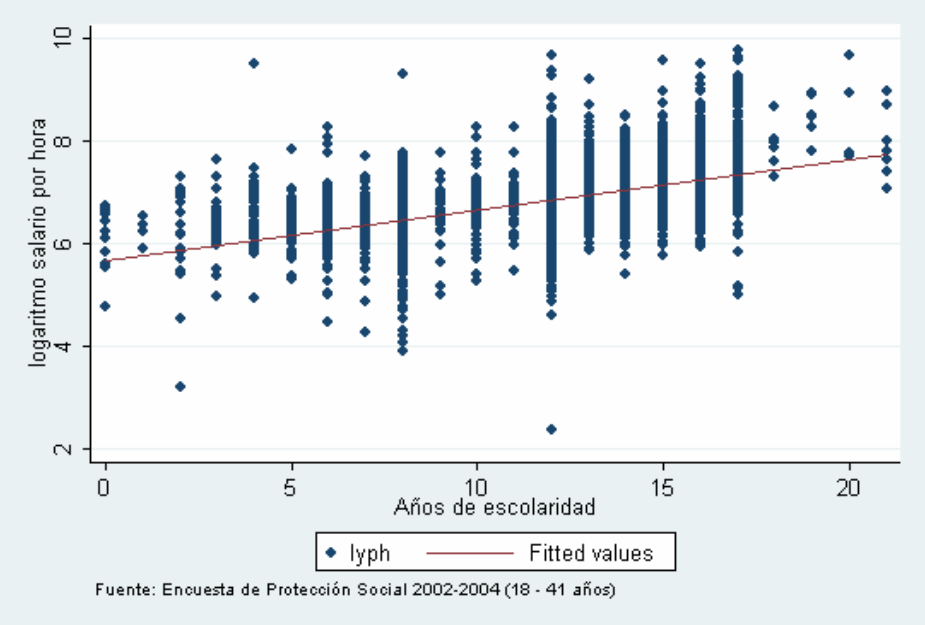
$$\text{Retorno a la educación: } \frac{\Delta\% yph}{\Delta esc}$$

La estimación de este modelo se facilita si como variable dependiente se utiliza el logaritmo del salario por hora y no el salario por hora, puesto que el cambio en el logaritmo del ingreso por hora corresponde aproximadamente al cambio porcentual en el ingreso por hora. De allí que el retorno a la educación pueda obtenerse directamente como el coeficiente de la variable escolaridad en la regresión.

La transformación logarítmica anterior es también conveniente porque la distribución empírica del logaritmo del ingreso es más cercana a una distribución normal que la distribución del ingreso.

Para estimar este modelo contamos con una muestra de 7.312 personas entrevistadas en la Encuesta de Protección Social[2], que en el año 2004 tenían entre 18 y 41 años. Se tomo este universo de personas ya que en la encuesta se pregunta por la historia laboral de las personas desde 1980. De esta forma, las personas mayores de 41 años en el año 2004 reportan una historia laboral censurada, la cual no nos permite obtener una medida apropiada de los años trabajados.

Para esta muestra se tiene la siguiente relación entre ingreso laboral por hora y años de escolaridad:



Nota:

```
use "ingresos_esp(18-41).dta", clear
twoway (scatter lyph esc04), ytitle(logaritmo salario por hora)
xtitle(Años de escolaridad) note(Fuente: Encuesta de Protección
Social 2002-2004 (18 - 41 años)) || lfit lyph esc04
```

[2] Para mayores antecedentes de esta encuesta visite www.proteccionsocial.cl.

Sin embargo, la escolaridad no es la única variable relevante para explicar el salario por hora. En efecto, al realizar una estimación MCO simple del logaritmo del salario por hora y los años de escolaridad, se obtiene el siguiente resultado:

```
. regress lyph esc04 [w=factor]
(analytic weights assumed)
(sum of wgt is   2.9330e+06)

      Source |       SS       df       MS              Number of obs =    4678
-------------+------------------------------           F(  1,  4676) = 1515.56
       Model |  525.129857     1  525.129857            Prob > F      =  0.0000
    Residual |  1620.19669  4676  .346492022           R-squared     =  0.2448
-------------+------------------------------           Adj R-squared =  0.2446
       Total |  2145.32655  4677  .458697146           Root MSE      =  .58864

------------------------------------------------------------------------------
        lyph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       esc04 |   .1064952   .0027355    38.93   0.000     .1011322    .1118581
       _cons |   5.574077   .0334675   166.55   0.000     5.508465    5.639689
------------------------------------------------------------------------------
```

De lo anterior se aprecia que la escolaridad puede explicar sólo un **24,5%** de la varianza total del salario por hora.

Cuando la variable dependiente está medida en logaritmos y la variable explicativa en nivel, el coeficiente estimado representa una semi-elasticidad, como se indicó anteriormente. En este caso, el coeficiente estimado para la variable escolaridad (esc04) mide el impacto en cambio porcentual de la variable dependiente asociado al cambio en una unidad de la variable explicativa. En este caso, el coeficiente tiene un valor de 0.1065; es decir, si la escolaridad aumenta en un año el salario se incrementa en **10,65%**. Lo que representa exactamente el retorno de la educación.
Otra variable importante para explicar el salario por hora es la experiencia laboral de la persona. Debido a falta de medidas precisas de esta variable, usualmente se aproxima la experiencia laboral como la edad menos la escolaridad menos 6. Sin embargo, la EPS entrega información autoreportada de los periodos de tiempo en que la persona ha estado trabajando. Esta información nos permite computar una medida más confiable de experiencia. A continuación se presenta el resultado del

modelo que incorpora, además de la escolaridad, la experiencia laboral efectiva (medida en meses):

```
. regress lyph experiencia esc04 [w=factor]
(analytic weights assumed)
(sum of wgt is   2.9330e+06)

      Source |       SS       df       MS              Number of obs =    4678
-------------+------------------------------           F(  2,  4675) =  822.19
       Model |  558.239095     2  279.119548            Prob > F      =  0.0000
    Residual |  1587.08746  4675  .339483948           R-squared     =  0.2602
-------------+------------------------------           Adj R-squared =  0.2599
       Total |  2145.32655  4677  .458697146           Root MSE      =  .58265

------------------------------------------------------------------------------
        lyph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  experiencia |    .001138   .0001152     9.88   0.000     .0009121    .0013639
       esc04 |   .1127432   .0027807    40.55   0.000     .1072918    .1181946
       _cons |   5.352633   .0400028   133.81   0.000     5.274209    5.431057
------------------------------------------------------------------------------
```

Ambas variables explican un 26% (R^2) del comportamiento del salario por hora. Adicionalmente, podemos observar lo siguiente:

- El coeficiente de la variable ***experiencia*** indica que, todo lo demás constante, un mes adicional de experiencia aumenta en un 0,114% el salario por hora; o, alternativamente, un año adicional de experiencia aumenta en un 1.37% el salario por hora (0,114%*12).
- El retorno a la educación estimado cambió con respecto a la estimación anterior. Ahora un año adicional de educación aumenta en un 11,3% el salario por hora. Esto nos muestra las consecuencias de la **omisión de variables relevantes** en la estimación del modelo. Al omitir la variable experiencia (primer modelo) el retorno a la educación estaba siendo subestimado (esto, porque la experiencia tiene una correlación negativa con la escolaridad). En general, la omisión de variables relevantes provoca un sesgo en la estimación MCO, el sesgo tiene la siguiente forma:

$$\hat{\beta}_{\text{esc}} = \beta_{\text{esc}} + \underbrace{\frac{Cov(\text{esc},\text{experiencia})}{V(\text{esc})}\beta_{\text{experiencia}}}_{\text{SESGO}}$$

Tal como se vio en la sección anterior, el output de STATA entrega los estadísticos para ver la significancia individual del modelo, los que nos indican que ambas variables son estadísticamente significativas (p-values igual a cero). Además se puede ver que el modelo es globalmente significativo a través del estadístico F (p-value es cero).

II.3. Omisión de variables relevantes

El problema de omisión de variables relevantes, tal como su nombre lo dice, se genera cuando una **variable relevante para explicar el comportamiento de la variable dependiente no ha sido incluida en la especificación**.

La omisión de variables relevantes provoca que el estimador MCO sea sesgado cuando dos condiciones se cumplen:

- La variable omitida esta correlacionada con los regresores incluidos
- La variable omitida explica el comportamiento de la variable dependiente, es decir, es una variable relevante.

Volvamos a la primera especificación (modelo simple) de la aplicación introducida en la primera clase:

$$(1) \quad \text{lyph}_i = \alpha + \beta_{\text{esc}} \cdot \text{esc}_i + \mu_i$$

Este modelo busca explicar el salario laboral por hora a través de los años de escolaridad. Sin embargo, tal como se vio la clase anterior, la variable experiencia también es relevante para explicar el salario por hora. Cuando se estimó el siguiente modelo, la variable experiencia resultó ser estadísticamente significativa:

$$(2) \quad \text{lyph}_i = \alpha + \beta_{\text{esc}} \cdot \text{esc}_i + \beta_{\text{exp}} \cdot \text{exp}_i + \mu_i$$

Así, si no es incluida en la especificación, se estaría omitiendo una variable relevante (segunda condición, la variable es significativa). Además, la siguiente matriz de correlación nos muestra, que la variable incluida (escolaridad) tiene una correlación negativa con la variable omitida (experiencia):

```
. correlate  experiencia esc04 if lyph!=. [w=factor]
(obs=4678)
             |    esc04 experi~a
-------------+------------------
       esc04 |   1.0000
 experiencia |  -0.2357   1.0000
```

Cuando la variable experiencia no es incluida, la variable forma parte término de error:

$$lyph_i = \beta_0 + \beta_{esc} \cdot esc04 + \underbrace{\varepsilon_i}_{\mu_i + \beta_{\text{exp}} \cdot \text{exp}}$$

Si la covarianza entre escolaridad y experiencia es distinta de cero, esto generará que el término de error de este modelo este correlacionado con la variable explicativa, rompiendo con uno de los supuestos del estimador MCO.

La omisión de la variable (relevante) experiencia genera el siguiente sesgo en la estimación MCO del retorno a la educación (coeficiente de la variable incluida en el modelo):

$$E\left[\hat{\beta}_{esc} \mid esc_i\right] = \beta_{esc} + \underbrace{\overbrace{\frac{Cov(esc, \exp)}{V(esc)}}^{-} \cdot \underbrace{\beta_{\exp}}_{+}}_{-}$$

Dado la covarianza negativa entre escolaridad y experiencia (mientras mayor es el tiempo dedicado a estudiar, en promedio, menos tiempo se ha participado en el mercado laboral), y a que la experiencia tiene un impacto positivo sobre salario por hora, el signo del sesgo es negativo. Esto significa que el estimador MCO esta subestimando el retorno a la educación cuando la variable experiencia es omitida. Tal como se aprecia en los output de ambas regresiones:

```
. regress lyph esc04 [w=factor]
(analytic weights assumed)
(sum of wgt is   2.9721e+06)

      Source |       SS       df       MS              Number of obs =    4727
-------------+------------------------------           F(  1,  4725) = 1513.93
       Model |  531.313164     1  531.313164            Prob > F      =  0.0000
    Residual |  1658.23313  4725   .35094881            R-squared     =  0.2427
-------------+------------------------------           Adj R-squared =  0.2425
       Total |  2189.54629  4726  .463297988            Root MSE      =  .59241

------------------------------------------------------------------------------
        lyph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       esc04 |   .1060137   .0027246    38.91   0.000     .1006721    .1113553
       _cons |   5.575595   .0332557   167.66   0.000     5.510398    5.640792
------------------------------------------------------------------------------

. regress lyph esc04 experiencia [w=factor]
(analytic weights assumed)
(sum of wgt is   2.9721e+06)

      Source |       SS       df       MS              Number of obs =    4727
-------------+------------------------------           F(  2,  4724) =  823.88
       Model |  566.226772     2  283.113386            Prob > F      =  0.0000
    Residual |  1623.31952  4724  .343632413            R-squared     =  0.2586
-------------+------------------------------           Adj R-squared =  0.2583
       Total |  2189.54629  4726  .463297988            Root MSE      =   .5862

------------------------------------------------------------------------------
        lyph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       esc04 |    .112605   .0027743    40.59   0.000     .1071662    .1180438
 experiencia |   .0011489    .000114    10.08   0.000     .0009254    .0013723
       _cons |   5.348351   .0398891   134.08   0.000      5.27015    5.426553
------------------------------------------------------------------------------
```

II.4. Inclusión de variables irrelevantes

Con el objetivo de eliminar el potencial problema de omisión de variables relevantes, siempre existe la tentación de incluir la mayor cantidad de variables explicativas posibles. Esto nos puede llevar a incluir variables irrelevantes.

La inclusión de variables irrelevantes no genera problemas de sesgo en la estimación, ya que el error sigue teniendo media cero y no está correlacionado con las variables explicativas del modelo. Sin embargo, incluir variables irrelevantes

genera un problema de ineficiencia, la varianza del estimador será mayor, provocando que la estimación sean menos precisa.

II.5. Multicolinealidad

El problema de multicolinealidad surge cuando se incluyen variables explicativas similares. Una de las dos variables es irrelevante, ya que no aporta información adicional con respecto a la otra.

Algunas fuentes de la multicolinealidad son:

- El método de recolección de información empleado
- Restricción de la población objeto de muestreo
- Especificación del modelo

La multicolinealidad, al igual que la inclusión de variables relevantes, genera problemas de eficiencia. La estimación MCO en presencia de variables colineales es imprecisa o ineficiente, pero sigue siendo insesgada.

El problema de multicolinealidad es fácil de detectar, pero no tiene más solución que eliminar la variable que no esta aportando información distinta de las otras.

Síntomas de la estimación en presencia de multicolinealidad:

1- El modelo tiene un ajuste bueno (R^2 alto), pero los parámetros resultan ser estadísticamente no significativos.
2- Pequeños cambios en los datos producen importantes cambios en las estimaciones.

3- Los coeficientes pueden tener signos opuestos a los esperados o una magnitud poco creíble.

Cuando existe multicolinealidad perfecta STATA automáticamente borra una de las dos variables.

El comando **`estat vif`** (post estimación) reporta el factor de inflación de varianza (VIF) de cada variable explicativa del modelo, y el promedio del modelo. Este factor mide el grado en que la varianza del coeficiente estimado para la variable ha sido inflada, como producto de que esta variable no es ortogonal (no es independiente) de las restantes variables del modelo.

$$VIF_k = \frac{1}{(1 - R_k^2)}$$

Donde R^2_k representa el R^2 (coeficiente de determinación) de la regresión entre la variable explicativa k y las restantes variables explicativas del modelo. Si R^2_k es grande significa que el comportamiento de la variable independiente k se puede explicar en gran medida con el comportamiento de las restantes variables de modelo, con lo cual esta variable no entrega información diferente a la que están entregando las restantes variables del modelo. La regla sobre este factor, es que existe multicolinealidad si el promedio de todos los VIF es mayor a 1 o el mayor es tiene un valor superior a 10.

Volvamos al modelo (2) que busca explicar el salario por hora a través de la experiencia y escolaridad, pero adicionalmente se incorporan tres variables: el IMC (índice de masa corporal)[3], la estatura, y el peso de la persona. Estas variables busca determinar si las características físicas de la persona tienen influencia sobre el

[3] IMC=peso/estatura^2

salario por hora, dado un nivel de escolaridad y experiencia constante. El modelo estimado es el siguiente:

$$\text{lyph}_i = \alpha + \beta_{esc} \cdot \text{esc}_i + \beta_{exp} \cdot \text{exp}_i + \beta_{imc} \cdot \text{IMC}_i + \beta_{est} \cdot \text{EST}_i + \beta_{peso} \cdot \text{PESO}_i + \mu_i$$

Los resultados de la estimación de este modelo son los siguientes:

```
. reg lyph esc04 experiencia imc estatura peso[w=factor]
(analytic weights assumed)
(sum of wgt is   2.8761e+06)

      Source |       SS       df       MS              Number of obs =    4578
-------------+------------------------------           F(  5,  4572) =  355.59
       Model |  594.863788     5  118.972758            Prob > F      =  0.0000
    Residual |  1529.7082   4572  .334581846           R-squared     =  0.2800
-------------+------------------------------           Adj R-squared =  0.2792
       Total |  2124.57199  4577  .464184398           Root MSE      =  .57843

------------------------------------------------------------------------------
        lyph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       esc04 |   .1115549   .0028177    39.59   0.000     .1060309     .117079
 experiencia |   .0011325   .0001163     9.74   0.000     .0009045    .0013605
         imc |   .0144427   .0088342     1.63   0.102    -.0028767     .031762
    estatura |   .0158268   .0032344     4.89   0.000     .0094859    .0221677
        peso |  -.0055682   .0034532    -1.61   0.107     -.012338    .0012017
       _cons |   2.753484   .5217005     5.28   0.000     1.730699    3.776269
------------------------------------------------------------------------------
```

Luego de la estimación podemos obtener el factor de inflación de varianza de cada una de las variables incluidas en la especificación, el que se obtiene haciendo una regresión de cada variable explicativa contra las restantes. Se reporta un VIF para cada variable, y el promedio de ellos.

. estat vif

Variable	VIF	1/VIF
peso	26.10	0.038314
imc	19.65	0.050893
estatura	12.25	0.081609
experiencia	1.10	0.912957
esc04	1.07	0.937564
Mean VIF	12.03	

$(1-R_k^2)$

¿Qué se puede concluir?

- El modelo tiene multicolinealidad, el promedio de los factores VIF es mayor a 1, y el mayor de estos factores es 26.1 (sobre el 10 sugerido).
- Las variables IMC, estatura y peso son las que presentan colinealidad. La segunda columna de esta tabla muestra el coeficiente de determinación de la regresión entre una variable explicativa, y las restantes, para cada una de ellas. En efecto, podemos observar que poca más de un 96% del comportamiento de la variable peso es explicado por las otras variables incluidas en el modelo. Algo similar sucede con las variables IMC y estatura, donde gran parte del comportamiento de estas variables puede ser explicado por las restantes variables del modelo, un 95% y 92% respectivamente.

- De los anterior se concluye que a pesar de que las variables resultan ser medianamente significativas (al 10%), están no pueden ser incluidas en forma conjunta en la especificación, ya que generan multicolinealidad.

- La escolaridad y experiencia, no tienen problema de colinealidad, un muy bajo porcentaje de su comportamiento se explica por el de las restantes variables explicativas, un 6% y 9% respectivamente.

La siguiente tabla presenta la estimación de cinco modelos distintos que fueron guardados en STATA para presentarlos en forma conjunta en una sola tabla.

```
reg lyph esc04 experiencia [w=factor]
estimates store model2
quietly reg lyph esc04 experiencia imc [w=factor]
estimates store model3
quietly reg lyph esc04 experiencia estatura [w=factor]
estimates store model4
quietly reg lyph esc04 experiencia peso [w=factor]
estimates store model5
quietly reg lyph esc04 experiencia peso estatura [w=factor]
estimates store model6
estimates table model2 model3 model4 model5 model6, stat(r2_a, rmse) b(%7.3g)
p(%4.3f)
```

Variable	model2	model3	model4	model5	model6
esc04	.113	.114	.112	.112	.111
	0.000	0.000	0.000	0.000	0.000
experiencia	.00115	.00122	.00112	.00103	.00113
	0.000	0.000	0.000	0.000	0.000
imc		-.0035			
		0.088			
estatura			.0108		.0108
			0.000		0.000
peso				.00407	-7.7e-05
				0.000	0.923
_cons	5.35	5.42	3.56	5.09	3.56
	0.000	0.000	0.000	0.000	0.000
r2_a	.258	.258	.28	.262	.279
rmse	.586	.587	.577	.585	.579

legend: b/p

Sabemos que por el problema de multicolinealidad no podemos incluir las tres variables en forma simultánea en el modelo. Los modelos 3, 4, y 5 son transformaciones del modelo (2), incorporando cada una de estas variables en forma independiente, y luego el modelo 6 incluye peso y estatura simultáneamente. En el modelo 3 apreciamos que la variable IMC sólo resulta significativa al 8.8% de significancia. En los modelos 4 y 5 concluimos que tanto la estatura como el peso

resultan ser variables significativas, si son incluidas en independientemente. El modelo con mejor ajuste (coeficiente de determinación) es el modelo 4 que incluye estatura, además es el que tiene menor error cuadrático medio:

$$(4) \quad \text{lyph}_i = \alpha + \beta_{\text{esc}} \cdot \text{esc}_i + \beta_{\text{exp}} \cdot \text{exp}_i + \beta_{\text{est}} \cdot \text{EST}_i + \mu_i$$

II.6. Variables categóricas o cualitativas como regresores

En gran parte de los modelos de regresión lineal las variables cualitativas son fundamentales para una correcta especificación. Hasta ahora hemos visto la incorporación de una o más variables explicativas, esencialmente cuantitativas y continuas.

Las variables cualitativas indican la presencia o ausencia de cierta cualidad, pueden tener dos o más categorías. Para la incorporación de variables cualitativas en el modelo de regresión esto siempre se debe hacer en forma de variable Dummy. Las variables Dummies (ficticias, dicotómicas, etc.) toman sólo valores 1 y 0, donde 1 indica la presencia de cierta característica y 0 que la característica no esta presente.

Por ejemplo, en la base de datos contamos con la variable género:

```
. tab  genero

     genero |      Freq.     Percent        Cum.
------------+-----------------------------------
     Hombre |      3,744       50.26       50.26
      Mujer |      3,705       49.74      100.00
------------+-----------------------------------
      Total |      7,449      100.00
```

Esta variable toma valor 1 cuando la persona es hombre y 2 cuando es mujer. La variable así definida no es una variable Dummy, la inclusión de la variable genero en el modelo, definida de esta forma, es incorrecta.

Debemos redefinir la variable para que la cualidad "Hombre" o, indistintamente, la cualidad "Mujer" tome el valor 1 y los restantes cero. De esta forma, se pueden definir dos variables dummies, pero una de ellas es redundante. En términos generales, si una variable tiene n categorías debo definir al menos n-1 dummies, siempre una de ellas es redundante, esta categoría que queda fuera se denomina **categoría base**.

Sigamos con el ejemplo de la variable género, puedo definir una dummy de la siguiente forma:

```
g sexo=1 if genero==1
replace sexo=0 if genero==2
```

Pero podría haber definido de la variable de esta otra forma:

```
g sexo_2=1 if genero==2
replace sexo_2=0 if genero==1
```

Continuemos con el modelo (4) estimado anteriormente. Ahora nos interesa, además, controlar por la característica género, esto significa que nuestro modelo esta controlando por las diferencias que pudiesen existir en salario por hora entre hombres y mujeres. Para esto debemos incluir la variable dummy de género en nuestra estimación. Cual de ellas se elija para formar parte del modelo no tiene mayor relevancia, sólo se debe tener claro para la interpretación del coeficiente asociado a la dummy, recuerde que la categoría que se deja fuera de la estimación se denomina categoría base, y por lo tanto la interpretación de el o los coeficientes siempre se hace tomando como referencia la categoría base.

Estimemos el siguiente modelo de regresión lineal:

$$(7) \quad \text{lyph}_i = \alpha + \beta_{\text{esc}} \cdot \text{esc}_i + \beta_{\text{exp}} \cdot \text{exp}_i + \beta_{\text{est}} \cdot \text{EST}_i + \beta_{\text{sexo}} \cdot \text{sexo}_i + \mu_i$$

```
.regress lyph esc04 experiencia estatura sexo [w=factor]
(analytic weights assumed)
(sum of wgt is   2.9240e+06)

      Source |       SS       df       MS              Number of obs =    4649
-------------+------------------------------           F(  4,  4644) =  455.87
       Model |  606.753782     4  151.688445            Prob > F      =  0.0000
    Residual |  1545.26191  4644  .332743735            R-squared     =  0.2819
-------------+------------------------------           Adj R-squared =  0.2813
       Total |  2152.01569  4648  .462998212            Root MSE      =  .57684

------------------------------------------------------------------------------
        lyph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       esc04 |   .1133031   .0028374    39.93   0.000     .1077406    .1188657
 experiencia |   .0010545   .0001153     9.14   0.000     .0008284    .0012806
    estatura |   .0084755   .0011773     7.20   0.000     .0061675    .0107835
       sexo |   .0731677   .0227193     3.22   0.001     .0286271    .1177083
       _cons |   3.896486   .1877199    20.76   0.000     3.528466    4.264506
```

¿Qué significa el coeficiente asociado a la variable Dummy?

Logaritmo del salario por hora esperado para los hombres:

$$E[\text{lyph}_i \mid \text{hombre}, X_i] = \alpha + \beta_{\text{esc}} \cdot \text{esc}_i + \beta_{\text{exp}} \cdot \text{exp}_i + \beta_{\text{est}} \cdot \text{EST}_i + \beta_{\text{sexo}}$$

Logaritmo del salario por hora esperado para las mujeres:

$$E[\text{lyph}_i \mid \text{mujer}, X_i] = \alpha + \beta_{\text{esc}} \cdot \text{esc}_i + \beta_{\text{exp}} \cdot \text{exp}_i + \beta_{\text{est}} \cdot \text{EST}_i$$

Si tengo dos personas con la misma escolaridad, la misma experiencia, y la misma estatura, pero la única diferencia es que uno es hombre y la otra es mujer, la

diferencia en el logaritmo del salario por hora promedio es exactamente igual a β_{sexo}. En este caso, el coeficiente nos indica que, todo lo demás constante, pasar de ser mujer a hombre incrementa el salario por hora en un 7.3%.

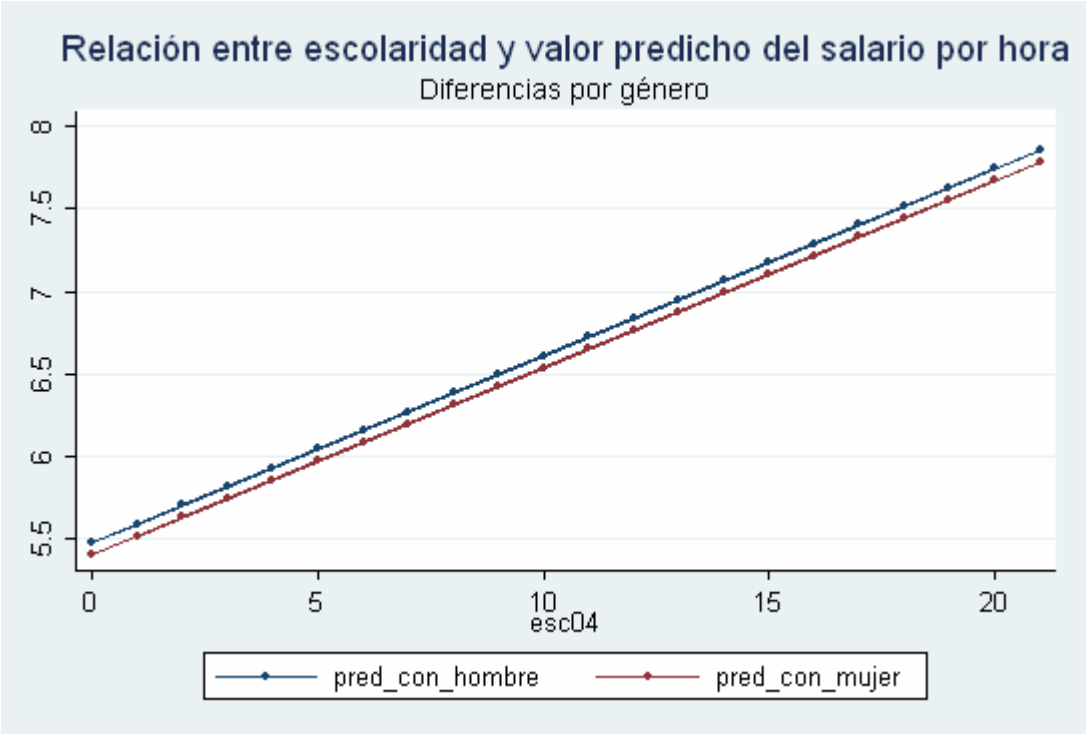
¿Qué sucede si hubiésemos incluido la otra dummy, la que se define como 1 cuando la persona es mujer y cero cuando es hombre?

```
. regress lyph esc04 experiencia estatura sexo_2 [w=factor]
(analytic weights assumed)
(sum of wgt is   2.9240e+06)

      Source |       SS       df       MS              Number of obs =    4649
-------------+------------------------------           F(  4,  4644) =  455.87
       Model |  606.753782     4  151.688445           Prob > F      =  0.0000
    Residual |  1545.26191  4644  .332743735           R-squared     =  0.2819
-------------+------------------------------           Adj R-squared =  0.2813
       Total |  2152.01569  4648  .462998212           Root MSE      =  .57684

------------------------------------------------------------------------------
        lyph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       esc04 |   .1133031   .0028374    39.93   0.000     .1077406    .1188657
 experiencia |   .0010545   .0001153     9.14   0.000     .0008284    .0012806
    estatura |   .0084755   .0011773     7.20   0.000     .0061675    .0107835
      sexo_2 |  -.0731677   .0227193    -3.22   0.001    -.1177083   -.0286271
       _cons |   3.969653   .2013815    19.71   0.000      3.57485    4.364457
------------------------------------------------------------------------------
```

El coeficiente asociado a la dummy género, que toma valor 1 si la persona es mujer y 0 si la persona es hombre, mide la disminución porcentual en el salario por hora de pasar de ser hombre a mujer. El coeficiente es exactamente igual al de la regresión anterior solo que con el signo opuesto, lo que se debe a que la dummy fue definida en forma inversa.

El siguiente gráfico muestra la relación entre escolaridad y el valor predicho del logaritmo del salario por hora, para hombres y mujeres. Podemos apreciar que este tipo de modelos nos permite captura sólo diferencias en intercepto (nivel) entre hombres y mujeres.

```
g
pred_con_mujer=_b[_cons]+_b[esc04]*esc04+_b[experiencia]*104.9436+_b[estatura]*
165.1241+_b[sexo_2]
g
pred_con_hombre=_b[_cons]+_b[esc04]*esc04+_b[experiencia]*104.9436+_b[estatura]
* 165.1241

twoway (connected pred_con_hombre esc04 if sexo==1, msize(small)),
title(Relación entre escolaridad y valor predicho del salario por hora)
subtitle(Diferencias por género) || (connected  pred_con_mujer esc04 if
sexo==0, msize(small))
```



La variable dummy también permite obtener impactos diferenciados, de acuerdo a alguna cualidad, de alguna variable explicativa sobre la variable dependiente. Por ejemplo, mediante la siguiente especificación podemos estimar un retorno a la educación diferenciado entre hombres y mujeres:

(8) $$\text{lyph}_i = \alpha + \beta_{\text{esc}} \cdot \text{esc}_i + \beta_{\text{exp}} \cdot \text{exp}_i + \beta_{\text{est}} \cdot \text{EST}_i + \beta_{\text{sexo}} \cdot \text{sexo}_i + \beta_{\text{esc-sexo}} \cdot \text{sexo}_i \cdot \text{esc}_i + \mu_i$$

Logaritmo del salario por hora esperado para los hombres:

$$E[lyph_i \mid hombre, X_i] = \alpha + \beta_{esc} \cdot esc_i + \beta_{exp} \cdot exp_i + \beta_{est} \cdot EST_i + \beta_{sexo} + \beta_{esc-sexo} \cdot esc_i$$

Logaritmo del salario por hora esperado para las mujeres:

$$E[lyph_i \mid mujer, X_i] = \alpha + \beta_{esc} \cdot esc_i + \beta_{exp} \cdot exp_i + \beta_{est} \cdot EST_i$$

Reescribiendo ambos valores esperados:

$$E[lyph_i \mid hombre, X_i] = (\alpha + \beta_{sexo}) + (\beta_{esc} + \beta_{esc-sexo}) \cdot esc_i + \beta_{exp} \cdot exp_i + \beta_{est} \cdot EST_i$$

$$E[lyph_i \mid mujer, X_i] = \alpha + \beta_{esc} \cdot esc_i + \beta_{exp} \cdot exp_i + \beta_{est} \cdot EST_i$$

Ahora obtengamos el retorno a la educación, que corresponde al cambio porcentual en el salario por hora como resultados de un cambio en los años de escolaridad:

Retorno a la educación de los hombres:

$$\frac{\partial E[lyph_i \mid hombre, X_i]}{\partial esc} = \beta_{esc} + \boxed{\beta_{esc-sexo}}$$

El coeficiente asociado a la variable interactiva representa las diferencias en retorno a la educación

Retorno a la educación de las mujeres:

$$\frac{\partial E[lyph_i \mid mujer, X_i]}{\partial esc} = \beta_{esc}$$

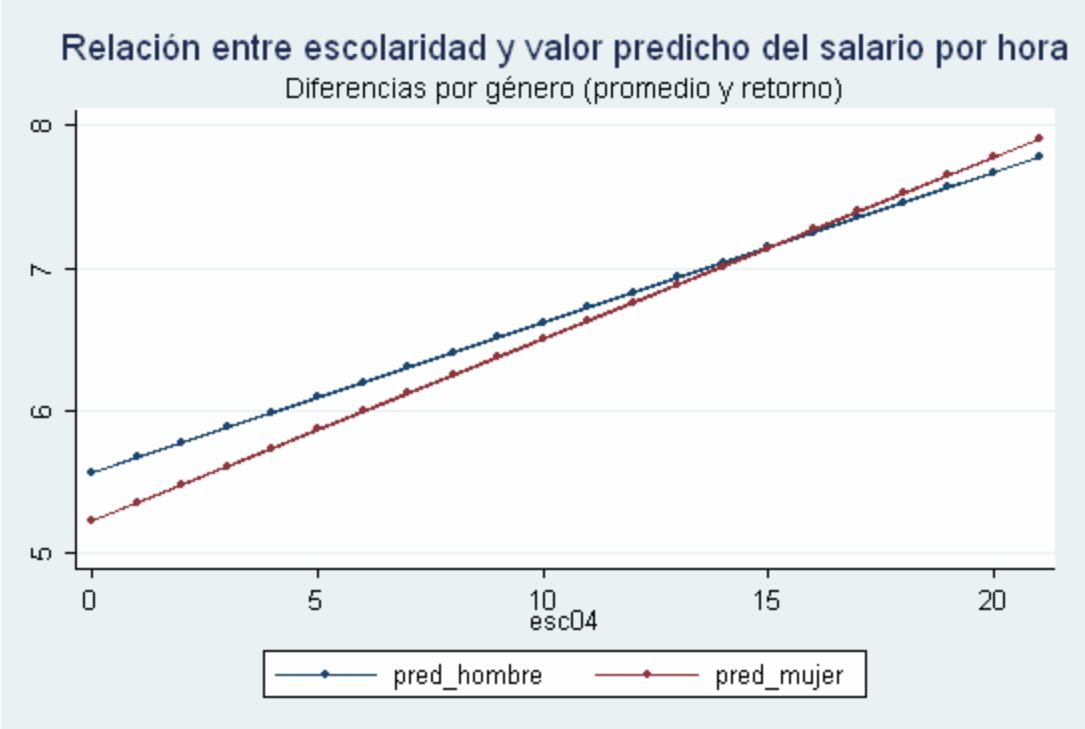
A continuación se presenta la estimación en STATA del modelo 8:

```
g sexo_esc=sexo*esc04
```

```
. regress lyph esc04 experiencia estatura sexo sexo_esc [w=factor]
(analytic weights assumed)
(sum of wgt is   2.9240e+06)

      Source |       SS       df       MS              Number of obs =    4649
-------------+------------------------------           F(  5,  4643) =  368.80
       Model |  611.734414     5  122.346883           Prob > F      =  0.0000
    Residual |  1540.28127  4643  .331742682           R-squared     =  0.2843
-------------+------------------------------           Adj R-squared =  0.2835
       Total |  2152.01569  4648  .462998212           Root MSE      =  .57597

------------------------------------------------------------------------------
        lyph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       esc04 |   .1275529   .0046423    27.48   0.000     .1184517    .1366541
 experiencia |   .0009996    .000116     8.62   0.000     .0007721     .001227
    estatura |   .0086395   .0011763     7.34   0.000     .0063335    .0109456
        sexo |   .3425243    .073124     4.68   0.000     .1991666     .485882
    sexo_esc |   -.022392    .005779    -3.87   0.000    -.0337216   -.0110625
       _cons |   3.700245   .1941592    19.06   0.000     3.319601    4.080889
------------------------------------------------------------------------------
```

El siguiente gráfico muestra la relación entre los años de escolaridad y el valor predicho del salario por hora, considerando diferencias en promedio y retorno entre hombres y mujeres:

```
g
pred_hombre=(_b[_cons]+_b[sexo])+_b[experiencia]*99.70346+_b[estatura]*165.1241
+(_b[esc04]+_b[sexo_esc])*esc04
g
pred_mujer=(_b[_cons])+_b[experiencia]*99.70346+_b[estatura]*165.1241+(_b[esc04
])*esc04

twoway (connected pred_hombre esc04 if sexo==1, msize(small)), title(Relación
entre escolaridad y valor predicho del salario por hora) subtitle(Diferencias
por género (promedio y retorno)) || (connected  pred_mujer esc04 if sexo==0,
msize(small))
```



El gráfico anterior nos muestra que con cero años de escolaridad los hombres obtienen, en promedio, un salario por hora mayor al de las mujeres. Sin embargo, el retorno a la educación de las mujeres (pendiente) es mayor que el de los hombres. Aproximadamente a los 15 años de escolaridad el salario promedio entre hombres y mujeres se iguala.

La siguiente tabla muestra el resumen de las estimaciones de los modelos 4, 7, y 8. Podemos notar que este último, donde controlamos por la característica género en promedio y retorno, tiene una mejor bondad de ajuste y menor error cuadrático medio.

```
. estimates table model4 model7 model8, stat(r2_a, rmse) b(%7.3g) p(%4.3f)

----------------------------------------------
    Variable | model4    model7    model8
-------------+--------------------------------
       esc04 |    .112      .113      .128
             |   0.000     0.000     0.000
 experiencia |  .00112    .00105      .001
             |   0.000     0.000     0.000
    estatura |   .0108    .00848    .00864
             |   0.000     0.000     0.000
        sexo |              .0732      .343
             |              0.001     0.000
    sexo_esc |                       -.0224
             |                        0.000
       _cons |    3.56       3.9       3.7
             |   0.000     0.000     0.000
-------------+--------------------------------
        r2_a |     .28      .281      .283
        rmse |    .577      .577      .576
----------------------------------------------
                                legend: b/p
```

Las variables dummies también nos permiten estimar efectos umbrales, por ejemplo, nos interesa saber no cuanto es el retorno de un año adicional de educación, sino de completar el nivel básico, medio y universitario.

Primero vamos a definir la variable nivel educacional:

```
g nivel=1 if esc04<8
replace nivel=2 if esc04>=8 & esc04<12
replace nivel=3 if esc04>=12 & esc04<17
replace nivel=4 if esc04>=17

label define nivellbl 1 "Ninguna" 2 "Básica Completa" 3 "Media Completa" 4
"Universitaria Completa"
label values nivel nivellbl
```

Esta variable tienen 4 categorías, pero no puede ser incluida tal cual esta definida en el modelo de regresión. Para cada una de estas categorías se puede definir una dummy:

$$DE\_1 = \begin{cases} 1 & \text{si nivel} = \text{ningua} \\ 0 & \text{sino} \end{cases}$$

$$DE\_2 = \begin{cases} 1 & \text{si nivel} = \text{básica completa} \\ 0 & \text{sino} \end{cases}$$

$$DE\_3 = \begin{cases} 1 & \text{si nivel} = \text{media completa} \\ 0 & \text{sino} \end{cases}$$

$$DE\_4 = \begin{cases} 1 & \text{si nivel} = \text{universitria completa} \\ 0 & \text{sino} \end{cases}$$

Pero solo se deben incluir tres de ellas en el modelo, la que no se incluye se denomina categoría base, y las interpretaciones de los coeficientes de las restantes dummies incluidas se deben hacer en función de esta categoría. Si por error todas las dummies son incluidas el programa borrará automáticamente una de ellas.

El comando **tabulate** de STATA tiene una opción para generar rápidamente las dummies de una variable categórica:

```
tab nivel, generate(DE_)
```

Ahora nos interesa estimar los efectos umbrales, es decir, el retorno de completar cada uno de los niveles educacionales. El modelo a estimar es el siguiente:

(9) $$\text{lyph}_i = \alpha + \beta_{\text{exp}} \cdot \text{exp}_i + \beta_{\text{est}} \cdot \text{EST}_i + \beta_{\text{sexo}} \cdot \text{sexo} + \beta_{\text{DE1}} \cdot \text{DE\_2} + \beta_{\text{DE1}} \cdot \text{DE\_3} + \beta_{\text{DE1}} \cdot \text{DE\_4} + \mu_i$$

En esta especificación hemos dejado fuera la dummy correspondiente al nivel educacional "ninguno", esto significa que los restantes coeficientes de nivel educacional deben ser interpretados tomando como base la característica "sin educación". A continuación la estimación del modelo en STATA:

```
. regress lyph experiencia estatura sexo DE_2 DE_3 DE_4 [w=factor]
(analytic weights assumed)
(sum of wgt is   2.9310e+06)

      Source |       SS       df       MS              Number of obs =    4662
-------------+------------------------------           F(  6,  4655) =  242.92
       Model |  513.910914     6  85.6518191           Prob > F      =  0.0000
    Residual |  1641.33738  4655  .352596644           R-squared     =  0.2384
-------------+------------------------------           Adj R-squared =  0.2375
       Total |  2155.24829  4661  .462400406           Root MSE      =   .5938

------------------------------------------------------------------------------
        lyph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 experiencia |   .0008591   .0001182     7.27   0.000     .0006274    .0010909
    estatura |   .0099157   .0012082     8.21   0.000     .0075471    .0122843
        sexo |   .0495806    .023364     2.12   0.034     .0037761    .0953851
        DE_2 |   .1535891   .0378905     4.05   0.000     .0793058    .2278725
        DE_3 |   .5520762   .0312586    17.66   0.000     .4907945    .6133578
        DE_4 |   1.457801   .0455263    32.02   0.000     1.368547    1.547054
       _cons |   4.522802   .1940015    23.31   0.000     4.142467    4.903137
------------------------------------------------------------------------------

. estimates store model9
```

Esta misma regresión se puede hacer con el siguiente comando:

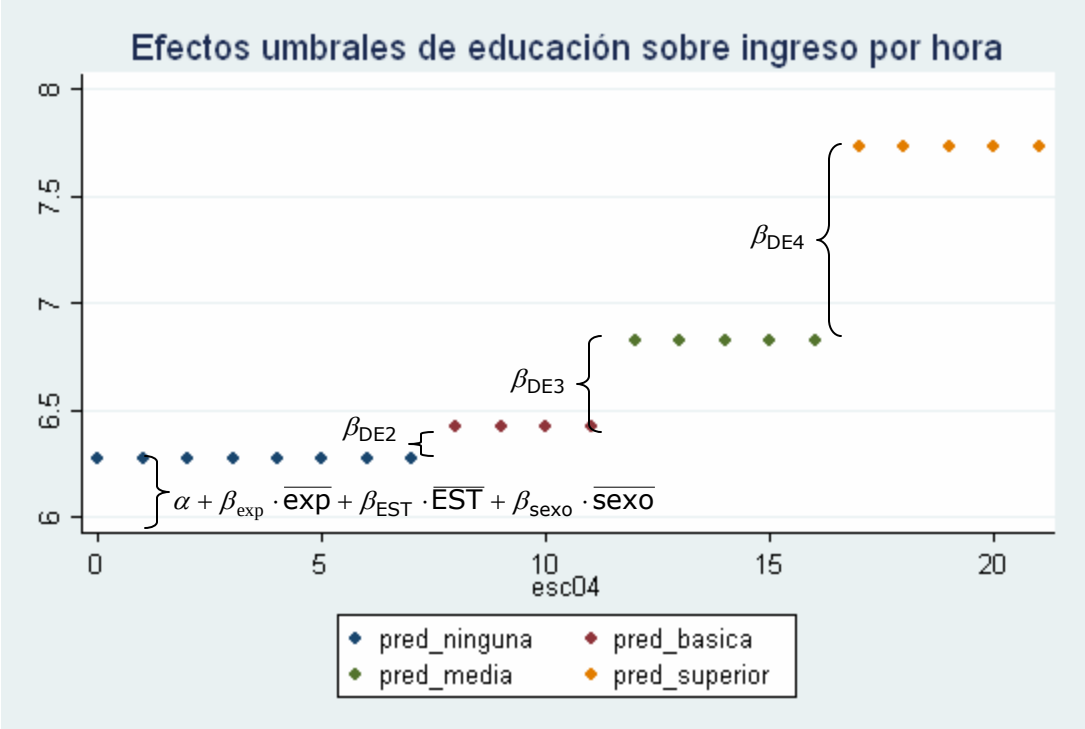
```
. xi: regress lyph experiencia estatura sexo i.nivel [w=factor]
i.nivel           _Inivel_1-4         (naturally coded; _Inivel_1 omitted)
(analytic weights assumed)
(sum of wgt is   2.9310e+06)

      Source |       SS       df       MS              Number of obs =    4662
-------------+------------------------------           F(  6,  4655) =  242.92
       Model |  513.910914     6  85.6518191           Prob > F      =  0.0000
    Residual |  1641.33738  4655  .352596644           R-squared     =  0.2384
-------------+------------------------------           Adj R-squared =  0.2375
       Total |  2155.24829  4661  .462400406           Root MSE      =   .5938

------------------------------------------------------------------------------
        lyph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 experiencia |   .0008591   .0001182     7.27   0.000     .0006274    .0010909
    estatura |   .0099157   .0012082     8.21   0.000     .0075471    .0122843
        sexo |   .0495806    .023364     2.12   0.034     .0037761    .0953851
   _Inivel_2 |   .1535891   .0378905     4.05   0.000     .0793058    .2278725
   _Inivel_3 |   .5520762   .0312586    17.66   0.000     .4907945    .6133578
   _Inivel_4 |   1.457801   .0455263    32.02   0.000     1.368547    1.547054
       _cons |   4.522802   .1940015    23.31   0.000     4.142467    4.903137
------------------------------------------------------------------------------
```

¿Cómo se interpretan los coeficientes asociados a DE_2, DE_3 y DE_4?

El coeficiente asociado a la dummy "básica completa (DE_2)" indica que, todo lo demás constante, pasar de no tener educación a tener educación básica completa aumenta, en promedio, el salario por hora en un 15.4%. El coeficiente asociado a la dummy "media completa (DE_3)" indica que, todo lo demás constante, pasar de no tener educación a tener educación media completa aumenta, en promedio, el salario por hora en un 55.2%. Y por último, el coeficiente asociado a la dummy "universitaria completa (DE_4)" indica que, todo lo demás constante, pasar de no tener educación a tener educación universitaria completa aumenta, en promedio, el salario por hora en un 145.8%.

El siguiente gráfico muestra la relación entre escolaridad y el valor predicho del logaritmo del salario por hora.

```
g
pred_ninguna=_b[_cons]+_b[experiencia]*99.70346+_b[estatura]*165.1241+_b[sexo]*
0.4895707

g
pred_basica=_b[_cons]+_b[experiencia]*99.70346+_b[estatura]*165.1241+_b[sexo]*0
.4895707+_b[DE_2]

g
pred_media=_b[_cons]+_b[experiencia]*99.70346+_b[estatura]*165.1241+_b[sexo]*0.
4895707+_b[DE_3]

g
pred_superior=_b[_cons]+_b[experiencia]*99.70346+_b[estatura]*165.1241+_b[sexo]
*0.4895707+_b[DE_4]

twoway (scatter pred_ninguna esc04 if nivel==1) || (scatter  pred_basica esc04
if nivel==2) || (scatter  pred_media esc04 if nivel==3) || (scatter
pred_superior esc04 if nivel==4), title(Efectos umbrales de educación sobre
ingreso por hora)
```



El modelo anterior tiene como hipótesis que sólo entrega retorno, en términos de salario por hora, completar los diferentes niveles educacionales, pero que al interior de cada nivel avanzar en años de escolaridad no significa un retorno adicional. Para poder estimar retornos a la educación diferenciados entre los niveles educacionales se deben interactuar las dummies de nivel educacional con la variable años de escolaridad. El modelo a estimar es el siguiente:

$$(10)\quad \begin{aligned} lyph_i = {} & \alpha + \beta_{exp} \cdot exp_i + \beta_{est} \cdot EST_i + \beta_{esc} \cdot esc_i + \beta_{sexo} \cdot sexo + \\ & \beta_{DE1} \cdot DE\_2 + \beta_{DE1} \cdot DE\_3 + \beta_{DE1} \cdot DE\_4 + \beta_{esc\text{-}DE2} \cdot esc_i \cdot DE\_2 + \\ & \beta_{esc\text{-}DE3} \cdot esc_i \cdot DE\_3 + \beta_{esc\text{-}DE4} \cdot esc_i \cdot DE\_4 + \mu_i \end{aligned}$$

Las variables interactivas se generan de la siguiente manera:

```
g DE2_esc=DE_2*esc04
g DE3_esc=DE_3*esc04
g DE4_esc=DE_4*esc04
```

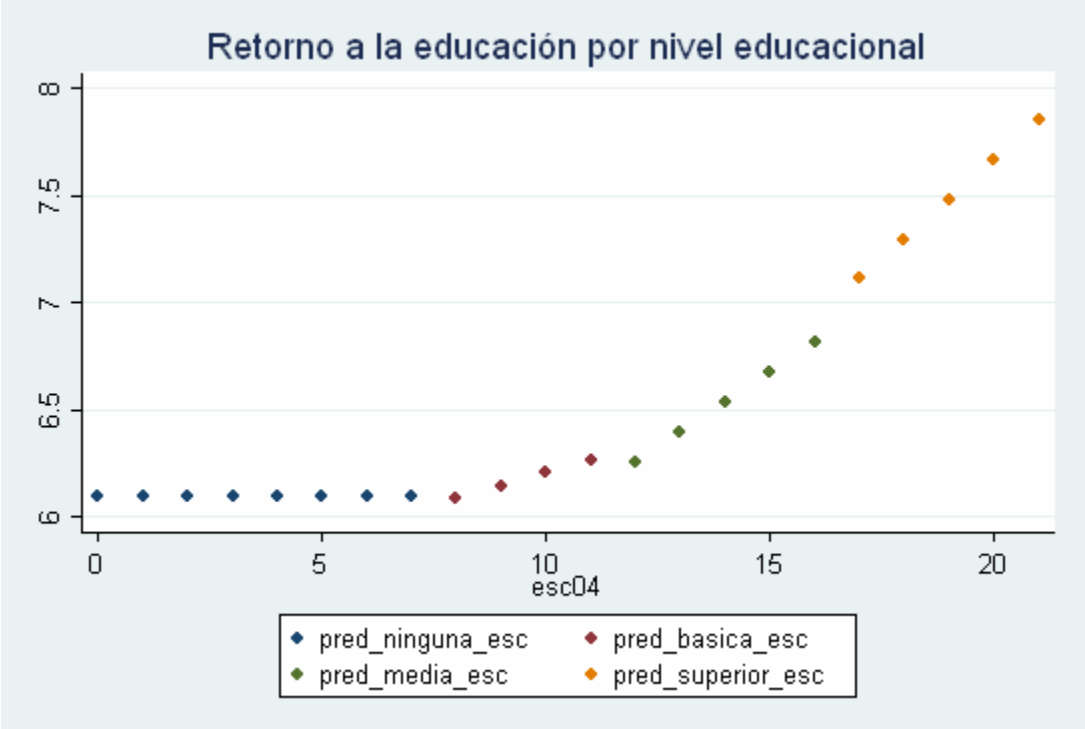
Estos son los resultados en STATA de la estimación del modelo (10):

```
. regress lyph experiencia estatura esc04 sexo DE_2 DE_3 DE_4 DE2_esc DE3_esc
DE4_esc [w=factor]
(analytic weights assumed)
(sum of wgt is   2.9240e+06)

      Source |       SS       df       MS              Number of obs =    4649
-------------+------------------------------           F( 10,  4638) =  236.34
       Model |  726.428785    10  72.6428785           Prob > F      =  0.0000
    Residual |   1425.5869  4638  .307371044           R-squared     =  0.3376
-------------+------------------------------           Adj R-squared =  0.3361
       Total |  2152.01569  4648  .462998212           Root MSE      =  .55441

------------------------------------------------------------------------------
        lyph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 experiencia |   .0010616   .0001111     9.55   0.000     .0008437    .0012795
    estatura |   .0077458   .0011336     6.83   0.000     .0055234    .0099682
       esc04 |    .038609   .0159501     2.42   0.016     .0073391    .0698789
        sexo |   .0822528   .0218629     3.76   0.000     .0393912    .1251145
        DE_2 |  -.4919427   .2173814    -2.26   0.024    -.9181136   -.0657717
        DE_3 |   -1.53227   .1261298   -12.15   0.000    -1.779544   -1.284995
        DE_4 |  -2.115875   .7445445    -2.84   0.005    -3.575537   -.6562139
     DE2_esc |   .0601242   .0282234     2.13   0.033     .0047929    .1154555
     DE3_esc |    .140969   .0175399     8.04   0.000     .1065823    .1753556
     DE4_esc |   .1841626   .0457869     4.02   0.000     .0943986    .2739267
       _cons |   4.630689   .1989754    23.27   0.000     4.240603    5.020775
------------------------------------------------------------------------------

estimates store model10
```

Y el siguiente gráfico muestra la relación entre los años de escolaridad y el valor estimado del salario por hora:

```
g
pred_ninguna_esc=_b[_cons]+_b[experiencia]*99.70346+_b[estatura]*165.1241+_b[se
xo]*0.4895707+_b[esc04]

g
pred_basica_esc=_b[_cons]+_b[experiencia]*99.70346+_b[estatura]*165.1241+_b[sex
o]*0.4895707+_b[DE_2]+_b[esc04]+_b[DE2_esc]*esc04
```

```
g
pred_media_esc=_b[_cons]+_b[experiencia]*99.70346+_b[estatura]*165.1241+_b[sexo
]*0.4895707+_b[DE_3]+_b[esc04]+_b[DE3_esc]*esc04

g
pred_superior_esc=_b[_cons]+_b[experiencia]*99.70346+_b[estatura]*165.1241+_b[s
exo]*0.4895707+_b[DE_4]+_b[esc04]+_b[DE4_esc]*esc04

twoway (scatter pred_ninguna_esc esc04 if nivel==1) || (scatter
pred_basica_esc esc04 if nivel==2) || (scatter  pred_media_esc esc04 if
nivel==3) || (scatter  pred_superior_esc esc04 if nivel==4), title(Retorno a la
educación por nivel educacional)
```



Por último, en la base de datos se cuenta con dos variables dummies más *dsoltero* que toma el valor 1 si la persona es soltera y 0 sino; y *jefe* que toma valor 1 si la persona es jefe de hogar y 0 sino. El resultado de la estimación del modelo incluyendo estas dos variables se presentan a continuación (modelo 11):

```
. estimates table model7 model10 model11, stat(r2_a, rmse) b(%7.3g) p(%4.3f)

-------------------------------------------
    Variable | model7    model10   model11
-------------+-----------------------------
        esc04 |    .113     .0386     .0387
              |   0.000     0.016     0.015
 experiencia |  .00105    .00106    .00075
              |   0.000     0.000     0.000
    estatura |  .00848    .00775    .00764
              |   0.000     0.000     0.000
         sexo |   .0732     .0823     .0623
              |   0.001     0.000     0.006
         DE_2 |             -.492     -.464
              |             0.024     0.032
         DE_3 |             -1.53     -1.55
              |             0.000     0.000
         DE_4 |             -2.12     -1.98
              |             0.005     0.007
     DE2_esc |             .0601     .0569
              |             0.033     0.043
     DE3_esc |              .141      .143
              |             0.000     0.000
     DE4_esc |              .184      .177
              |             0.000     0.000
  dsoltero04 |                      -.0735
              |                      0.000
        jefe |                      .0779
              |                      0.000
       _cons |     3.9      4.63      4.69
              |   0.000     0.000     0.000
-------------+-----------------------------
        r2_a |    .281      .336      .343
        rmse |    .577      .554      .552
-------------------------------------------
                               legend: b/p
```

El modelo con mejor ajuste y menor error cuadrático medio es aquel donde se han incluido las dos dummies adicionales sobre estado civil y jefatura de hogar.

II.7. Incorporación de no linealidades

El estimador MCO asume que la relación entre la variable dependiente y la(s) variable(s) explicativa(s) es lineal. Sin embargo, en algunos casos esta relación es no lineal, y la estimación lineal se comportará relativamente bien para ciertos valores de las variables pero para otros no. La omisión de no linealidades genera un problema de especificación equivalente a la omisión variables relevantes, que se puede solucionar incorporando potencias de las variables explicativas al modelo de regresión lineal.

El comando post-estimación **`estat ovtest`** computa el test RESET de omisión de no linealidades. La idea de este test es bastante simple, el test RESET hace una nueva regresión aumentada donde incluye los regresores originales y además potencias de los valores predichos a través de la especificación original:

$$Y_i = \alpha + \beta X_i + \varphi_1 \cdot \hat{Y}_i^2 + \varphi_2 \cdot \hat{Y}_i^3 + \varphi_3 \cdot \hat{Y}_i^4 + \varepsilon_i$$

La hipótesis nula es que no existen problemas de especificación, es decir, que no existen no linealidades omitidas. Para testear esta hipótesis se hace un test de hipótesis conjunta de que todos los coeficientes de las potencias del valor predicho de la variable dependiente son cero. Si no se puede rechazar la hipótesis nula, los coeficientes asociados a las potencias incluidas en la especificación aumentada son iguales a cero.

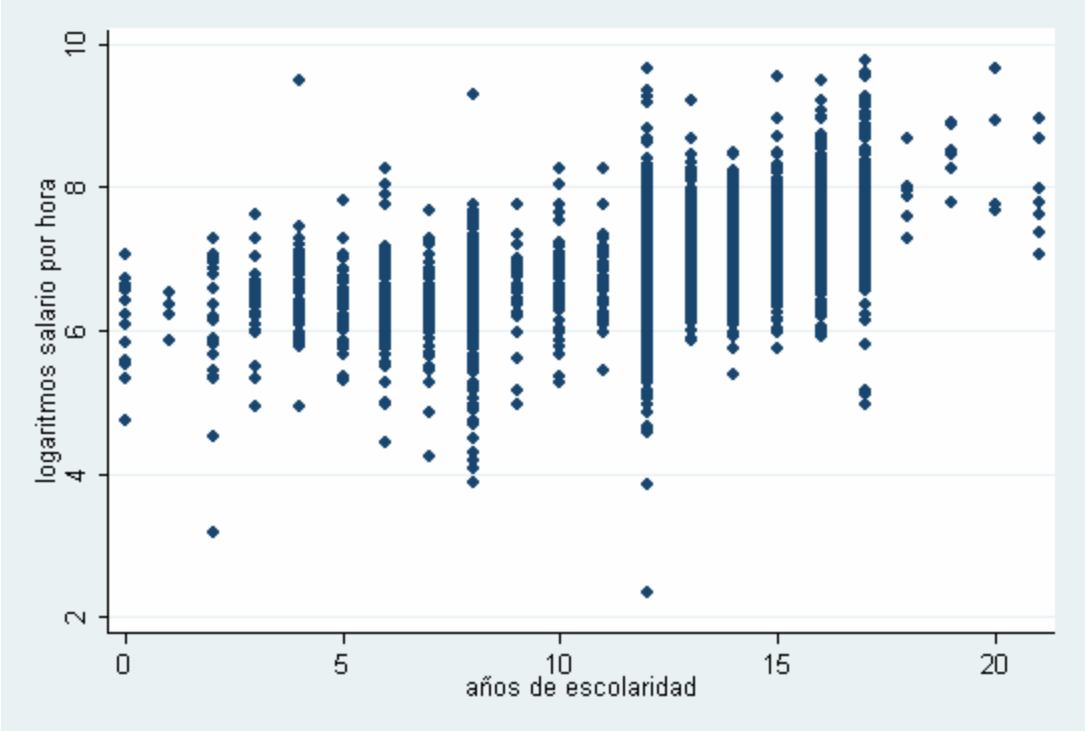
Al hacer el test de no linealidades omitidas después de la estimación del modelo 11 obtenemos el siguiente resultado:

```
. estat ovtest

Ramsey RESET test using powers of the fitted values of lyph
       Ho:  model has no omitted variables
                 F(3, 4633) =     10.19
                   Prob > F =      0.0000
```

Se rechaza la hipótesis nula de que no existen variables no lineales omitidas, es decir, en esta especificación existe un problema de no linealidad. Para solucionar el problema se deben probar diferentes especificaciones que considerando potencias de las variables explicativas, logaritmos, y productos entre ellas. Cuando el test no se pueda rechazar, significa que el modelo ya esta libre de no linealidades omitidas.

II.8. Heteroscedasticidad

En datos de corte transversal el problema de heteroscedasticidad es bastante común. La heteroscedasticidad se produce cuando la varianza del error difiere para distintos valores de la(s) variable(s) explicativa(s). Por ejemplo, para niveles bajos de escolaridad la varianza en el logaritmo del salario por hora es más baja que para niveles de escolaridad más elevados.



La presencia de heteroscedasticidad **no** genera problemas de sesgo en el estimador MCO, es decir, se sigue cumpliendo la propiedad de insesgamiento de este estimador:

$$E\left[\hat{\beta}\right] = \beta$$

Pero, al tener los errores una varianza no constante, la matriz de varianzas y covarianzas del estimar MCO deja de ser la mínima o la más eficiente. La varianza de los coeficientes estimados es mayor, por lo cual toda inferencia basada en la varianza MCO es incorrecta, los estadísticos t se están computando con una varianza menor a la que se debería, por lo tanto son mayores y existe mayor probabilidad de rechazar la hipótesis nula cuando esta no debería ser rechazada. Es decir, la significancia de los parámetros se puede ver afectada, mostrando significancia cuando en realidad no la hay.

Para solucionar el problema de heterocedasticidad se debe conocer el patrón de heterocedasticidad o que variables generan el problema, ya que el estimar MCG o MCF tienen como espíritu "quitar" la heterocedasticidad de las variables explicativas y dependiente, mediante una transformación que consiste en dividir cada observación de la variable dependiente y las variables explicativas por la desviación estándar del error asociado a esa observación. Y luego aplicando MCO a este modelo transformado se obtiene una estimación insesgada y eficiente que cumple con la propiedad MELI (Mejor Estimador Lineal e Insesgado).

La transformación de las variables es la siguiente:

$$\frac{y_1}{\sigma_1}, \frac{x_{11}}{\sigma_1}, \ldots, \frac{x_{k1}}{\sigma_1}$$
$$\frac{y_2}{\sigma_2}, \frac{x_{12}}{\sigma_2}, \ldots, \frac{x_{k2}}{\sigma_2}$$
$$\vdots$$
$$\frac{y_n}{\sigma_n}, \frac{x_{1n}}{\sigma_n}, \ldots, \frac{x_{kn}}{\sigma_n}$$

Después de realizar la estimación del modelo en STATA el comando **`estat hettest`** permite testear la presencia de heterocedasticidad. La hipótesis nula de este test es la homocedasticidad del error. Este comando computa el test de heterocedasticidad de Breusch-Pagan (BP) el que consiste en un test de Wald a la hipótesis nula de que las variables explicativas del modelo original no son significativas en explicar el comportamiento del término de error estimado al cuadrado, para esto se estima una regresión auxiliar del error estimado al cuadrado en función de las variables explicativas originales del modelo.

Por ejemplo, después de la estimar el modelo 11 podemos testear la presencia de heteroscedasticidad de la siguiente forma:

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lyph

         chi2(1)      =    64.89
         Prob > chi2  =   0.0000
```

De este test podemos concluir que se rechaza la hipótesis nula de homocedasticidad. También podemos realizar el test de White que es más amplio que el anterior al considerar no sólo las variables del modelo original como variables explicativas de la regresión auxiliar, sino también los productos y cuadrados de cada una de ellas.

```
whitetst

White's general test statistic :  117.0304  Chi-sq(63)  P-value =  4.2e-05
```

Del cual podemos concluir lo mismo, se rechaza la hipótesis nula de homocedasticidad.

Para poder solucionar el problema y obtener una estimación insesgada y eficiente a través de la metodología de Mínimos Cuadrados Generalizados (MCG) o Mínimos Cuadrados Factibles (MCF) es necesario conocer el patrón de Heterocedasticidad, es decir, conocer la verdadera matriz de varianzas y covarianzas del término de error, conocer las desviaciones estándar de cada error para poder realizar la transformación. Esto en la práctica es poco probable.

La solución más sensata es la planteada por White, que consiste en quedarse con la estimación menos eficiente pero insesgada de MCO, pero estimar el forma correcta la matriz de varianzas y covarianzas de los coeficientes estimados, de forma tal de que los test de hipótesis y la inferencia este realizada en forma apropiada. Esto se hace en STATA simplemente introduciendo la opción **robust** al comando regress.

A continuación vemos las diferencias entre la estimación del modelo 11 sin corregir por heteroscedasticidad y utilizando la opción robust que estima la matriz correcta de varianzas y covarianzas del estimador MCO en presencia de heteroscedasticidad:

Variable	model11	model12
experiencia	.00075	.00075
	1.2e-04	1.5e-04
	0.000	0.000
estatura	.00764	.00764
	.0011	.0014
	0.000	0.000
esc04	.0387	.0387
	.0159	.0212
	0.015	0.068
sexo	.0623	.0623
	.0225	.0259
	0.006	0.016
DE_2	-.464	-.464
	.216	.241
	0.032	0.054
DE_3	-1.55	-1.55
	.126	.173
	0.000	0.000
DE_4	-1.98	-1.98
	.741	1.26
	0.007	0.115
DE2_esc	.0569	.0569
	.0281	.0322
	0.043	0.077
DE3_esc	.143	.143
	.0175	.0235
	0.000	0.000
DE4_esc	.177	.177
	.0456	.076
	0.000	0.020
dsoltero04	-.0735	-.0735
	.0186	.0222
	0.000	0.001

```
        jefe |   .0779     .0779
             |   .0195     .0239
             |   0.000     0.001
       _cons |    4.69      4.69
             |    .199       .25
             |   0.000     0.000
-------------+----------------------
        r2_a |    .343      .343
        rmse |    .552      .552
---------------------------------
                   legend: b/se/p
```

De la comparación de ambos modelos debemos notar lo siguiente:

- Los coeficientes estimados son exactamente iguales.
- La bondad de ajuste del modelo tampoco se ve afectada
- Las varianzas estimadas de los coeficientes son mayores en el modelo que incorpora la presencia de heteroscedasticidad, confirmando lo que los test estadísticos BP y White indicaban sobre la presencia de este problema.
- Algunas variables que resultaban ser significativas al 5%, utilizando la estimación correcta de la varianza y por ende computando en forma apropiada los test estadísticos ya no lo son.

En resumen, siempre utilice la opción robust del comando regress de stata. Si existe el problema de Heterocedasticidad, con esta opción Ud. estará seguro de que los test estadísticos son correctos y así las conclusiones sobre la significancia de los parámetros. Si es que no hay Heterocedasticidad, obtendrá exactamente el mismo resultado que sin ocupar esta opción, ya que sin Heterocedasticidad la matriz de varianzas y covarianzas robusta (o de White), en este caso, sería la misma que la del estimador MCO.

II.9. Selección de modelos anidados

Al final de día puede que más de un modelo satisfaga todos los requerimientos teóricos y econométricos, pero Ud. deberá escoger sólo uno de estos modelos para poder concluir, hacer predicciones y tomar decisiones de política.

Los modelos sobre los cuales tiene que elegir pueden estar anidados o no. Se dice que dos modelos están anidados cuando uno de ellos corresponde al anterior imponiendo cierta restricciones sobre los parámetros.

II.9.1. Selección entre modelos anidados

Los criterios de información de Akaike (AIC) y Schwarz (BIC) son medidas consistentes para ver el mejor modelo. El mejor modelo es aquel que tiene menor valor del criterio de información.

Volvamos al modelo (11), y supongamos que no estamos seguros si incluir o no la variable estatura. Podemos ver los criterios de información. Debemos escoger el modelo que minimiza los criterios de información.

La siguiente tabla resume la estimación del modelo (11) con y sin la variable estatura, ambos con la opción robust:

```
regress lyph experiencia estatura esc04 sexo DE_2 DE_3 DE_4 DE2_esc DE3_esc
DE4_esc dsoltero jefe [w=factor], robust
estimates store model12

regress lyph experiencia esc04 sexo DE_2 DE_3 DE_4 DE2_esc DE3_esc DE4_esc
dsoltero jefe [w=factor], robust
estimates store model13

estimates table model13 model12, stat(r2_a, rmse, aic, bic) b(%7.3g) p(%4.3f)
```

```
----------------------------------
    Variable | model13   model12
-------------+--------------------
 experiencia |  .00065    .00075
             |   0.000     0.000
       esc04 |   .0365     .0387
             |   0.081     0.068
        sexo |    .152     .0623
             |   0.000     0.016
        DE_2 |    -.47     -.464
             |   0.046     0.054
        DE_3 |   -1.59     -1.55
             |   0.000     0.000
        DE_4 |   -1.46     -1.98
             |   0.215     0.115
     DE2_esc |   .0599     .0569
             |   0.058     0.077
     DE3_esc |     .15      .143
             |   0.000     0.000
     DE4_esc |     .15      .177
             |   0.035     0.020
  dsoltero04 |  -.0766    -.0735
             |   0.001     0.001
        jefe |     .08     .0779
             |   0.001     0.001
    estatura |             .00764
             |             0.000
       _cons |    5.91      4.69
             |   0.000     0.000
-------------+--------------------
        r2_a |    .336      .343
        rmse |    .555      .552
         aic |    7855      7677
         bic |    7932      7761
----------------------------------
                      legend: b/p
```

Ambos criterios nos indican que nos debemos quedar con el modelo que incluye la variable estatura.

II.9.2. Selección de modelos no anidados

Cuando tenemos dos modelos que no están anidados, por ejemplo, los modelos (8) y (10), la simple comparación de la bondad de ajuste del modelo a través del R^2 o R^2 ajustado, o del error cuadrático medio del modelo no son una forma apropiada, y no nos permitirá concluir sobre el mejor modelo.

(8) $\text{lyph}_i = \alpha + \beta_{\text{esc}} \cdot \text{esc}_i + \beta_{\text{exp}} \cdot \text{exp}_i + \beta_{\text{est}} \cdot \text{EST}_i + \beta_{\text{sexo}} \cdot \text{sexo}_i + \beta_{\text{esc-sexo}} \cdot \text{sexo}_i \cdot \text{esc}_i + \mu_i$

(10) $\text{lyph}_i = \alpha + \beta_{\text{exp}} \cdot \text{exp}_i + \beta_{\text{est}} \cdot \text{EST}_i + \beta_{\text{esc}} \cdot \text{esc}_i + \beta_{\text{sexo}} \cdot \text{sexo} +$
$\beta_{\text{DE1}} \cdot \text{DE\_2} + \beta_{\text{DE1}} \cdot \text{DE\_3} + \beta_{\text{DE1}} \cdot \text{DE\_4} + \beta_{\text{esc-DE2}} \cdot \text{esc}_i \cdot \text{DE\_2} +$
$\beta_{\text{esc-DE3}} \cdot \text{esc}_i \cdot \text{DE\_3} + \beta_{\text{esc-DE4}} \cdot \text{esc}_i \cdot \text{DE\_4} + \mu_i$

¿Cómo escogemos entre el modelo (8), que estima un retorno a la educación diferenciado por sexo pero igual para todos los niveles educacionales, y el modelo (10) que estima un retorno diferenciado por nivel educacional?

Davidson y MacKinnon (1981) propusieron el test J para poder seleccionar entre modelos no anidados. Este test consiste definir uno de los modelos como aquel bajo la hipótesis nula y el otro como bajo la alternativa, se estiman ambos modelos y se obtiene el valor predicho de la variable dependiente, luego el valor predicho con el modelo de la hipótesis alternativa se incluye como variable explicativa del modelo bajo la hipótesis nula, y se testea la significancia estadística de esta nueva variable, si es estadísticamente significativa se rechaza el modelo de la hipótesis nula. Luego se invierten los modelos definidos bajo la hipótesis nula y se repite el procedimiento. Se pueden dar cuatro soluciones:

1- Se rechaza el modelo (8) y no el modelo (10)
2- Se rechaza el modelo (10) y no el modelo (8)
3- Se rechazan ambos modelos
4- No se puede rechazar ninguno de los dos modelos

Sólo en los primeros dos casos el test J nos permite concluir sobre el modelo que debemos preferir.

El comando para realizar este test no viene en STATA pero puede ser instalado ejecutando el siguiente comando: **ssc install nnest**.

La ejecución de este comando sobre los dos modelos anteriores se debe realizar de la siguiente forma:

```
. nnest lyph esc04 experiencia estatura sexo sexo_esc (experiencia estatura
esc04 sexo DE_2 DE_3 DE_4 DE2_esc DE3_esc D
> E4_esc)
```

El comando nos entrega dos resultados, el del tes J de Davidson y MacKinnon y el del test de Cox-Pearsan, que esa bastante similar. El resultado se presenta a continuación:

```
M1 : Y = a + Xb with X = [esc04 experiencia estatura sexo sexo_esc]
M2 : Y = a + Zg with Z = [experiencia estatura esc04 sexo DE_2 DE_3 DE_4
DE2_esc DE3_esc DE4_esc]

J test for non-nested models

H0 : M1  t(4642)    19.19419
H1 : M2  p-val       0.00000

H0 : M2  t(4637)    -0.93573
H1 : M1  p-val       0.34946

Cox-Pesaran test for non-nested models

H0 : M1  N(0,1)    -5.48e+02
H1 : M2  p-val       0.00000

H0 : M2  N(0,1)    -75.18041
H1 : M1  p-val       0.00000
```

El test J nos permite concluir que el modelo (8) se rechaza, pero no así el modelo (10). Con lo cual nos debemos quedar con el último de los modelos estimados.

Capítulo III. Estimador de Variables Instrumentales

III.1. Introducción

Uno de los supuestos claves para que el estimador MCO sea insesgado es que el término de error no debe estar correlacionado con las variables explicativas o regresores del modelo:

$$Cov(\mu_i, X_i) = 0$$

Existen tres situaciones que pueden invalidar este supuesto:

- **Omisión de variables relevantes**
- **Simultaneidad:** determinación simultánea de la variable de interés (variable dependiente) y los regresores.
- **Error de medición** en los regresores

A pesar de que estos problemas son generados por diferentes razones, la solución a ellos es una sola y se llama **Estimador de Variables Instrumentales (IV).**

Este estimador nos permite obtener una estimación consistente de los coeficientes de interés cuando no se cumple el supuesto de correlación cero entre el término de error y una o más variables explicativas. Para entender como funciona el estimador IV, pensemos que una de las variables explicativas esta compuesta por una parte que esta correlacionada con el error (por cualquiera de las tres razones antes mencionadas), y otra parte que no esta correlacionada con el error. Si se tiene información suficiente para aislar la segunda parte de la variable, luego nos podemos enfocar en como la variación en esta parte de la variable explicativa afecta la variación de la variable dependiente. De esta forma, se elimina el sesgo en la estimación MCO considerando sólo la parte de la variable explicativa que no esta

correlacionada con el error. Esto es exactamente lo que hace el estimador de variables instrumentales. La información sobre los movimientos de la variable explicativa que no están correlacionados con el término de error se captura a través de una o más variables llamadas **variables instrumentales** o simplemente **instrumentos**.

En resumen, la regresión por variables instrumentales usa estas variables como herramientas o "instrumentos" para aislar del comportamiento de la variable explicativa la parte no correlacionada con el término de error, lo que permite una estimación consistente de los coeficientes de regresión.

III.2. Endogeneidad

El estimador MCO asume que la causalidad es en un sentido, de la variable explicativa a la variable dependiente. Pero la causalidad podría, en algunos casos, también funcionar en ambos sentidos. Por ejemplo, generalmente en el modelo de la ecuación de Mincer se asume que la escolaridad afecta el nivel de ingresos, pero la relación entre estas variables también podría ser inversa, el nivel de ingresos determina el nivel de educación. Otro ejemplo es el relacionado con el tamaño de los cursos (o número de alumnos por profesor) y los resultados académicos (prueba SIMCE), en general, se asume que la causalidad es en el sentido de que cursos más pequeños tienen mejores logros educacionales, pero se podría esperar también una relación inversa, mientras menores son los logros el gobierno entrega mayores recursos y menor es el número de alumnos por profesor. En ambos casos se dice que la variable explicativa, años de escolaridad o número de alumnos por profesor, es endógena.

$$\text{Modelo original: } Y_i = \alpha + \beta X_i + \mu_i \quad (1)$$

$$X_i = \phi + \varphi Y_i + \nu_i \quad (2)$$

Veamos que sucede cuando hay simultaneidad o endogeneidad de la variable explicativa. Supongamos que para un individuo cualquiera el término de error es negativo, es decir, el valor puntual de la variable dependiente esta por debajo del valor estimado, es decir, un valor negativo de μ_i disminuye el valor de Y_i. En la segunda ecuación si φ fuese negativo podemos ver que mientras menor es Y_i mayor es X_i, con lo cual podemos apreciar que existe una correlación negativa entre μ_i y X_i. De esta forma, la endogeneidad en la variable explicativa rompe con el supuesto de no correlación entre el término de error y las variables explicativas.

III.3. Error de medición

El error de medición es un problema con la recolección de los datos, el error de medición sólo genera problemas de sesgo e inconsistencia cuando las variables explicativas están medidas con error, cuando la variable dependiente es la que esta medida con error no se genera problema de sesgo.

A continuación podemos apreciar que cuando la variable explicativa esta medida con error, el término de error del modelo esta correlacionado con la variable explicativa incluida (variable con error), lo que invalida, nuevamente, el supuesto del estimador MCO sobre la no correlación entre el error y las variables explicativas.

Supongamos que en el siguiente modelo no observamos la variable explicativa X_i que debiésemos, sino una que esta medida con error, que llamaremos X_i^*. De esta forma:

$$X_i^* = X_i + \varepsilon_i$$

donde ε_i es el error de medición

Modelo "verdadero": $Y_i = \alpha + \beta X_i + \mu_i$

Modelo estimado: $Y_i = \alpha + \beta X_i^* + \underbrace{\nu_i}_{\mu_i - \beta\varepsilon_i}$

El modelo estimado no cumple con los supuestos MCO, ya que existe correlación distinta de cero entre el término de error compuesto ν_i y la variable medida con error X_i^*. El estimador MCO será sesgado e inconsistente.

III.4. Estimador de Variables Instrumentales (IV)

Supongamos el modelo básico de la ecuación de Mincer para estimar retornos a la educación:

$$lyph = \alpha + \beta \cdot esc_i + \mu_i$$

Si la correlación entre el término de error y la variable años de escolaridad es distinta de cero (por cualquiera de las tres razones antes mencionada), la estimación del retorno a la educación será sesgada e inconsistente.

La idea del estimador IV es buscar una variable **Z**, denominada **instrumento**, que permita aislar o separar la parte de los años de escolaridad que esta correlacionada con el error de la que no esta correlacionada con el error. Y luego utilizar sólo la parte de los años de escolaridad no correlacionada con el error para estimar correctamente el parámetro de interés a través de MCO.

El instrumento debe satisfacer dos condiciones para que sea un instrumento válido:

- Condición de relevancia: $Cov(esc_i, Z_i) \neq 0$
- Condición de exogeneidad: $Cov(Z_i, \mu_i) = 0$

Si el instrumento es relevante, entonces la variación del instrumento esta relacionada con la variación en la variable años de escolaridad. Adicionalmente, si el instrumento es exógeno, la parte de años de escolaridad que esta siendo capturada por el instrumento es justamente la parte exógena (o no correlacionada con el error) de años de escolaridad. De esta forma, un instrumento que es relevante y exógeno puede capturar el comportamiento de años de escolaridad que es exógeno, y esto puede ser utilizando para estimar consistentemente el retorno a la educación.

III.4.1. Estimador Mínimos Cuadrados Ordinarios en dos etapas (MCO2E)

El estimado MCO2E, tal como su nombre sugiere, es un estimador que consta de dos etapas. En la primera etapa se descompone la variable que tiene el problema de endogeneidad en dos partes, la que no esta correlacionada con el término de error y la que esta correlacionada con el término de error. De esta forma, la primera etapa consiste en hacer una regresión de la variable con problemas, en este caso años de escolaridad, con el instrumento:

PRIMERA ETAPA:

$$esc_i = \pi_0 + \pi_1 Z_i + \nu_i$$

Esta regresión permite hacer la descomposición de la variable escolaridad de la forma que necesitamos: una parte exógena ($\pi_0+\pi_1 Z_i$), que es la parte predicha por Z_i, si el instrumento cumple la condición de exogenidad esta predicción será justamente la parte exógena de la variable escolaridad, y otra parte que esta correlacionada con el error y es la que genera el problema de endogeneidad.

SEGUNDA ETAPA:

La segunda etapa consiste en estimar el modelo original, pero en vez de utilizar la variable escolaridad con problema, se utiliza la predicción del modelo de la primera etapa, a la cual se le ha "quitado" la parte que esta correlacionada con el término de error:

$$lyph = \alpha + \beta \cdot \hat{esc}_i + \mu_i$$

En un modelo de regresión múltiple, que incluye más de una variable explicativa, puede haber un grupo de ellos con problemas de endogeneidad y un grupo que no (exógenas). Para que se pueda utilizar el estimador de variables instrumentales debe haber al menos tantos instrumentos como variables con problema de endogeneidad. Si el número de instrumentos es exactamente igual al número variables endógenas se dice que los coeficientes de la regresión están exactamente identificados. Si el número de instrumentos es superior al número de variables con problemas se dice que los coeficientes están sobreidentificados.

III.5. Ejemplos de variables instrumentales

III.5.1. ¿Afecta la obligatoriedad de educación a la escolaridad e ingresos?, Angrist y Krueger (1991)

El artículo publicado en el año 1991 por Angrist y Krueger estima retornos a la educación a través de variables instrumentales. Como se explico anteriormente, la escolaridad e ingresos pueden tener un problema de endogeneidad. Además existe un problema potencial de omisión de variables relevante, habilidad, ambos aspectos genera que la escolaridad en el modelo de regresión este correlacionada con le término de error, lo que provoca sesgo e inconsistencia en el estimado MCO.

Este artículo explota la característica de experimento natural de la fecha de nacimiento, y como esto determina los años de escolaridad logrados, para estimar correctamente el retorno a la educación mediante variables instrumentales.

El instrumento utilizado en este caso para separar la parte de escolaridad exógena de la endógena, consiste en el **trimestre de nacimiento de la persona**.

¿Por qué el trimestre de nacimiento se puede utilizar como instrumento de los años de escolaridad?

La ley educacional en EEUU obliga a los estudiantes a permanecer en el colegio hasta la edad de 16 años, en el minuto que estos alumnos cumplen esta edad pueden abandonar el colegio. Sin embargo, para ingresar al colegio deben tener los seis años cumplidos al 1 de Enero del año de ingreso al colegio.
De esta forma, si comparamos dos niños uno nacido el 15 de Diciembre y otro el 15 de Enero del siguiente año, a pesar de que la diferencia en edad es sólo un mes, el segundo de ellos deberá esperar un año completo para poder ingresar al colegio, ingresando cuando tenga 7 años de edad y no a los 6 años de edad como el primero de los niños. Sin embargo, la ley permite que ambos abandonen el colegio a los 16 años, si ambos decidieran abandonar el colegio a los 16 años, el primero tendrá un año de educación más que el segundo de ellos.

Así, a priori, el trimestre de nacimiento es un instrumento que cumple con la condición de relevancia, y la condición de exogeneidad, debido a que el cumpleaños es poco probable que este correlacionado con otros atributos personales que puedan determinar el ingreso de la persona, sólo tiene influencia a través de su impacto en el nivel educacional logrado.

El modelo estimado por los autores tiene como variable dependiente el logaritmo del salario por hora, y como variables explicativas los años de escolaridad, dummy de

raza, variable dummy de área metropolitana, variable dummy si esta casado, 9 dummies para año de nacimiento, 8 dummies para región de residencia, 49 dummies de estado, edad y edad al cuadrado.

La estimación por MCO entrega un retorno a la educación de 5.7%. Cuando se realiza la estimación MCO2E, en la primera etapa se hace una regresión de los años de escolaridad (variable endógena) contra raza, área, y todas las variables explicativas incluidas en el modelo original distintas de la escolaridad, más tres dummies correspondientes al trimestre de nacimiento 1, 2 y 3, dummies que corresponden a los instrumentos de los años de escolaridad. El retorno a la educación estimado por esta metodología es de 3.9%.

III.5.2. Using Geographic Variation in College Proximity to Estimate the Return to Schooling, Card (1993)

Este artículo tiene como objetivo estimar el retorno a la educación, sin embargo, como ya mencionamos antes el nivel educacional y los ingresos presentan endogeneidad, el nivel educacional no es entregado aleatoriamente en la población, sino que depende de las decisiones tomadas sobre invertir o no en educación, las que dependen en parte del nivel de ingresos. De esta forma, para identificar correctamente el impacto que tiene la escolaridad sobre los ingresos se requiere una variación exógena en los años de escolaridad, es decir, requiere una variable instrumental que permita descomponer la escolaridad en la parte correlacionada con el término de error (endógena), y la parte no correlacionada con el error (exógena).
Este artículo utiliza como variable instrumental en la estimación de retornos a la educación una variable la presencia de una universidad en el área de residencia de la persona. Los estudiantes que crecieron en áreas donde no existen universidades presentan mayores costos de educación, ya que no tienen la posibilidad de seguir viviendo en sus casas. De esta forma, se espera que estos costos reduzcan la inversión en educación, al menos en las familias de menores ingresos.

En este artículo se estima la siguiente ecuación de salarios por hora:

$$lyph_i = \alpha + \beta_1 \cdot esc_i + \beta_2 \cdot \exp_i + \beta_3 \cdot \exp_i^2 + \beta_k \cdot X_k + \mu_i$$

Donde X_k incluye una serie de controles como: raza, área geográfica, educación de los padres, y estructura familiar.

La estimación por MCO del modelo anterior estima un retorno a la educación de 7.3%.

El estimador MCO2E utiliza como instrumento para la escolaridad una variable que indica que existe una universidad en el área donde vive la persona. El retorno a la educación estimado en este caso es de 13.2%.

III.5.3. Estimating the payoff to schooling using the Vietnam-era Daft lottery, Angrist y Krueger (1992)

Estos autores, nuevamente con el objetivo de estimar el retorno a la educación en forma correcta eliminando el problema de endogeneidad utilizan la metodología de variables instrumentales. Entre 1970 y 1973 la prioridad para servicio militar fue seleccionada aleatoriamente mediante una lotería. Muchos de los hombres que estimaban que podían ser seleccionados para el servicio militar se matricularon en los colegios para evadir el servicio militar, generando un mayor nivel educacional. Este artículo ocupa esta lotería como experimento natural para estimar el retorno a la educación.

El modelo estimado tiene como variable dependiente el logaritmo del salario por hora y como variable explicativa la escolaridad más un conjunto de regresores como estatus de veterano, raza, cuidad metropolitana, estado civil, dummies de año de

nacimiento, y dummies de regiones. La estimación MCO de este modelo entrega un valor estimado del retorno a la educación de 5.9%.
Luego para solucionar el problema de endogeneidad de los años de escolaridad, se estima primero un modelo de regresión entre los años de escolaridad como variable dependiente, y 130 dummies con la fecha de nacimiento para la lotería. Luego incorporando la predicción de escolaridad a partir de esta primera etapa, se obtiene un estimar del retorno a la educación de 6.5%.

III.6. Aplicación: "Wages of a Very Young Men", Griliches (1976)

Este autor tiene como objetivo estimar el retorno a la educación, para lo cual utiliza una muestra de 758 hombres jóvenes de la base de datos del estudio *National Longitudinal Survey of Young Men (NLS)*. Con el objetivo de eliminar el problema de omisión de variable relevante al no incluir la variable habilidad, el autor incorpora la variable IQ[4], que corresponde al test de coeficiente intelectual. El problema es que esta variable esta medida con error. Por lo cual, utiliza la metodología de variables instrumentales para solucionar el problema de sesgo que se genera en la estimación MCO cuando uno de los regresores esta medido con error.

La especificación estimada por el autor es la siguiente:

$$LW_i = \alpha + \beta_1 \cdot IQ_i + \beta_2 \cdot S_i + \beta_3 \cdot EXP_i + \beta_4 \cdot TENURE_i + \beta_5 \cdot RNS_i + \beta_6 \cdot SMSA_i + \beta_7 \cdot D67_i + \beta_8 \cdot D68_i + \beta_9 \cdot D69_i + \beta_{10} \cdot D70_i + \beta_{11} \cdot D71_i + \beta_{12} \cdot D73_i + \mu_i$$

LW: logaritmo del salario
IQ: coeficiente intelectual
S: años de escolaridad

[4] IQ corresponde al coeficiente intelectual. Este número resulta de la realización de un test estandarizado para medir habilidades cognitivas de las personas o "inteligencia", en relación con su grupo de edad. Se expresa en forma estandarizada para que la media en un grupo de edad sea 100.

EXP: experiencia laboral

TENURE: antigüedad en el empleo

RNS: variable dummy que toma valor 1 si la persona reside en el Sur

SMSA: variable dummy que indica si vive en zona urbana (metropolitan area)

D67-D73: dummies de año

La estimación MCO del modelo anterior es la siguiente:

```
. xi: reg lw  iq s expr tenure rns smsa i.year
i.year            _Iyear_66-73        (naturally coded; _Iyear_66 omitted)

      Source |       SS       df       MS              Number of obs =     758
-------------+------------------------------           F( 12,   745) =   46.86
       Model |  59.9127611    12  4.99273009           Prob > F      =  0.0000
    Residual |  79.3733888   745  .106541461           R-squared     =  0.4301
-------------+------------------------------           Adj R-squared =  0.4210
       Total |   139.28615   757  .183997556           Root MSE      =  .32641

------------------------------------------------------------------------------
          lw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |   .0027121   .0010314     2.63   0.009     .0006873    .0047369
           s |   .0619548   .0072786     8.51   0.000     .0476658    .0762438
        expr |   .0308395   .0065101     4.74   0.000     .0180592    .0436198
      tenure |   .0421631   .0074812     5.64   0.000     .0274763    .0568498
         rns |  -.0962935   .0275467    -3.50   0.001    -.1503719   -.0422151
        smsa |   .1328993   .0265758     5.00   0.000     .0807268    .1850717
   _Iyear_67 |  -.0542095   .0478522    -1.13   0.258    -.1481506    .0397317
   _Iyear_68 |   .0805808   .0448951     1.79   0.073    -.0075551    .1687168
   _Iyear_69 |   .2075915   .0438605     4.73   0.000     .1214867    .2936963
   _Iyear_70 |   .2282237   .0487994     4.68   0.000      .132423    .3240245
   _Iyear_71 |   .2226915   .0430952     5.17   0.000     .1380889     .307294
   _Iyear_73 |   .3228747   .0406574     7.94   0.000     .2430579    .4026915
       _cons |   4.235357   .1133489    37.37   0.000     4.012836    4.457878
------------------------------------------------------------------------------
```

De esta regresión se obtiene que para este grupo de jóvenes el retorno a la educación es de un 6.2%. Por otra parte, un punto adicional en el coeficiente intelectual aumenta el salario en un 0.27%.

Sin embargo, el autor indica que la variable IQ puede estar medida con error. El error de medición genera que la variable explicativa (IQ en este caso) y el término

de error estén correlacionados lo que genera problemas de sesgo en la estimación MCO de todos los parámetros del modelo. Para solucionar este problema se debe utilizar la metodología de variables instrumentales, que consiste en no utilizar directamente la variable IQ sino un instrumento que cumpla con la condición de relevancia y exogenidad.

El comando en STATA que permite la estimación de MCO2E (Variables Instrumentales) es el siguiente:

```
ivreg depvar [varlist1] (varlist2=instlist)
```

depvar: es la variable dependiente
varlist1: variables explicativas exógenas
varlist2: variables explicativas endógenas
instlist: variables instrumentales

Este comando lo que hace es estimar en una primera etapa una o más regresiones (porque las variables endógenas pueden ser más de una) de cada una de las variables endógenas (*varlist2*) contra las variables instrumentales (*instlist: excluded instruments*) y las variables explicativas exógenas del modelo original (*varlist1: included instruments*). Luego obtiene la predicción de las variables endógenas las que son incluidas como regresores junto con las variables explicativas exógenas *varlist1* en el modelo que tiene como variable dependiente a *depvar*.

En la base de datos se tiene información del nivel educacional de la madre (med), puntaje de otra prueba estandarizada (kww)[5], edad del entrevistado (age), y estado civil (mrt) que toma valor 1 si la persona esta casada. Estas cuatro variables pueden ser utilizadas como instrumentos para el coeficiente intelectual.

[5] KWW es una prueba estandarizada denominada Knowledge of the World of Work. Tal como su nombre lo dice es una pruba enfocada a información ocupacional y no a inteligencia propiamente tal.

A continuación se presenta la estimación del modelo anterior utilizando la metodología de MCO2E:

```
. ivreg lw s expr tenure rns smsa _I* (iq= med kww age mrt), first
```

La opción **first** del comando indica que muestre tanto la estimación de la primera como de la segunda etapa.

```
First-stage regressions
-----------------------

      Source |       SS       df       MS              Number of obs =     758
-------------+------------------------------           F( 15,   742) =   25.03
       Model |  47176.4676    15  3145.09784           Prob > F      =  0.0000
    Residual |  93222.8583   742  125.637275           R-squared     =  0.3360
-------------+------------------------------           Adj R-squared =  0.3226
       Total |  140399.326   757  185.468066           Root MSE      =  11.209

------------------------------------------------------------------------------
          iq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |    2.497742   .2858159     8.74   0.000     1.936638    3.058846
        expr |    -.033548   .2534458    -0.13   0.895    -.5311042    .4640082
      tenure |    .6158215   .2731146     2.25   0.024     .0796522    1.151991
         rns |   -2.610221   .9499731    -2.75   0.006    -4.475177   -.7452663
        smsa |    .0260481   .9222585     0.03   0.977    -1.784499    1.836595
   _Iyear_67 |    .9254935   1.655969     0.56   0.576    -2.325449    4.176436
   _Iyear_68 |    .4706951   1.574561     0.30   0.765    -2.620429     3.56182
   _Iyear_69 |    2.164635   1.521387     1.42   0.155    -.8221007     5.15137
   _Iyear_70 |    5.734786   1.696033     3.38   0.001     2.405191    9.064381
   _Iyear_71 |    5.180639   1.562156     3.32   0.001     2.113866    8.247411
   _Iyear_73 |    4.526686    1.48294     3.05   0.002     1.615429    7.437943
         med |    .2877745   .1622338     1.77   0.077    -.0307176    .6062665
         kww |    .4581116   .0699323     6.55   0.000     .3208229    .5954003
         age |   -.8809144   .2232535    -3.95   0.000    -1.319198   -.4426307
         mrt |    -.584791    .946056    -0.62   0.537    -2.442056    1.272474
       _cons |    67.20449   4.107281    16.36   0.000     59.14121    75.26776
------------------------------------------------------------------------------
```

Instrumental variables (2SLS) regression

```
      Source |       SS       df       MS              Number of obs =     758
-------------+------------------------------           F( 12,   745) =   45.91
       Model |  59.2679161    12  4.93899301           Prob > F      =  0.0000
    Residual |  80.0182337   745  .107407025           R-squared     =  0.4255
-------------+------------------------------           Adj R-squared =  0.4163
       Total |   139.28615   757  .183997556           Root MSE      =  .32773

------------------------------------------------------------------------------
          lw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |   .0001747   .0039374     0.04   0.965    -.0075551    .0079044
           s |   .0691759    .013049     5.30   0.000     .0435587    .0947931
        expr |    .029866    .006697     4.46   0.000     .0167189    .0430132
      tenure |   .0432738   .0076934     5.62   0.000     .0281705     .058377
         rns |  -.1035897   .0297371    -3.48   0.001    -.1619682   -.0452111
        smsa |   .1351148   .0268889     5.02   0.000     .0823277    .1879019
   _Iyear_67 |   -.052598   .0481067    -1.09   0.275    -.1470388    .0418428
   _Iyear_68 |   .0794686   .0451078     1.76   0.079     -.009085    .1680222
   _Iyear_69 |   .2108962   .0443153     4.76   0.000     .1238984    .2978939
   _Iyear_70 |   .2386338   .0514161     4.64   0.000     .1376962    .3395714
   _Iyear_71 |   .2284609   .0441236     5.18   0.000     .1418396    .3150823
   _Iyear_73 |   .3258944   .0410718     7.93   0.000     .2452642    .4065247
       _cons |    4.39955   .2708771    16.24   0.000     3.867777    4.931323
------------------------------------------------------------------------------
Instrumented:  iq
Instruments:   s expr tenure rns smsa _Iyear_67 _Iyear_68 _Iyear_69
               _Iyear_70 _Iyear_71 _Iyear_73 med kww age mrt
------------------------------------------------------------------------------
```

De la primera etapa podemos apreciar que 3 de los 4 instrumentos que no son parte del modelo original son significativos, la excepción es la dummy de estatus marital. Sin embargo, en la segunda etapa la variable IQ resulta no significativa, condicional a todos los demás factores IQ parece no tener un rol determinante en los salarios. Recordemos que los instrumentos debía cumplir dos criterios: relevancia y exogeneidad.

III.6.1. Validez de los instrumentos

Relevancia:

El criterio de relevancia se puede testear mediante el test F (significancia global) del modelo de la primera etapa, la regla es que este debe ser mayor a 10. Cuando los

instrumentos no pueden explicar el comportamiento de la variable endógena se dice que el instrumento es débil, y cuando esto sucede la distribución normal supuesta para la inferencia del estimador MCO2E no se cumple, por lo cual la inferencia es incorrecta. Con instrumentos débiles el estimar MCO2E no es confiable. En la estimación anterior, el Test F es superior a 10, lo que muestra la relevancia de los instrumentos seleccionados

Exogenidad:

Por otro lado, debemos chequear la exogenidad de los instrumentos, cuando esta no se cumple el estimador MCO2E será inconsistente, es decir, por más grande que sea el tamaño muestral el estimador no se acercará al "verdadero" valor del coeficiente. Además recuerde que la idea de las variables instrumentales es justamente que esta no tenga relación con el error para de esta forma poder estimar correctamente el impacto de la variable que se esta instrumentalizando.
No existe un test que permita testear explícitamente la exogenidad de los instrumentos, saber si un instrumento es o no exógeno depende de la opinión de los expertos y del conocimiento personal del problema empírico. Sin embargo, existen los llamados ***Test de sobreindentificación***. Lo que hace el test es computar el error de la estimación por MCO2E $\hat{\mu}_i^{MCO2E}$ y hacer una regresión de estos errores contra los instrumentos y las variables exógenas del modelo, si se rechaza la hipótesis nula de que los coeficientes de estas variables son cero, entonces se rechaza la hipótesis nula de que los instrumentos no están correlacionados con el término de error, es decir, se rechaza la hipótesis nula de exogeneidad de los instrumentos.

Una vez que se ha realizado la estimación MCO2E en STATA el comando **`overid`** computa el test de sobreidentificación Sargan y Basmann:

```
. overid

Tests of overidentifying restrictions:
Sargan N*R-sq test       87.655  Chi-sq(3)    P-value = 0.0000
Basmann test             97.025  Chi-sq(3)    P-value = 0.0000
```

En este caso ambos test muestran un rechaza de la hipótesis nula de exogeneidad de los instrumentos, es decir, no cumplen la segunda condición de validez de instrumentos.

Al igual que el estimador MCO, el estimador MCO2E o de variables instrumentales funciona bajo el supuesto de que los errores son *iid*. Si este supuesto no se cumple, la estimación sigue siendo insesgada pero deja de ser eficiente. Existe otro estimador denominado GMM que permite obtener estimaciones consistente y eficientes cunado los errores no son iid. El comando **ivreg2** es una extensión del comando anterior pero que permite la opción de estimar por el método de GMM. Adicionalmente, este comando entrega el test de sobreidentificación de Hansen (J test) inmediatamente.

Veamos los resultados:

```
. ivreg2 lw s expr tenure rns smsa _I* (iq= med kww age mrt), gmm

GMM estimation
--------------
                                                      Number of obs =      758
Total (centered) SS     =  139.2861498                Centered R2   =   0.4166
Total (uncentered) SS   =  24652.24662                Uncentered R2 =   0.9967
Residual SS             =  81.26217887                Root MSE      =      .33
------------------------------------------------------------------------------
             |               Robust
          lw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |  -.0014014   .0041131    -0.34   0.733     -.009463    .0066602
           s |   .0768355   .0131859     5.83   0.000     .0509915    .1026794
        expr |   .0312339   .0066931     4.67   0.000     .0181157    .0443522
      tenure |   .0489998   .0073437     6.67   0.000     .0346064    .0633931
         rns |  -.1006811   .0295887    -3.40   0.001    -.1586738   -.0426884
        smsa |   .1335973   .0263245     5.08   0.000     .0820021    .1851925
   _Iyear_67 |  -.0210135   .0455433    -0.46   0.645    -.1102768    .0682498
   _Iyear_68 |   .0890993    .042702     2.09   0.037     .0054049    .1727937
   _Iyear_69 |   .2072484   .0407995     5.08   0.000     .1272828     .287214
   _Iyear_70 |   .2338308   .0528512     4.42   0.000     .1302445    .3374172
   _Iyear_71 |   .2345525   .0425661     5.51   0.000     .1511244    .3179805
   _Iyear_73 |   .3360267   .0404103     8.32   0.000     .2568239    .4152295
       _cons |   4.436784   .2899504    15.30   0.000     3.868492    5.005077
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):        74.165
                                                  Chi-sq(3) P-val =   0.00000
------------------------------------------------------------------------------
Instrumented:  iq
Instruments:   s expr tenure rns smsa _Iyear_67 _Iyear_68 _Iyear_69 _Iyear_70
               _Iyear_71 _Iyear_73 med kww age mrt
------------------------------------------------------------------------------
```

En ambas estimaciones se rechaza la hipótesis nula de que los instrumentos son exógenos. El test Hansen J, o los dos reportados anteriormente (Sargan y Basmann) evalúan el conjunto completo de instrumentos. Por lo cual el test se puede ver afectado por una parte de estos instrumentos que realmente no son exógenos. El test C permite testear la exogeneidad de un subconjunto de los instrumentos. El estadístico de este test se obtiene de la diferencia de dos J test, el del modelo original con todos los instrumentos, y el del modelo en que sólo un subconjunto de instrumentos. En STATA se indica que instrumentos no deben ser considerados a

través de la opción **orthog()**. La hipótesis nula del test C es que las variables instrumentales especificadas en la opción **orthog()** son instrumentos apropiados.

```
. ivreg2 lw s expr tenure rns smsa _I* (iq= med kww age mrt), gmm orthog(age
mrt)

GMM estimation
--------------
                                                      Number of obs =      758
Total (centered) SS     =  139.2861498                Centered R2   =   0.4166
Total (uncentered) SS   =  24652.24662                Uncentered R2 =   0.9967
Residual SS             =  81.26217887                Root MSE      =      .33
------------------------------------------------------------------------------
             |               Robust
          lw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |  -.0014014   .0041131    -0.34   0.733     -.009463    .0066602
           s |   .0768355   .0131859     5.83   0.000     .0509915    .1026794
        expr |   .0312339   .0066931     4.67   0.000     .0181157    .0443522
      tenure |   .0489998   .0073437     6.67   0.000     .0346064    .0633931
         rns |  -.1006811   .0295887    -3.40   0.001    -.1586738   -.0426884
        smsa |   .1335973   .0263245     5.08   0.000     .0820021    .1851925
   _Iyear_67 |  -.0210135   .0455433    -0.46   0.645    -.1102768    .0682498
   _Iyear_68 |   .0890993    .042702     2.09   0.037     .0054049    .1727937
   _Iyear_69 |   .2072484   .0407995     5.08   0.000     .1272828     .287214
   _Iyear_70 |   .2338308   .0528512     4.42   0.000     .1302445    .3374172
   _Iyear_71 |   .2345525   .0425661     5.51   0.000     .1511244    .3179805
   _Iyear_73 |   .3360267   .0404103     8.32   0.000     .2568239    .4152295
       _cons |   4.436784   .2899504    15.30   0.000     3.868492    5.005077
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):        74.165
                                                     Chi-sq(3) P-val =  0.00000
C statistic (exogeneity/orthogonality of specified instruments):        72.989
                                                     Chi-sq(2) P-val =  0.00000
Instruments tested: age mrt
------------------------------------------------------------------------------
Instrumented:  iq
Instruments:   s expr tenure rns smsa _Iyear_67 _Iyear_68 _Iyear_69 _Iyear_70
               _Iyear_71 _Iyear_73 med kww age mrt
------------------------------------------------------------------------------
```

En esta estimación se testea si el subconjunto de instrumentos (age y mrt) son instrumentos válidos. El test C rechaza la hipótesis nula de que los instrumentos son apropiados. Por lo cual, la edad y la dummy de estado civil deben ser excluidos como instrumentos.

A continuación se presenta la estimación del modelo anterior pero sin considerar estos dos instrumentos que no cumplen con la condición de exogeneidad. Se puede apreciar tres diferencias importantes con respecto a la estimación anterior: los instrumentos incluidos cumplen con la condición de exogeneidad, el coeficiente asociada a IQ estimado por MCO2E es estadísticamente significativo, un punto adicional en el coeficiente intelectual aumenta el salario, en promedio, en 2.4%, y el coeficiente asociado a la escolaridad no es significativo.

```
. ivreg2 lw s expr tenure rns smsa _I* (iq= med kww), gmm

GMM estimation
--------------

                                                      Number of obs =       758
Total (centered) SS     =  139.2861498                Centered R2   =    0.1030
Total (uncentered) SS   =  24652.24662                Uncentered R2 =    0.9949
Residual SS             =  124.9413508                Root MSE      =       .41

------------------------------------------------------------------------------
             |               Robust
          lw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |   .0240417   .0060961     3.94   0.000     .0120936    .0359899
           s |   .0009181   .0194208     0.05   0.962    -.0371459     .038982
        expr |   .0393333   .0088012     4.47   0.000     .0220833    .0565834
      tenure |   .0324916   .0091223     3.56   0.000     .0146122     .050371
         rns |  -.0326157   .0376679    -0.87   0.387    -.1064433     .041212
        smsa |    .114463   .0330718     3.46   0.001     .0496434    .1792825
   _Iyear_67 |  -.0694178   .0568781    -1.22   0.222    -.1808968    .0420613
   _Iyear_68 |   .0891834   .0585629     1.52   0.128    -.0255977    .2039645
   _Iyear_69 |   .1780712   .0532308     3.35   0.001     .0737407    .2824016
   _Iyear_70 |    .139594   .0677261     2.06   0.039     .0068533    .2723346
   _Iyear_71 |   .1730151   .0521623     3.32   0.001      .070779    .2752512
   _Iyear_73 |    .300759   .0490919     6.13   0.000     .2045407    .3969772
       _cons |   2.859113   .4083706     7.00   0.000     2.058721    3.659504
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         0.781
                                                  Chi-sq(1) P-val =    0.37681
------------------------------------------------------------------------------
Instrumented:  iq
Instruments:   s expr tenure rns smsa _Iyear_67 _Iyear_68 _Iyear_69 _Iyear_70
               _Iyear_71 _Iyear_73 med kww
------------------------------------------------------------------------------
```

III.6.2. Test de endogeneidad Durbin-Wu-Hausman

La estimación por el método de variables instrumentales (MCO2E o GMM) tiene la ventaja de generar una estimación consistente en la presencia de endogeneidad. Sin embargo, la estimación a través de esta metodología genera una pérdida de eficiencia comparado con el estimador MCO. De esta forma, se debe buscar el equilibrio entre ganancias de consistencia y pérdida de eficiencia. El test de endogeneidad de Durbin-Wu-Hausman justamente testea la endogeneidad del modelo mediante la comparación de ambas estimaciones. La hipótesis nula del test es que el coeficiente MCO y el de variables instrumentales son similares, es decir, no existe un problema de endogeneidad.

Luego de estimar el modelo por el método de variables instrumentales se puede computar este test utilizando el comando **ivendog**:

```
. ivreg lw s expr tenure rns smsa _I* (iq= med kww)

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =     758
-------------+------------------------------           F( 12,   745) =   30.47
       Model |  13.1515039    12  1.09595866           Prob > F      =  0.0000
    Residual |  126.134646   745   .16930825           R-squared     =  0.0944
-------------+------------------------------           Adj R-squared =  0.0798
       Total |   139.28615   757  .183997556           Root MSE      =  .41147

------------------------------------------------------------------------------
          lw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |   .0243202   .0060513     4.02   0.000     .0124405    .0361998
           s |   .0004625   .0191587     0.02   0.981    -.0371489    .0380738
        expr |    .039129   .0085141     4.60   0.000     .0224146    .0558435
      tenure |   .0327048   .0097792     3.34   0.001     .0135067    .0519029
         rns |  -.0341617   .0386606    -0.88   0.377    -.1100584    .0417351
        smsa |   .1140326   .0338968     3.36   0.001      .047488    .1805771
   _Iyear_67 |  -.0679321   .0604394    -1.12   0.261     -.186584    .0507198
   _Iyear_68 |   .0900522   .0566543     1.59   0.112    -.0211689    .2012733
   _Iyear_69 |   .1794505    .055824     3.21   0.001     .0698595    .2890415
   _Iyear_70 |   .1395755   .0661226     2.11   0.035     .0097668    .2693843
   _Iyear_71 |   .1735613   .0559634     3.10   0.002     .0636966    .2834259
   _Iyear_73 |   .2971599   .0517334     5.74   0.000     .1955994    .3987204
       _cons |   2.837153   .4082421     6.95   0.000     2.035711    3.638595
------------------------------------------------------------------------------
```

```
Instrumented:  iq
Instruments:   s expr tenure rns smsa _Iyear_67 _Iyear_68 _Iyear_69 _Iyear_70
               _Iyear_71 _Iyear_73 med kww
--------------------------------------------------------------------------------

. ivendog

Tests of endogeneity of: iq
H0: Regressor is exogenous
    Wu-Hausman F test:                 21.83742  F(1,744)    P-value = 0.00000
    Durbin-Wu-Hausman chi-sq test:     21.61394  Chi-sq(1)   P-value = 0.00000
```

En este caso se rechaza la hipótesis nula de exogeneidad, la estimación MCO difiere significativamente de la estimación por variables instrumentales, indicando que la estimación MCO presenta problemas de inconsistencia producto de la endogeneidad en el modelo.

III.7. Referencias

Joshua D. Angrist and Alan B. Krueger (1991). "Does Compulsory School Attendance Affect Schooling and Earnings". *The Quarterly Journal of Economics*, Vol. 106, No. 4. (Nov., 1991), pp. 979-1014.

Joshua D. Angrist and Alan B. Krueger (1992). "Estimating the payoff to schooling using the Vietnam-era Daft lottery". NBER Working Paper N°4067.

Christopher F. Baum (2006). "An Introduction to Modern Econometrics Using STATA". Capítulo 8: Instrumental variables estimator.

David E. Card (1995). "Using Geographic Variation in College Proximity to Estimate the Return to Schooling". Aspects of Labour Market Behaviour: Essays in Honor of John Vanderkamp, eds. L. N. Christofides et al., Toronto: University of Toronto Press, pp. 201-221.

Zvi Griliches (1976). "Wages of a Very Young Men". The Journal of Political Economy, Vol. 84, N°4, Part 2: Essays in Labor Economics in Honor of H. Gregg Lewis. pp. S69-S86.

James H. Stock y Mark W. Watson. "Introducción to Econometrics". Capítulo 10: Instrumental Variables Regression.

Capítulo IV. Variable Dependiente Discreta

IV.1. Introducción

A pesar de que el modelo de regresión lineal visto hasta ahora, es el método más utilizado en las ciencias sociales, este método no logra satisfacer estructuras de modelos donde la relación entre la variable dependiente y las variables explicativas no es lineal. Vimos como la incorporación de potencias de las variables explicativas, de transformaciones logarítmicas, y de variables binarias interactuadas con variables continuas podían de alguna forma capturar no linealidades presentes, pero bajo el mismo contexto del modelo de regresión lineal.

Volvamos al comienzo del curso, donde se empezó hablando del análisis de regresión. El análisis de regresión busca estimar el promedio poblacional de la variable dependiente para valores fijos de las variables explicativas, es decir, estimar el promedio de la variable dependiente condicional en X, lo que se traduce a que el promedio condicional de variable dependiente en X se puede escribir como una función de X:

$$E(Y \mid X) = f(X)$$

El modelo de regresión lineal asume que esta función f(•) es lineal, es decir:

$$E(Y \mid X) = \alpha + \beta X$$

Sin embargo, este es un caso particular del análisis de regresión. El estimador MCO sólo es capaz de estimar modelos de regresiones lineales (primer supuesto), existen

otros estimadores para estimar modelos no lineal, donde la función f(•) toma cualquier otra forma, por ejemplo, el estimador **Máxima Verosímil**.

El estimador Máximo Verosimil es otro método para estimar la relación que existe entre la o las variables explicativas y la variable dependiente, la idea de este estimador es que la variable dependiente al ser una variable aleatoria tiene asociada una función de probabilidad la que depende de ciertos parámetros, por ejemplo, en el caso de una distribución normal estos parámetros son la media y la varianza. Entonces asumiendo una cierta distribución de la variable dependiente (normal, logística, etc.) se tiene que determinar los parámetros de esa distribución que hicen más probable la muestra que observamos. A la función de densidad en este contexto de le llama función de verosimilitud, ya que se toma como dado los daos y desconocidos los parámetros, contrario a lo que uno usualmente entiende por función de densidad, donde se conocen los parámetros y se generan datos a partir de ella. En resumen, el estimador de máxima verosmilitud parte asumiendo una función de densidad de la variable dependiente, la que se llama función de verosimilitud, y el objetivo es determinar o estimar los parámentros de esta función que hicieron que la muestra que observamos sea la más probable.

Cuando la variable dependiente es binaria (toma valores 1 y 0 solamente), la relación entre las variables explicativas y la variable dependiente es claramente no lineal, de hecho el valor esperado de la variable dependiente representa la probabilidad de que la variable sea igual a 1, y como han visto en sus cursos de estadística, la función de probabilidad acumulada tiene una forma no lineal, de esta forma, lo más apropiado para la correcta estimación de un modelo con variable dependiente discreta es utilizar le método de máxima verosimilitud, modelos que reciben el nombre de **probit** (si el error se supone con distribución normal) y **logit** (si se asume un error con distribución logística).

Algunos ejemplos de modelos de variable dependiente discreta son:

- Decisión de estudiar en colegios privados versus públicos
- Decisión de otorgar o no un crédito a una empresa
- Decisión de las personas de capacitarse o no
- Decisión de las personas de ahorrar o no (o de endeudarse o no)
- Factores asociados a la depresión
- Decisión de contribuir o no al sistema de pensiones
- Decisión de tener o no un seguro

Para la presente clase se utilizará la encuesta CASEN 2006 para estudiar los determinantes de que una persona realice o no una capacitación laboral. En esta encuesta se pregunta a las personas si entre Noviembre de 2005 y Octubre de 2006 han asistido a algún curso de capacitación laboral. Plantearemos un modelo simple para analizar la relación entre la realización de capacitación laboral y un conjunto de variables demográficas y características del empleo de los ocupados, por lo cual sólo se tomará como muestra de análisis los ocupados como asalariados. Según los datos de la ENCUESTA CASEN 2006, un 49,7% de los mayores de 15 años (población en edad de trabajar) se encuentran ocupados. Del total de personas ocupadas, un 69,1% trabaja como asalariado, y de los asalariados un 21,4% ha realizado algún curso de capacitación en el último año. Las características individuales que se utilizarán en la estimación son: género, edad, escolaridad, estado civil, y condición de jefe de hogar. Además se utilizarán algunas características del empleo como: ingreso laboral por hora, tamaño de la empresa y rama de actividad económica.

IV.2. Modelo de probabilidad lineal

El modelo de probabilidad lineal consiste en aplicar MCO, es decir, estimar un modelo lineal cuando la variable dependiente es discreta. En nuestra aplicación la variable dependiente tiene la siguiente forma:

$$Y_i = \begin{cases} 1 & \text{si ha asistido a alguna capacitación} \\ 0 & \text{si no ha asistido a alguna capacitación} \end{cases}$$

El modelo de regresión lineal estimado en este caso es[6]:

$$Y_i = \beta_0 + \beta_1 \cdot esc_i + \beta_2 \cdot edad_i + \beta_3 \cdot casado_i + \beta_4 \cdot jefe_i + \beta_5 \cdot genero_i + \beta_6 \cdot lyph_i + \beta_7 \cdot Egrande_i + \beta_8 \cdot Emediana_i + \beta_9 \cdot industria_i + \beta_{10} \cdot \min eria_i + \beta_{11} \cdot comercio_i + \beta_{12} \cdot servicios_i + \beta_{13} \cdot construccion_i + \beta_{14} \cdot transporte_i + \beta_{15} \cdot electr_i + \beta_{16} \cdot sevcomu_i + \mu_i$$

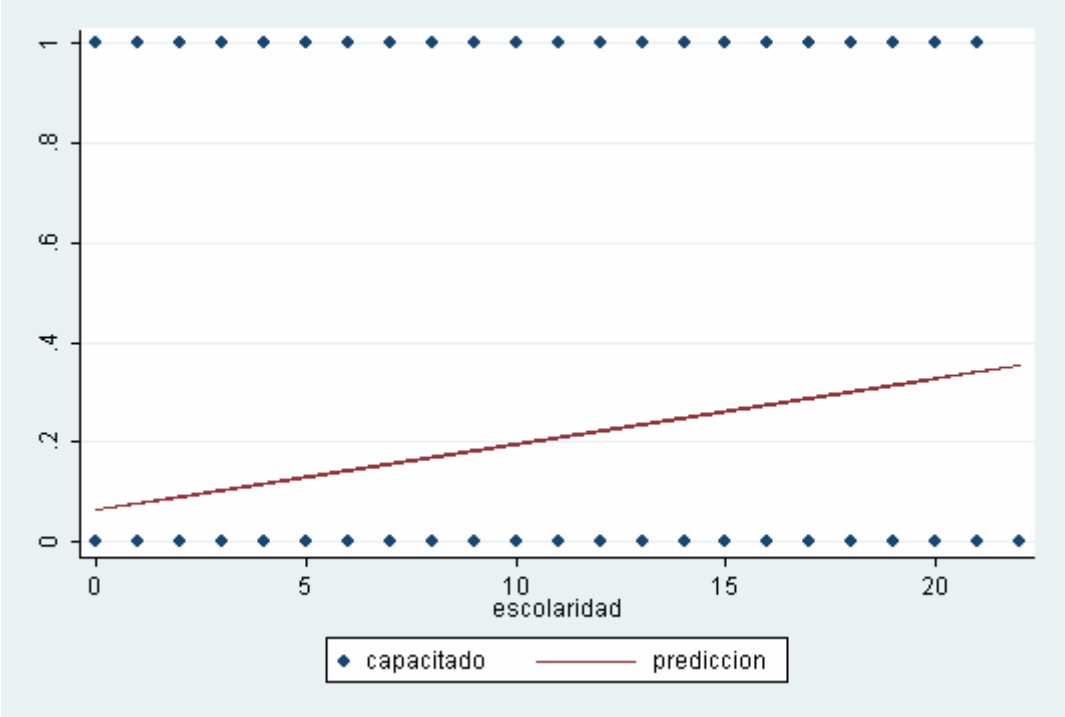
La siguiente tabla nos muestra los resultados de la estimación del modelo anterior por MCO utilizando la opción robust:

[6] donde:
esc: años de escolaridad
edad: edad en años cumplidos
casado: dummy que toma valor 1 si la persona esta casada y cero sino
jefe: dummy qye toma valor 1 si la persona es jefe de hogar y cero sino
genero: dummy que toma valor 1 si la persona es hombre y cero si es mujer
lyph: logaritmo natural del ingreso por hora
Emediana: dummy que toma valor 1 si la empresa tiene entre 10 y 49 trabajadores
Egrande: dummy que toma valor 1 si la empresa tiene 50 trabajadores y más
industria: dummy que toma valor 1 si trabaja en sector industria
mineria: dummy que toma valor 1 si trabaja en mineria
comercio: dummy que toma valor 1 si trabaja en sector comercio
servicios: dummy que toma valor 1 si trabaja en servicios financieros
construccion: dummy que toma valor 1 si trabaja en construcción
transporte: dummy que toma valor 1 si trabaja en sector transportes
electr: dummy que toma valor 1 si trabaja en sector electricidad, gas y agua.
servcomu: dummy que toma valor 1 si trabaja en sector servicios comunales

```
Linear regression                               Number of obs =   54066
                                                F( 16, 54049) =  427.62
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1192
                                                Root MSE      =  .36319

------------------------------------------------------------------------------
             |               Robust
  capacitado |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         esc |   .0131787   .0005115    25.76   0.000     .0121761    .0141812
        edad |  -.0002417   .0001516    -1.59   0.111    -.0005388    .0000554
      casado |   .0350264   .0037339     9.38   0.000      .027708    .0423447
        jefe |   .0616054   .0040919    15.06   0.000     .0535852    .0696256
      genero |  -.0322881   .0040074    -8.06   0.000    -.0401427   -.0244335
        lyph |   .0652031   .0030554    21.34   0.000     .0592145    .0711917
     Egrande |   .1271726   .0035704    35.62   0.000     .1201746    .1341707
    Emediana |   .0430007   .0035718    12.04   0.000     .0359999    .0500014
     mineria |   .1356752   .0120397    11.27   0.000     .1120773    .1592731
   industria |   .0287321   .0052282     5.50   0.000     .0184849    .0389794
      electr |   .0927762   .0202167     4.59   0.000     .0531512    .1324011
construccion |  -.0393508   .0051154    -7.69   0.000     -.049377   -.0293247
    comercio |   .0298023   .0053396     5.58   0.000     .0193367     .040268
  transporte |   .0056154   .0069195     0.81   0.417    -.0079468    .0191776
   servicios |   .0480401   .0088693     5.42   0.000     .0306562     .065424
    servcomu |   .0700014   .0053381    13.11   0.000     .0595386    .0804642
       _cons |  -.5082504   .0185669   -27.37   0.000    -.5446416   -.4718592
------------------------------------------------------------------------------
```

Tomando todo las variables explicativas constante (evaluados en algún valor) excepto la variable escolaridad podemos graficar la relación estimada entre la variable dependiente binaria que toma valor 1 si la persona ha realizado cursos de capacitación y los años de escolaridad. Por ejemplo, tomemos la edad en 34, una persona casada, jefe de hogar, hombre, que trabaja en una empresa mediana en sector comercio, el siguiente gráfico muestra la relación entre el valor observado de la variable dependiente, el valor predicho de la variable dependiente, y los años de escolaridad.

```
g prediccion=_b[_cons]+_b[esc]*esc+_b[edad]*34+_b[casado]*1+_b[jefe]*1
+_b[genero]*1+_b[lyph]*6.8+_b[Emediana]*1+_b[comercio] if e(sample)

twoway (scatter capacitado esc if prediccion!=.) (line prediccion esc)
```



¿Qué representa el valor estimado de la variable dependiente?

Notemos que cuando realizamos la estimación por MCO, lo que obtenemos es el valor estimado para la media poblacional de Y condicional en X. Cuando la variable dependiente es binaria, la media de esta variable representa la probabilidad de que la variable dependiente tome valor igual a 1. Por lo cual, el valor estimado para la variable dependiente condicional en las variables explicativas representa justamente la probabilidad estimada de haber realizado una capacitación:

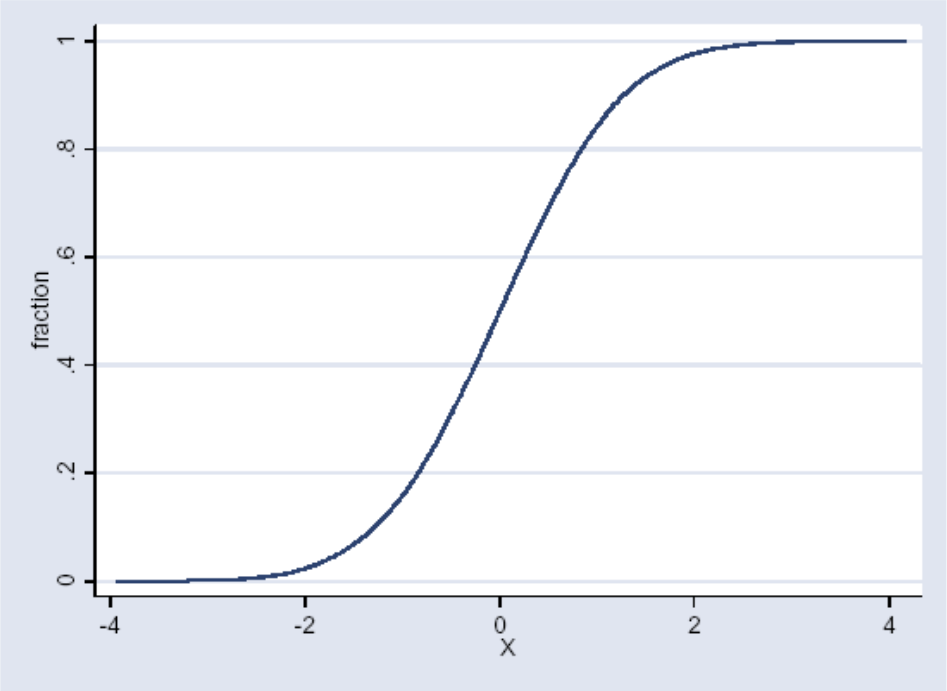
$$E(Y_i \mid X_i) = 1 \cdot \Pr(Y_i = 1 \mid X_i) + 0 \cdot [1 - \Pr(Y_i = 1 \mid X_i)] = \Pr(Y_i = 1 \mid X_i)$$

¿Qué miden los coeficientes β?

Los coeficientes del modelo miden el impacto de la variable explicativa que acompañan sobre la probabilidad de que la variable dependiente sea igual a 1:

$$\frac{\partial E(Y_i \mid X_i)}{\partial X_k} = \frac{\partial \Pr(Y_i = 1 \mid X_i)}{\partial X_k} = \beta_k$$

La recta de regresión estimada indica como va aumentando la probabilidad de que la variable de dependiente sea igual a 1 a medida que aumenta el valor de la variable explicativa X, el concepto estadístico asociado a esto es la función de distribución acumulada. Por esta razón, y algunas otras que mencionaremos a continuación, la estimación lineal realizada por MCO no es apropiada en este tipo de problemas, ya que la función de distribución acumulada es claramente no lineal.



Del modelo estimado anteriormente se obtienen los siguientes resultados:

- Un año adicional de escolaridad aumenta la probabilidad de haber realizado un curso de capacitación en 1.3 puntos porcentuales, es decir, si por ejemplo la probabilidad de haber realizado una capacitación con 10 años de escolaridad es 15.3%, tener 11 años de escolaridad aumenta la probabilidad de haber realizado una capacitación a 16.6%.
- La edad no tiene un impacto significativo sobre la probabilidad de haber realizado una capacitación.
- Pasar de no estar casado a estar casado aumenta la probabilidad de realizar una capacitación en 3.5 puntos porcentuales.
- Ser jefe de hogar aumenta la probabilidad de haber realizado una capacitación en 6.2 puntos porcentuales, con respecto a no ser jefe de hogar.

- Ser hombre disminuye la probabilidad de haber realizado una capacitación en 3.2 puntos porcentuales, con respecto a las mujeres.
- Un aumento de un 1% en el salario por hora, aumenta la probabilidad de haber realizado una capacitación en 6.5 puntos porcentuales.
- Trabajar en una empresa grande (50 trabajadores y más) aumenta la probabilidad de haber realizado una capacitación en 12.7 puntos porcentuales respecto a las personas que trabajan en una empresa pequeña (menos de 10 trabajadores).
- Trabajar en una empresa mediana (entre 10 y 49 trabajadores) aumenta la probabilidad de haber realizado una capacitación en 4.3 puntos porcentuales respecto a las personas que trabajan en una empresa pequeña (menos de 10 trabajadores).
- Todas los coeficientes de las dummies de actividad económica de la empresa se deben interpretar en función de la dummy que fue retirada de la estimación (agricultura), con respecto a estas variables se puede concluir que: el sector minería tienen 13.6 puntos porcentuales más de probabilidad de haber realizado una capacitación que el sector agricultura, el sector industria 2.9 puntos más, electricidad 9.3 puntos más, construcción 3.9 puntos menos, comercio 3.0 puntos porcentuales más, transporte 0.6 puntos porcentuales más pero no significativo, servicios financieros 4.8 puntos porcentuales más de probabilidades de haber realizado una capacitación, y servicios comunales 7 puntos porcentuales más.

¿Cuáles son los problemas del modelo de probabilidad lineal?

(i) El término de error es heterocedástico (DEMOSTRAR)
(ii) Los errores no tienen una distribución normal
(iii) Las predicciones que se obtiene pueden no estar dentro del rango posible de interpretación.
(iv) El modelo supone linealidad.

Como se mencionó anteriormente, cuando la variable dependiente es binaria, el valor estimado para esta variable condicional en las variables explicativas representa la probabilidad de que la variable dependiente tome valor 1. Así, la mejor forma de modelar este tipo de modelos no es a través de una función lineal, sino a través de una función de probabilidad acumulada:

$$\Pr(Y_i \mid X_i) = F(X_i\beta)$$

Si se asume una función de probabilidad normal el modelo de denomina PROBIT, si se asume una función de probabilidad logística el modelo se denomina LOGIT.

Notemos que los coeficientes β estimados no miden directamente el efecto marginal de la variable explicativa sobre la probabilidad de que la variable dependiente sea igual a uno, ya que el coeficiente β forma parte del argumento de la función de distribución acumulada F(•). El efecto marginal de la variable explicativa sobre la probabilidad se debe obtener de la siguiente forma:

$$\frac{\partial \Pr(Y=1 \mid X)}{\partial X_k} = \frac{\partial F(X\beta)}{\partial X_k} \cdot \beta_k = f(X\beta) \cdot \beta_k$$

Donde f(•) es la función de densidad, la cual debe ser evaluada en Xβ, por lo cual se deben elegir valores de las variables explicativas sobre las cuales evaluar.

IV.3. Los modelos PROBIT y LOGIT

El modelo de probabilidad lineal analizado en la sección anterior es simple de estimar y usar, sin embargo, tiene características que lo hace poco apropiado: la probabilidad predicha puede estar fuera del rango (0,1), y el efecto marginal de las variables explicativas es constante. Por esta razón se utilizan los modelos de PROBIT o LOGIT (modelos de variable dependiente binaria), los que son estimados por máxima verosimilitud.

Recordemos que para poder obtener el estimador de máxima verosimilitud debemos computar la función de densidad conjunta de las N observaciones en la muestra. Para nuestro modelo de variable dependiente discreta, esta función de probabilidad conjunta de la muestra se puede expresar de la siguiente forma:

$$\Pr(Y \mid X) = \prod_{i=1}^{N} \Pr(Y_i \mid X_i)$$

En el modelo de variable dependiente binaria los valores de la variable dependiente sin sólo dos, uno y cero. Hemos visto también que la probabilidad de que la variable dependiente tome valor 1 condicional en X se puede escribir como la función de distribución acumulada evaluada en Xβ:

$$\Pr(Y_i = 1 \mid X_i) = F(X_i \beta)$$
$$\Pr(Y_i = 0 \mid X_i) = 1 - F(X_i \beta)$$

Así, la probabilidad de observar la muestra se puede escribir de la siguiente forma:

$$\Pr(Y \mid X) = \prod_{y_i=1} F(X_i\beta) \prod_{y_i=0} [1 - F(X_i\beta)]$$

Expresión que también se puede escribir de la siguiente forma:

$$\Pr(Y \mid X) = \prod_{i=1}^{N} [F(X_i\beta)]^{y_i} [1 - F(X_i\beta)]^{1-y_i}$$

Ambas expresiones anteriores, representan la función de verosimilitud que se debe optimizar con respecto a los parámetros desconocidos (β). Tomando logaritmo natural de la expresión anterior se obtiene:

$$\ln L = \sum_{i=1}^{N} y_i \ln F(X_i\beta) + (1 - y_i) \ln(1 - F(X_i\beta))$$

Maximizando la expresión con respecto a β, obtenemos la siguiente condición de primer orden:

$$\frac{\partial \ln L}{\partial \beta_k} = \sum_{i=1}^{N} \left[\frac{y_i f(X_i\beta)}{F(X_i\beta)} - \frac{(1 - y_i) f(X_i\beta)}{(1 - F(X_i\beta))} \right] \cdot X_k = 0$$

La ecuación no tiene una solución matemática cerrada, por lo cual se deben utilizar métodos iterativos para encontrar la solución, estos procesos los realizan los softwares computacionales.

La única diferencia entre el modelo PROBIT y el modelo LOGIT es la función de densidad supuesta. En el modelo PROBIT se supone una función de densidad normal estándar:

$$f(X_i\beta) = \phi(X_i\beta) = \frac{1}{\sqrt{2\pi}} e^{\frac{-\mu_i^2}{2}}$$

$$F(X_i\beta) = \Phi(X_i\beta) = \int_{-\infty}^{\mu_i} \frac{1}{\sqrt{2\pi}} e^{\frac{-\mu_i^2}{2}}$$

El modelo LOGIT supone una función de densidad logística:

$$f(X_i\beta) = \lambda(X_i\beta) = \frac{e^{\mu_i}}{(1+e^{\mu_i})^2}$$

$$F(X_i\beta) = \Lambda(X_i\beta) = \frac{e^{\mu_i}}{(1+e^{\mu_i})}$$

IV.4. Estimación de la probabilidad de capacitarse con modelos de variable dependiente discreta

Volvamos al modelo estimado en la sección IV.2, para estudiar los determinantes de los asalariados para realizar una capacitación. Primero utilicemos el modelo PROBIT:

```
Iteration 0:   log likelihood = -25755.968
Iteration 1:   log likelihood =  -22474.46
Iteration 2:   log likelihood = -22374.012
Iteration 3:   log likelihood = -22373.367
Iteration 4:   log likelihood = -22373.367

Probit regression                                 Number of obs   =      54066
                                                  LR chi2(16)     =    6765.20
                                                  Prob > chi2     =     0.0000
Log likelihood = -22373.367                       Pseudo R2       =     0.1313

------------------------------------------------------------------------------
  capacitado |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         esc |   .0571299   .0024712    23.12   0.000     .0522864    .0619734
        edad |   -.002937   .0006914    -4.25   0.000    -.0042922   -.0015819
      casado |   .1719956   .0156521    10.99   0.000     .1413181    .2026731
        jefe |   .2758226    .016909    16.31   0.000     .2426816    .3089636
      genero |     -.1654   .0168968    -9.79   0.000    -.1985171    -.132283
        lyph |   .2187762   .0124129    17.62   0.000     .1944474    .2431049
     Egrande |   .6335182   .0197989    32.00   0.000      .594713    .6723235
    Emediana |   .3068984   .0219413    13.99   0.000     .2638943    .3499025
     mineria |   .4962219   .0371671    13.35   0.000     .4233757     .569068
   industria |   .1911305   .0236132     8.09   0.000     .1448494    .2374116
      electr |   .4233926   .0647481     6.54   0.000     .2964887    .5502966
construccion |  -.1276754   .0294515    -4.34   0.000    -.1853992   -.0699516
    comercio |   .2018357   .0244174     8.27   0.000     .1539785    .2496928
  transporte |   .1027201   .0310052     3.31   0.001     .0419509    .1634892
   servicios |   .2536718   .0321939     7.88   0.000     .1905729    .3167707
    servcomu |   .2944947   .0227859    12.92   0.000     .2498352    .3391542
       _cons |  -3.631647   .0734629   -49.44   0.000    -3.775632   -3.487663
------------------------------------------------------------------------------
```

El impacto de cada variable explicativa sobre la probabilidad de haberse capacitado no se puede interpretar directamente de la estimación anterior, los valores de este output son los coeficientes, los que no representan el efecto marginal en este tipo de modelos. Para obtener los efectos marginales debemos utilizar el comando **mfx** después de haber estimado el modelo utilizando el comando PROBIT o LOGIT:

```
Marginal effects after probit
      y  = Pr(capacitado) (predict)
         =  .14891232
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.      z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
     esc |   .0132558      .00057   23.27   0.000   .012139  .014373   10.2672
    edad |  -.0006815      .00016   -4.25   0.000  -.000996 -.000367   38.1447
  casado*|   .0391724      .00349   11.22   0.000   .032329  .046015   .603836
    jefe*|   .0644845      .00397   16.24   0.000   .056702  .072267   .474328
  genero*|  -.0396499      .00417   -9.50   0.000  -.047827 -.031473    .69221
    lyph |   .0507624      .00289   17.57   0.000   .045099  .056426   6.74595
 Egrande*|   .1498894      .00466   32.19   0.000   .140763  .159016   .471072
Emediana*|   .0761954      .00575   13.25   0.000   .064924  .087466   .280417
 mineria*|   .1423699      .01254   11.36   0.000   .117797  .166942   .029316
indust~a*|   .0476349      .00627    7.59   0.000   .035337  .059933   .127474
  electr*|   .1191014      .02122    5.61   0.000   .077507  .160696   .008878
constr~n*|  -.0280096      .00609   -4.60   0.000  -.039954 -.016065   .087745
comercio*|   .0504774      .00653    7.73   0.000   .037681  .063274   .129397
transp~e*|   .0249543      .00786    3.17   0.002    .00954  .040369   .060944
servic~s*|   .0658449      .00922    7.14   0.000   .047781  .083909   .049125
servcomu*|    .074579      .00623   11.97   0.000   .062364  .086794   .200884
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

¿Qué se puede concluir de la estimación anterior?

- Todas las variables resultan ser estadísticamente significativas
- Aumentar la escolaridad en un año aumenta la probabilidad de haber realizado una capacitación en 1.3 puntos porcentuales.
- La edad disminuye la probabilidad de haber realizado una capacitación en 0.07 puntos porcentuales
- Estar casado aumenta la probabilidad en 3.9 puntos porcentuales
- Ser jefe de hogar también aumenta la probabilidad de realizar capacitación en 6.4 puntos porcentuales.
- Ser hombre disminuye la probabilidad en 4 puntos porcentuales.

- Un 1% más de salario por hora aumenta la probabilidad de capacitarse en 5 puntos porcentuales.
- Trabajar en una empresa grande versus una empresa pequeña aumenta la probabilidad de capacitarse en 15 puntos porcentuales.
- Trabajar en una empresa mediana versus una empresa pequeña aumenta la probabilidad de capacitarse en 7.6 puntos porcentuales.
- Con respectos a los sectores económicas (todos evaluados versus el sector agricultura) se concluye que: mineria aumenta la probabilidad en 14.2 puntos porcentuales, industria aumenta la probabilidad en 4.8 puntos porcentuales, electricidad la aumenta en 11.9 puntos porcentuales, construcción disminuye la probabilidad en 2.8 puntos porcentuales, comercio aumenta la probabilidad en 5 puntos porcentuales, transporte aumenta la probabilidad en 2.5 puntos porcentuales, servicios financieros aumenta la probabilidad en 6.6 puntos porcentuales, y servicios comunales aumenta la probabilidad en 7.5 puntos porcentuales.

La siguiente tabla muestra los efectos marginales del modelo utilizando LOGIT:

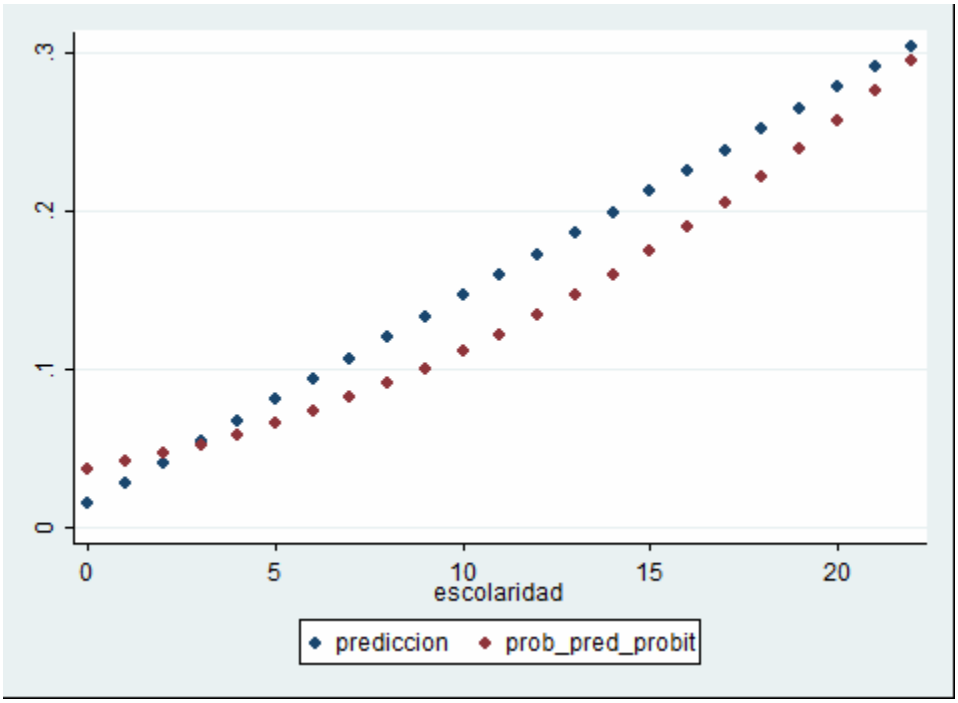
```
Marginal effects after logit
      y  = Pr(capacitado) (predict)
         =  .14396798
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
     esc |   .0122621      .00055   22.48   0.000   .011193  .013331   10.2672
    edad |  -.0007206      .00015   -4.76   0.000  -.001017 -.000424   38.1447
  casado*|   .0381315      .00328   11.62   0.000   .031698  .044565   .603836
    jefe*|   .0613547      .00378   16.24   0.000   .053949  .068761   .474328
  genero*|  -.0384177        .004   -9.62   0.000  -.046248 -.030587    .69221
    lyph |   .0444915      .00271   16.39   0.000   .039171  .049812   6.74595
 Egrande*|   .1519367      .00491   30.95   0.000   .142315  .161559   .471072
Emediana*|   .0838158      .00622   13.48   0.000   .071628  .096004   .280417
 mineria*|   .1416881      .01259   11.25   0.000    .11701  .166366   .029316
indust~a*|   .0500673      .00634    7.90   0.000   .037649  .062486   .127474
  electr*|   .1185053      .02144    5.53   0.000   .076482  .160529   .008878
constr~n*|  -.0249567      .00601   -4.16   0.000  -.036727 -.013186   .087745
comercio*|   .0524467      .00661    7.94   0.000   .039495  .065398   .129397
transp~e*|   .0280152      .00783    3.58   0.000   .012665  .043366   .060944
servic~s*|   .0664822       .0092    7.23   0.000   .048451  .084514   .049125
servcomu*|   .0737426      .00625   11.81   0.000     .0615  .085985   .200884
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

La siguiente tabla muestra la comparación de ambos modelos:

	PROBIT	LOGIT
esc	0.0133	0.0123
edad	-0.0007	-0.0007
casado*	0.0392	0.0381
jefe*	0.0645	0.0614
genero*	-0.0396	-0.0384
lyph	0.0508	0.0445
Egrande*	0.1499	0.1519
Emediana*	0.0762	0.0838
mineria*	0.1424	0.1417
industria*	0.0476	0.0501
electr*	0.1191	0.1185
construccion*	-0.0280	-0.0250
comercio*	0.0505	0.0524
transporte*	0.0250	0.0280
servicios*	0.0658	0.0665
servcomu*	0.0746	0.0737

Al igual que con el modelo de probabilidad lineal se puede graficar la relación entre la probabilidad de capacitarse y los años de escolaridad a a través de los siguientes comandos:

```
g
pred_probit=_b[_cons]+_b[esc]*esc+_b[edad]*33.8+_b[casado]*0.42+_b[jefe]*0.27+_
b[genero]*0.5+_b[lyph]*6.8+_b[Egrande]*0.31+_b[Emediana]*0.18+_b[mineria]*0.02+
_b[industria]*0.11+_b[electr]*0.006+_b[construccion]*0.08+_b[comercio]*0.15+_b[
transporte]*0.06+_b[servicios]*0.04+_b[servcomu]*0.22 if e(sample)

g prob_pred_probit=normal(pred_probit)

twoway (scatter prediccion esc) (scatter prob_pred_probit esc)
```



Con respecto a la bondad de ajuste del modelo, existen dos medidas que nos permiten ver que tan bueno es el modelo estimado: porcentaje predicho correctamente, y el Pseudos-R^2.

(a) Proporción de las observaciones predichas correctamente: para poder calcular el porcentaje de observaciones para las cuales se predijo correctamente, primero se debe computar la probabilidad predicha por el modelo (**`predict pred, p`**):

$$\hat{p} = F(X_i \hat{\beta})$$

Luego generar una variable que tome valor 1 si p-gorro es mayor o igual a 0.5 y que tome valor cero si es menor a 0.5. Luego generar una variable que indique si la observación fue predicha en forma correcta, esta variable tomará valor 1 cuando la observación de la variable dependiente toma valor 1 y fue predicho que tomaba valor 1 (p-gorro mayor o igual a 0.5), también será igual a 1 cuando la variable dependiente toma valor cero y fue predicho que tomaba valor cero (p-gorro menor a 0.5), y tomará valor cero en los casos que se predijo en forma incorrecta.

(b) Pseudos-R^2: esta medida de bondad de ajuste compara la función de verosimilitud con las variables explicativas del modelo y la función de verosimilitud sin las variables del modelo (sólo la constante):

$$Pseudo - R^2 = 1 - \frac{\ln(L_{NR})}{\ln(L_0)}$$

Donde:

L_{NR}: representa la función de verosimilitud no restringida, incluyendo todas las variables del modelo.

L_0: representa la función de verosimilitud sólo con la constante.

Notemos que si las variables explicativas del modelo no explican nada, la razón de verosimilitudes es 1 y el pseudos-R^2 es igual a cero, en la medida

que las variables explican algo de la variable dependiente, la verosimilitud no restringida es menos negativa que la restringida, y esta razón es menor a 1, y la medida de bondad de ajuste mayor a cero.

Capítulo V. Variable Dependiente Categórica ordinal y no ordinal

V.1. Introducción

Continuando con la discusión de la clase anterior, que nos permitió modelar y estimar casos en que la variable dependiente era discreta y de carácter binario, en esta clase abordamos el caso en que esta variable es categórica y ordinal.

Es decir, nos interesa ahora una variable dependiente que está dividida en varias categorías, las que pueden ser ordenadas (puestas en un ranking).

Por ejemplo, queremos conocer cuáles son los aspectos que explican el autoreporte de salud de los individuos. En la Encuesta de Protección Social, por ejemplo, los entrevistados ordenan su salud en 6 categorías: 1) Excelente; 2) Muy buena; 3) Buena; 4) Regular; 5) Mala; y 6) Muy mala. Se trata, obviamente, de una variable ordinal puesto que al ir desde 1 a 6 claramente la salud de la persona es peor, y viceversa. Otra variable ordinal que podría ser objeto de nuestro interés, es el número de personas que trabajan en la empresa. En la encuesta CASEN, por ejemplo, la respuesta a esta pregunta consta de las siguientes categorías: A) 1 persona; B) de 2 a 5 personas; C) de 6 a 9 personas; D) de 10 a 49 personas; E) de 50 a 199 personas; F) 200 y más personas.

Todas estas variables ordinales, normalmente, son codificadas con valores 1, 2, 3,…,etc; quizás con el objetivo de utilizar la variable codificada de este forma para realizar un análisis de regresión lineal. Sin embargo, esta forma de estimar asume implícitamente que los intervalos entre las categorías son iguales. Por ejemplo, la diferencia entre estar muy de acuerdo y de acuerdo, se asume igual que a la diferencia entre estar de acuerdo e indiferente. Esto hace que los resultados de la

estimación por MCO tengan resultados no apropiados[7]. De esta forma, cuando la variable dependiente es de carácter ordinal, es mejor no asumir que las diferencias entre categorías son las mismas. En estos casos la metodología de estimación apropiada es **Ordinal Regression Models.**

Otro tipo de variables categóricas son las variables no ordinales, como por ejemplo, estatus marital, estatus laboral, dependencia de los colegios, etc. La variable, tipo de colegio, que toma el valor 1 si el colegio en el que se estudió es municipal; 2 si se trató de un establecimiento particular subvencionado; y 3 si se estudió en un colegio particular pagado, es una variable categórica pero no ordinal, no tienen un orden determinado. En algunos casos cuando las variables son ordinales pero hay categorías no sabe o no responde, muchas estudios prefieren no eliminar la categoría no sabe, y tratar la variable como no ordinal. En este caso la metodología es **Multinomial Logit Model.**

V.2. Modelos de regresión ordinal (oprobit y ologit)

Supongamos que la variable dependiente tiene J categorías que pueden ser ordenadas. En este caso, el modelo de regresión lineal no puede ser utilizado porque simplemente supondría que habría igual distancia entre las categorías, lo que no necesariamente es correcto.

El análisis lo realizaremos suponiendo un modelo que incluye una variable latente (no observada) y^* que, a diferencia de nuestra variable dependiente, es continua y que se puede escribir como una función lineal de un vector de variables X:

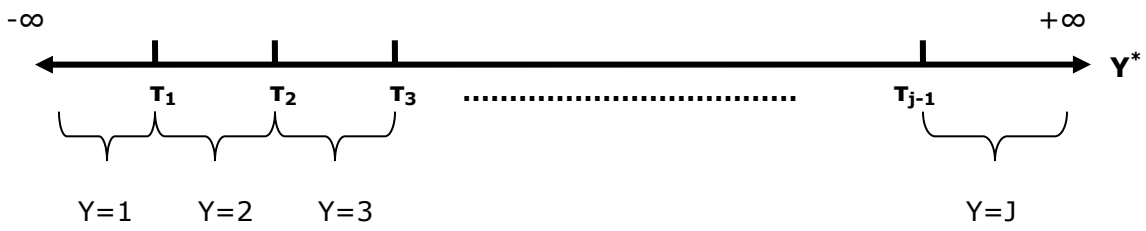
$$(1) \quad Y_i^* = \beta_1 + \beta_2 X_2 + ... + \beta_k X_k + \mu_i$$

[7] McKelvey y Zavoina (1975) y Winship y Mare (1984)

Por otra parte, la variable que observamos es una variable que está en categorías, y que está ordenada de acuerdo a la variable Y^* del modo siguiente:

$$(2) \quad Y_i = \begin{cases} 1 & si\ \tau_0 = -\infty \le Y_i^* < \tau_1 \\ 2 & si\ \tau_1 \le Y_i^* < \tau_2 \\ & \cdots\cdots\cdots\cdots \\ J & si\ \tau_{J-1} \le Y_i^* < \tau_J = +\infty \end{cases}$$

Nótese que la variable latente fluctúa entre -∞ y +∞.



Como en el caso del modelo de regresión binario, se puede realizar una estimación por medio del método de Máxima Verosimilitud para aproximarse a la regresión de Y* respecto del vector X.

Si suponemos que los errores del modelo μ tienen una distribución normal, tenemos el modelo **PROBIT ORDENADO**.

Si los errores, en cambio, suponemos que siguen una distribución logística, estamos ante el modelo **LOGIT ORDENADO**.

Siguiendo con el análisis, la probabilidad de que la variable dependiente sea igual al número asociado a la primera categoría se puede escribir como:

$$
\begin{aligned}
(3) \quad \Pr(Y_i = 1 \mid X_i) &= \Pr(\tau_0 \le Y_i^* < \tau_1 \mid X_i) \\
&= \Pr(\tau_0 \le X_i\beta + \mu_i < \tau_1 \mid X_i) \\
&= \Pr(\tau_0 - X_i\beta \le \mu_i < \tau_1 - X_i\beta \mid X_i) \\
&= F(\tau_1 - X_i\beta) - F(\tau_0 - X_i\beta)
\end{aligned}
$$

donde F es la función de distribución acumulada (normal o logística, según corresponda).

En general, la probabilidad de que Y sea igual al valor de la categoría m se puede escribir como:

$$(4) \quad \Pr(Y_i = m \mid X_i) = F(\tau_m - X_i\beta) - F(\tau_{m-1} - X_i\beta)$$

donde:

$$\tau_0 < \tau_1 < \tau_2 < \ldots < \tau_J$$

Nótese también que:

$$
\begin{aligned}
F(\tau_o - X_i\beta) &= F(-\infty - X_i\beta) = 0 \\
F(\tau_J - X_i\beta) &= F(+\infty - X_i\beta) = 1
\end{aligned}
$$

La estimación procede por máxima verosimilitud. La función de verosimilitud de este problema de estimación es la siguiente:

$$(5) \quad \ln L(\beta, \tau \mid Y, X) = \sum_{j=1}^{J} \sum_{Y_i = j} \ln[F(\tau_j - X_i\beta) - F(\tau_{j-1} - X_i\beta)]$$

El proceso de optimización implica escoger los valores de los parámetros β de manera tal que maximicen la expresión (5). Al igual que en el caso de las variables dependientes dicotómicas, no se tienen expresiones algebraicas para el resultado de esta maximización, puesto que se deben usar métodos numéricos para su solución, los que son proporcionados rápidamente por un software como STATA.

Antes de seguir, es necesario establecer que estos modelos presentan un problema de identificación. En esta estimación existen muchos parámetros libres, no se pueden estimar al mismo tiempo los J-1 umbrales y la constante. Para poder identificar este modelo se necesita asumir algo sobre la constante o alguno de los umbrales. En STATA, Ordinal Regresión Model (ORM) es identificado asumiendo que la constante es igual a cero, y se estiman los valores de todos los umbrales. Los coeficientes β no se verán afectados por la opción que se adopte.

V.3. Aplicación modelos ordinales

Consideremos la variable autopercepción de salud que reporta cada entrevistado en la Encuesta CASEN 2003. Podemos ver en la pregunta 6 del módulo de salud, que cada entrevistado clasifica su estado de salud en las siguientes categorías: 1) muy buena; 2) buena; 3) regular; 4) mala; y 5) muy mala. La tabulación de esta variable para los que responden, en la encuesta CASEN, indica que un 12.5% considera su salud como muy buena; un 51.6% de los individuos indica que es buena y un 28.4% que es regular. Por otra parte, una salud mala es indicada por 6.4% de los entrevistados y una muy mala por un 1.1% de ellos.

```
. tab s6 [aw=expr]

 percepcion |
   de salud |      Freq.     Percent        Cum.
------------+-----------------------------------
  muy buena |11,991.1229       12.50       12.50
      buena | 49,534.921       51.64       64.14
    regular | 27,211.186       28.37       92.51
       mala | 6,137.0227        6.40       98.91
   muy mala | 1,049.7472        1.09      100.00
------------+-----------------------------------
      Total |     95,924      100.00
```

Supongamos que queremos usar un modelo en que la educación y la edad son las variables determinantes de este autoreporte. Comenzaremos usando un modelo probit ordenado, para lo cual el comando en STATA es "oprobit" seguido de la variable dependiente (s6) y de las variables explicativas. Previamente, eliminamos las observaciones en que las personas indican no saber su percepción de salud, con el objeto que STATA no considere a ese grupo como otra categoría.

```
drop if s6==9
(1781 observations deleted)

. oprobit s6 esc edad [aw=expr]

(sum of wgt is   5.7985e+06)
Iteration 0:   log likelihood = -113285.38
Iteration 1:   log likelihood = -104788.18
Iteration 2:   log likelihood = -104754.42
Iteration 3:   log likelihood = -104754.41

Ordered probit regression                         Number of obs   =      95701
                                                  LR chi2(2)      =   17061.95
                                                  Prob > chi2     =     0.0000
Log likelihood = -104754.41                       Pseudo R2       =     0.0753

------------------------------------------------------------------------------
          s6 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         esc |  -.0711402   .0009141   -77.83   0.000    -.0729318   -.0693486
        edad |   .0151264   .0002283    66.24   0.000     .0146788    .0155739
-------------+----------------------------------------------------------------
       /cut1 |  -1.325288   .0165261                     -1.357678   -1.292897
       /cut2 |   .3443102   .0160589                      .3128352    .3757851
       /cut3 |    1.56539   .0168951                      1.532276    1.598504
```

```
       /cut4 |   2.507543   .0202028                      2.467946     2.54714
------------------------------------------------------------------------------
```

En primer lugar, puede observarse que no se estima el coeficiente para la constante. Esto es porque, para efectos de identificar el modelo, STATA hace automáticamente el supuesto de que la constante es 0.

En segundo lugar, el output de STATA muestra los coeficientes para las variables explicativas y, en un recuadro inferior, los valores _cut1, _cut2, _cut3 y _cut4. Estos corresponden a los que, en la nomenclatura de esta clase, serían τ_1, τ_2, τ_3 y τ_4, respectivamente.

En efecto, en este caso podemos escribir:

$$\Pr(Y_i = 1 | X_i) == \Pr(-\infty \leq X_i\beta + \mu_i \leq -1.33 | X_i)$$
$$\Pr(Y_i = 2 | X_i) == \Pr(-1.33 \leq X_i\beta + \mu_i < 0.34 | X_i)$$
$$\Pr(Y_i = 3 | X_i) == \Pr(0.34 \leq X_i\beta + \mu_i < 1.57 | X_i)$$
$$\Pr(Y_i = 4 | X_i) == \Pr(1.57 \leq X_i\beta + \mu_i < 2.51 | X_i)$$
$$\Pr(Y_i = 5 | X_i) == \Pr(2.51 \leq X_i\beta + \mu_i < +\infty | X_i)$$

Por ejemplo, considérese la probabilidad de que la salud autoreportada esté en la categoría "Muy Buena".

$$\Pr(Y_i = 1 | X_i) == \Pr(\mu_i \leq -1.325288 - (-0.71104 \cdot esc_i + 0.01513 \cdot edad_i))$$

Para poder obtener una estimación de esta probabilidad, se requiere asignar un valor para la escolaridad y otro para la variable edad. Supondremos que esta probabilidad se evalúa en las medias: 9.8 para escolaridad y 42.8 para edad, de la muestra para la cual sea ha realizado la estimación anterior:

```
. sum esc edad [aw=expr] if e(sample)

    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------------------
         esc |   95701     5798520    9.823057   4.351644          0         23
        edad |   95701     5798520    42.77247   17.31619         15        107
```

$$\Pr(Y_i = 1 | X_i) == \Pr(\mu \leq -1.2755) = 0.10106206$$

Para obtener lo anterior, se puede escribir lo siguiente:

```
. di normal(_b[/cut1]-(_b[esc]*9.8+_b[edad]*42.8))
.1010621
```

Para los otros tramos, evaluando en las medias se tiene:

```
.       di       normal(_b[/cut2]-(_b[esc]*9.8+_b[edad]*42.8))-normal(_b[/cut1]-
(_b[esc]*9.8+_b[edad]*42.8))
.55217506

.         di         norm(_b[/cut3]-(_b[esc]*9.8+_b[edad]*42.8))-norm(_b[/cut2]-
(_b[esc]*9.8+_b[edad]*42.8))
.29362425

.       di       normal(_b[/cut4]-(_b[esc]*9.8+_b[edad]*42.8))-normal(_b[/cut3]-
(_b[esc]*9.8+_b[edad]*42.8))
.04786429

. di 1-normal(_b[/cut4]-(_b[esc]*9.8+_b[edad]*42.8))
.00527429
```

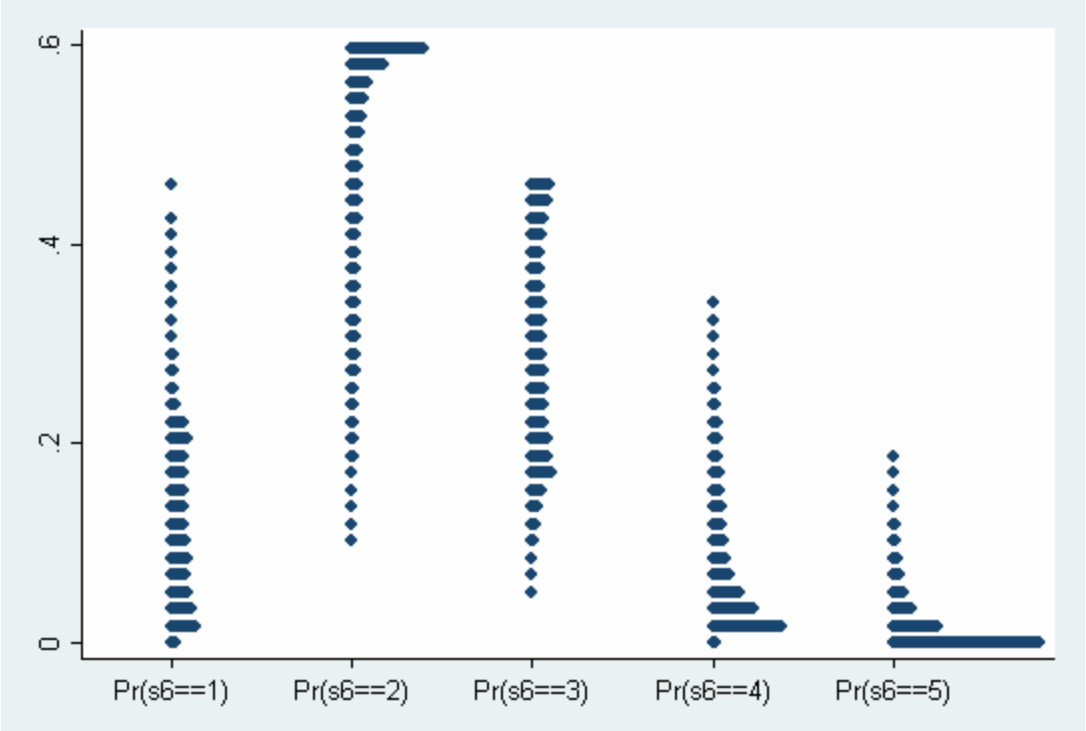
Como se aprecia, la función "normal(z)" en STATA entrega la probabilidad acumulada en una función de densidad normal estándar, es decir, la probabilidad de que una variable aleatoria normal tome un valor igual a z o menos.

Para obtener una predicción de la variable dependiente, después de la regresión se usa el comando "predict" seguido de un nombre para cada una de las variables asociadas a los 5 tramos de la variable dependiente. Es útil examinar la media, el mínimo y máximo de las probabilidades predichas:

```
. predict muybuena buena regular mala muymala
(option pr assumed; predicted probabilities)
(68169 missing values generated)

. sum  muybuena buena regular mala muymala if e(sample)

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    muybuena |     95701    .1037124    .0747025   .0022704   .4551731
       buena |     95701    .4911572     .114717   .1190668   .5961691
     regular |     95701    .3116579    .1025757   .0570016   .4584956
        mala |     95701    .0778752    .0645772   .0026346   .3190735
     muymala |     95701    .0155972    .0209159   .0000995   .1598913
```

La diferencia entre utilizar el comando **predict**, y hacerlo manualmente como en el paso anterior, es que el comando **predict** no evalúa en el promedio sino que calcula para cada individuo su probabilidad predicha, utilizando las observaciones individuales de las variables escolaridad y edad. De esta forma, la tabla anterior nos muestra el promedio de las probabilidades predichas para cada individuo en cada una de las categorías de respuesta del estado de salud.

El siguiente gráfico nos muestra la distribución de las probabilidades estimadas para cada uno de los estados de salud:



El cual se obtuvo utilizando el siguiente comando:

```
dotplot muybuena buena regular mala muymala if e(sample)
```

Los efectos marginales de esta estimación se deben computar en forma separada para cada una de las categorías de la variable dependiente. El comando que se utiliza para esto es **`mfx`**, comando que computa los efectos marginales de cualquier estimación. En particular, para los modelos de variable dependiente categórica, los efectos marginales de deben computar en forma separada para cada una de las categorías. Esto se hace de la siguiente forma:

```
. mfx, predict(p outcome(1))

Marginal effects after oprobit
      y  = Pr(s6==1) (predict, p outcome(1))
         =  .10142631
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
     esc |   .0126146      .00017   72.39   0.000   .012273  .012956   9.82306
    edad |  -.0026822      .00004  -62.58   0.000  -.002766 -.002598   42.7725
------------------------------------------------------------------------------
```

Primero, este comando lo que hace es obtener la probabilidad predicha para la categoría 1 (Muy buena), y segundo obtiene los efectos marginales de esta categoría con respecto a las variables explicativas del modelo. De esta forma, se puede concluir que un año adicional de escolaridad aumenta la probabilidad de que el individuo reporte salud muy buena en 1.26%, y un año adicional de edad disminuye la probabilidad de que la persona reporte un estado de salud muy buena en 0.27%.

A continuación se presentan los efectos marginales estimados para todas las demás categorías:

```
. mfx, predict(p outcome(2))

Marginal effects after oprobit
      y  = Pr(s6==2) (predict, p outcome(2))
         =  .55256973
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
     esc |   .0136246      .00021   63.90   0.000   .013207  .014042   9.82306
    edad |   -.002897      .00005  -57.00   0.000  -.002997 -.002797   42.7725
```

```
. mfx, predict(p outcome(3))

Marginal effects after oprobit
      y  = Pr(s6==3) (predict, p outcome(3))
         =  .29308764
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
     esc |  -.0185637      .00026  -71.38   0.000  -.019073 -.018054   9.82306
    edad |   .0039472      .00006   61.73   0.000   .003822  .004072   42.7725
------------------------------------------------------------------------------
```

```
. mfx, predict(p outcome(4))

Marginal effects after oprobit
      y  = Pr(s6==4) (predict, p outcome(4))
         =  .04767313
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
     esc |  -.0066024      .00011  -61.16   0.000  -.006814 -.006391   9.82306
    edad |   .0014038      .00003   55.18   0.000   .001354  .001454   42.7725
------------------------------------------------------------------------------

. mfx, predict(p outcome(5))

Marginal effects after oprobit
      y  = Pr(s6==5) (predict, p outcome(5))
         =  .00524319
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
     esc |  -.0010731      .00004  -29.53   0.000  -.001144 -.001002   9.82306
    edad |   .0002282      .00001   29.07   0.000   .000213  .000244   42.7725
------------------------------------------------------------------------------
```

Todo lo realizado hasta ahora se puede hacer en forma equivalente asumiendo una función de distribución logística por el método de logit ordenado.

Se puede también apreciar que en el output de STATA se proveen los errores estándar y test t y p-values para la hipótesis nula de que los coeficientes de las variables explicativas, a nivel individual, son iguales a cero, los que se interpretan de la manera habitual.

V.4. Multinomial Logit

La metodología *Multinomial Logit (MNLM)* es aplicada cuando la variable dependiente es categórica, pero estas categorías no tienen un orden. Esta metodología puede ser pensada como una estimación simultánea de modelos logit binarios para todas las comparaciones posibles de las categorías de resultados. En este sentido, MNLM es una simple extensión del modelo logit binario, sin embargo, la dificultad viene del echo de que son muchas categorías que comparar.

Por ejemplo, si la variable categórica tiene tres categorías, MNLM equivalente a hacer tres modelos logia binarios que compare las categorías 1 y 2, 1 y 3, y 2 y 3. Si se agrega una categoría más, sin embargo, se deben incluir tres comparaciones más: 1 y 4, 2 y 4, y 3 y 4.

Considere una variable categórica no ordinal con tres categorías: A, B, y C, las que tienen N_A, N_B, y N_C cantidad de observaciones cada una. Para analizar la relación entre la variable dependiente categórica (y) y la variable explicativa (x) se debe correr una serie de regresiones binarias logia. Para analizar los efectos de la variable explicativa x en las chances (odds) de A versus B, se deben seleccionar las N_A+N_B observaciones y estimar el siguiente modelo logit binario:

$$\ln\left[\frac{\Pr(A \mid X_i)}{\Pr(B \mid X_i)}\right] = \alpha_{A/B} + \beta_{A/B} X_i$$

La variable dependiente de este modelo es el logaritmo de las chances de A versus B. Un incremento de una unidad de X aumenta las chances de A versus B en $e^{\beta_{A/B}}$.

Las otras comparaciones pueden ser obtenidas de la misma forma:

$$\ln\left[\frac{\Pr(B \mid X_i)}{\Pr(C \mid X_i)}\right] = \alpha_{B/C} + \beta_{B/C} X_i$$

$$\ln\left[\frac{\Pr(A \mid X_i)}{\Pr(C \mid X_i)}\right] = \alpha_{A/C} + \beta_{A/C} X_i$$

Sin embargo, una de estas es irrelevante, si sabemos como X afecta las chantes de B versus C, y sabemos como X afecta las chances de A versus C, es razonable esperar saber el impacto de X sobre las chances de A versus C.

Para efectos de estimar el modelo, finalmente se toma una de las categorías como base, y se estiman todas las demás en función de esta. En términos de probabilidades el modelo puede ser reescrito de la siguiente forma:

$$\Pr(y = m \mid X_i) = \frac{\exp(\beta_{m|b} X_i)}{\sum_{j=1}^{J} \exp(\beta_{j|b} X_i)}$$

V.5. Aplicación Multinomial Logit

Ahora analizaremos los determinantes del estatus laboral reportado en la encuesta CASEN 2003. A partir de la variable o9 se genera una nueva variable de estatus laboral:

```
g estatus=1 if o9==1
replace estatus=2 if o9==2
replace estatus=3 if o9==3 | o9==4
replace estatus=4 if o9==5

label define o9lbl 1 "Empleador" 2 "Cuenta Propia" 3 "Empleador S.
Publico" 4 "E. Sector Privado"
label values estatus o9lbl

g genero=1 if sexo==1
replace genero=0 if sexo==2

. tab estatus [aw=expr]

             estatus |      Freq.     Percent        Cum.
---------------------+-----------------------------------
           Empleador | 3,670.7364        4.33        4.33
       Cuenta Propia | 18,773.272       22.14       26.47
Empleador S. Publico | 8,701.0031       10.26       36.73
   E. Sector Privado | 53,650.988       63.27      100.00
---------------------+-----------------------------------
               Total |     84,796      100.00
```

```
. mlogit estatus edad esc genero [aw=expr]

(sum of wgt is   5.4402e+06)
Iteration 0:   log likelihood = -83863.306
Iteration 1:   log likelihood = -75998.235
Iteration 2:   log likelihood = -75504.147
Iteration 3:   log likelihood = -75499.609
Iteration 4:   log likelihood = -75499.608

Multinomial logistic regression                 Number of obs   =      84442
                                                LR chi2(9)      =   16727.40
                                                Prob > chi2     =     0.0000
Log likelihood = -75499.608                     Pseudo R2       =     0.0997

------------------------------------------------------------------------------
     estatus |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Empleador    |
        edad |   .0799784   .0014199    56.33   0.000     .0771954    .0827614
         esc |   .1597634   .0045325    35.25   0.000     .1508798     .168647
      genero |   .0993261   .0392921     2.53   0.011      .022315    .1763371
       _cons |   -8.00148    .100142   -79.90   0.000    -8.197754   -7.805205
-------------+----------------------------------------------------------------
Cuenta Pro~a |
        edad |   .0438387   .0007387    59.35   0.000     .0423908    .0452865
         esc |  -.0708369    .002294   -30.88   0.000    -.0753331   -.0663406
      genero |  -.2440134   .0193707   -12.60   0.000    -.2819792   -.2060476
       _cons |  -1.911952   .0468815   -40.78   0.000    -2.003838   -1.820066
-------------+----------------------------------------------------------------
Empleador ~o |
        edad |   .0516767   .0010375    49.81   0.000     .0496432    .0537101
         esc |   .2016091   .0035161    57.34   0.000     .1947177    .2085006
      genero |  -.8844475   .0247347   -35.76   0.000    -.9329267   -.8359682
       _cons |  -5.845748   .0689473   -84.79   0.000    -5.980882   -5.710614
------------------------------------------------------------------------------
(estatus==E. Sector Privado is the base outcome)
```

La estimación anterior muestra tres estimaciones de logit binario, estimando las chances de cada alternativa versus la categoría base, que en este caso STATA la ha seleccionado automáticamente como Sector Pivado. Los parámetros estimados en cada regresión representan los betas para cada modelo, y por si sólo no tienen interpretación. Para mayor análisis debemos ver los efectos marginales. Pero antes, la siguiente tabla muestra el mismo modelo anterior pero tomando como base a los empleadores:

```
. mlogit estatus edad esc genero, baseoutcome(1) nolog

Multinomial logistic regression                 Number of obs   =      84512
                                                LR chi2(9)      =   18932.79
                                                Prob > chi2     =     0.0000
Log likelihood = -75390.042                     Pseudo R2       =     0.1116

------------------------------------------------------------------------------
     estatus |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Cuenta Pro~a |
        edad |  -.0427713   .0016151   -26.48   0.000    -.0459367   -.0396058
         esc |   -.236924   .0049962   -47.42   0.000    -.2467164   -.2271316
      genero |   -.294665   .0476776    -6.18   0.000    -.3881113   -.2012187
       _cons |   6.609551   .1147018    57.62   0.000     6.384739    6.834362
-------------+----------------------------------------------------------------
Empleador ~o |
        edad |  -.0359512   .0017585   -20.44   0.000    -.0393978   -.0325046
         esc |    .032728    .005432     6.03   0.000     .0220816    .0433745
      genero |  -1.100984   .0497899   -22.11   0.000    -1.198571   -1.003398
       _cons |   2.979489   .1218546    24.45   0.000     2.740659     3.21832
-------------+----------------------------------------------------------------
E. Sector ~o |
        edad |  -.0886659   .0016018   -55.35   0.000    -.0918054   -.0855265
         esc |  -.1798783   .0048512   -37.08   0.000    -.1893865   -.1703702
      genero |  -.0412019   .0465784    -0.88   0.376     -.132494    .0500901
       _cons |   8.557451   .1126263    75.98   0.000     8.336707    8.778194
------------------------------------------------------------------------------
(estatus==Empleador is the base outcome)
```

A continuación se presenta la estimación de los efectos marginales de cada una de las variables explicativas sobre la probabilidad de pertenecer a cada una de las categories ocupaciones.

```
. mfx, predict(p outcome(1))

Marginal effects after mlogit
      y  = Pr(estatus==1) (predict, p outcome(1))
         =  .02391565
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
    edad |   .0016826      .00004   45.73   0.000    .00161  .001755   39.8841
     esc |    .004217      .00011   37.22   0.000   .003995  .004439    9.5322
  genero*|   .0046209      .00098    4.72   0.000   .002704  .006538   .716289
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

La tabla anterior nos muestral que un año adicional de edad aumenta la probabilidad de ser empleador en 0.17%, un año adicional de escolaridad aumenta la probabilidad de ser empleador en 0.42%, y pasar de ser mujer a ser hombre aumenta la probabilidad de ser empleador en 0.46%.

```
. mfx, predict(p outcome(2))

Marginal effects after mlogit
      y  = Pr(estatus==2) (predict, p outcome(2))
         =  .27279488
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
    edad |   .0075244      .00013   56.62   0.000   .007264  .007785   39.8841
     esc |  -.0165297      .00042  -39.12   0.000  -.017358 -.015702    9.5322
  genero*|  -.0252176      .00375   -6.72   0.000  -.032576 -.017859   .716289
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1

. mfx, predict(p outcome(3))

Marginal effects after mlogit
      y  = Pr(estatus==3) (predict, p outcome(3))
         =  .06965042
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
    edad |   .0023962      .00007   35.03   0.000   .002262   .00253   39.8841
     esc |    .014561       .0002   71.29   0.000   .014161  .014961    9.5322
  genero*|  -.0778299      .00233  -33.43   0.000  -.082393 -.073267   .716289
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1

. mfx, predict(p outcome(4))

Marginal effects after mlogit
      y  = Pr(estatus==4) (predict, p outcome(4))
         =  .63363904
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
    edad |  -.0116032      .00015  -78.33   0.000  -.011893 -.011313   39.8841
     esc |  -.0022483      .00045   -4.99   0.000  -.003131 -.001365    9.5322
  genero*|   .0984267      .00402   24.46   0.000   .090539  .106315   .716289
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Capítulo VI. Variable Dependiente Limitada: Censura, Truncamiento, y Sesgo de Selección

VI.1. Introducción

En el modelo de regresión lineal, los valores de todas las variables son conocidos para la muestra completa. Dentro del trabajo empírico, existen muchos casos donde debido a la forma de recolección de los datos, disponemos de información incompleta de las variables de interés. En esta clase se considerarán los casos en que la variable dependiente es limitada, es decir, esta censurada o truncada.

Censura: esto sucede cuando la variable dependiente, pero no así las variables explicativas, son observadas dentro de un rango restringido En esta situación todas las observaciones de la variable dependiente que están en o bajo (sobre) cierto umbral son tratadas como si estuvieran **en** el umbral. Por ejemplo, cuando para ingresos superiores a 1.000.000 todas las observaciones tienen 1.000.000. Nosotros sabemos que deberían haber ingresos superiores a 1,000,000 pero sólo observamos valores hasta 1.000.000.

Truncamiento: en este caso los datos son limitados en una forma más severa, ya que las observaciones con ingresos superiores al 1.000.000 son eliminadas de la muestra. En este caso, tanto los datos de la variable dependiente como de las variables explicativas están perdidos.

En ambos casos la variable de interés es observada en forma incompleta o limitada. La estimación de estos modelos por MCO entrega un estimador inconsistente debido a que la muestra no es representativa de la población. De esta forma, se requiere de una metodología apropiada de estimación donde se realicen los supuestos

pertinentes sobre la distribución, para lograr una estimación consistente de los parámetros.

Sesgo de selección o truncamiento incidental: en este caso la muestra esta truncada, sólo tenemos observaciones de las variable dependiente y explicativas para un subconjunto de la muestra, pero la forma en que esta muestra fue truncada no es aleatoria sino que esta relacionada con la variable dependiente o la variable de interés. Por ejemplo, cuando estimamos un modelo de salario, sólo observamos a las mujeres que participan en el mercado del trabajo, la decisión de participar o no en el mercado del trabajo no es aleatoria, sino que depende entre otras cosas de variables que determinan el ingreso, que es justamente la variable de interés. Este problema fue planteado por primera vez por Heckman (1979), y su metodología es la más apropiada para este tipo de problema.

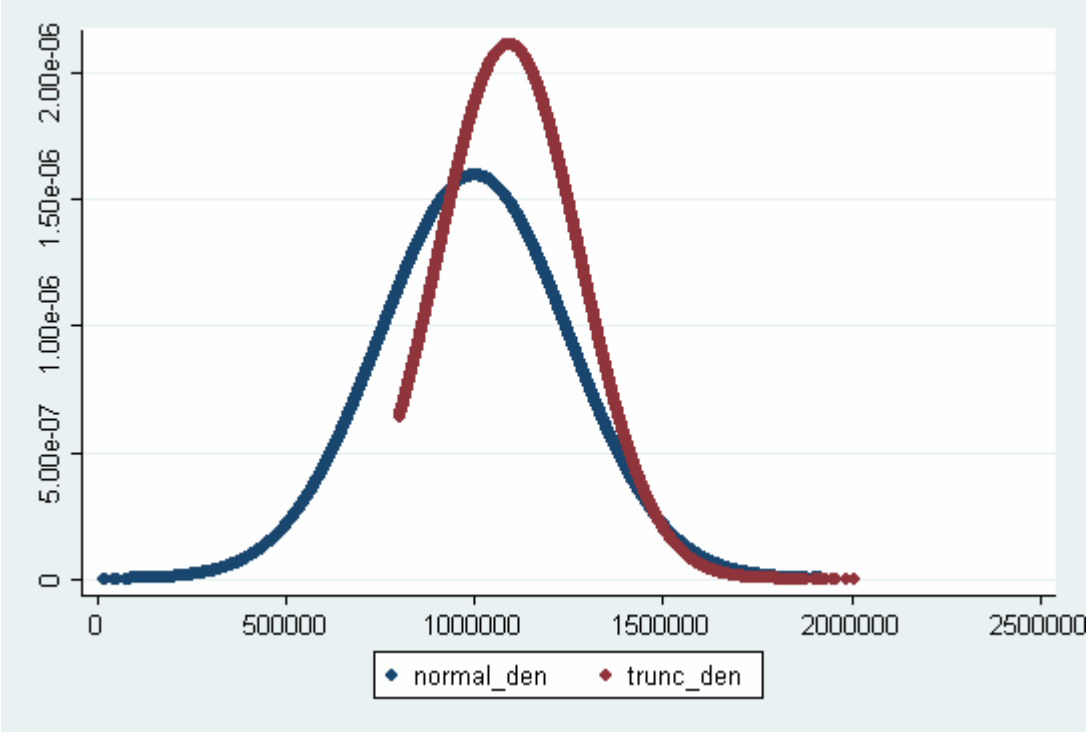
VI.2. Datos Truncados

VI.2.1. Estimador para el problema de Truncamiento

Cuando tenemos datos truncados se tiene acceso sólo a una parte de la muestra. Para las personas que son excluidas no tenemos observaciones ni de la variable dependiente ni de las variables explicativas del modelo. Por ejemplo, si tenemos una encuesta que sólo entrevista a los egresados de Ingeniería Comercial, y queremos estudiar el retorno a la educación. Esta es una muestra truncada, relativa a la población, ya que han sido excluidos todos los individuos que no han egresado de la carrera de Ingeniería Comercial. Los individuos excluidos probablemente no tengan las mismas características de los individuos que han sido incluidos. En efecto, deberíamos esperar que el ingreso promedio de las personas que están quedando fuera sea menor que los ingresos de los graduados de ingeniería comercial.

Los efectos de truncar la distribución es clara, el promedio de la variable se de interés se mueve hacia la derecha desde el punto de truncamiento, y la desviación estándar de la variable se reduce.

El siguiente gráfico muestra la diferencia entre una distribución de probabilidad de la muestra completa y la muestra truncada, y a continuación la media y desviación estándar de los datos truncados y no truncados.



Variable	Obs	Mean	Std. Dev.	Min	Max
normal	100000	1000029	250156.1	0	2070634
trunc	78789	1091829	189679.9	800012.8	2070634

La función de densidad truncada toma la siguiente forma:

$$(1) \quad f(y \mid y > \tau) = \frac{f(y)}{\Pr(y > \tau)}$$

Esta forma de la función de densidad asegura que la función de distribución acumulada sume uno sobre el rango de observaciones truncadas.

Si la variable ***y*** se distribuye normal la expresión en (1) se puede reescribir de la siguiente forma:

$$(2) \quad \phi(z \mid z > \alpha) = \frac{\phi(z)}{[1 - \Phi(\alpha)]}$$

$$donde:$$

$$z = \frac{y - E(y)}{V(y)}$$

$$\alpha = \frac{\tau - E(y)}{V(y)}$$

La muestra truncada no puede ser utilizada para hacer inferencia sobre la población completa sin corregir por aquellos individuos que fueron excluidos en forma no aleatoria de la población. Pero tampoco podemos utilizar la estimación obtenidas de esta sub-muestra para hacer inferencia sobre el sub-grupo para el cual tenemos observaciones. La regresión MCO cuando estamos estimando en una sub muestra será sesgado hacia cero, y se subestimará la varianza del error (σ_u^2).

Cuando estamos tratando con una distribución normal truncada, donde $Y_i = X_i + \mu_i$ es observada sólo para valores de Y_i superiores a cierto umbral τ, se puede definir:

$$\alpha_i = \frac{\tau - X_i\beta}{\sigma_u}$$

$$\lambda(\alpha_i) = \frac{\phi(\alpha_i)}{[1-\Phi(\alpha_i)]} \quad cuando \quad Y_i > \tau$$

$$\lambda(\alpha_i) = \frac{-\phi(\alpha_i)}{\Phi(\alpha_i)} \quad cuando \quad Y_i < \tau$$

La expresión $\lambda(\alpha_i)$ se denomina *inverse Mills ratio* (IMR), el cual representa la probabilidad de observar α_i condicional a que α_i esta en la muestra truncada. Con manipulación de las variables y de la distribución normal, se obtiene que:

$$E\left[Y_i \mid Y_i > \tau, X_i\right] = X_i\beta + \sigma_u\lambda(\alpha_i) + \mu_i$$

De la ecuación anterior podemos deducir que la simple regresión MCO de $\boldsymbol{Y_i}$ sobre $\boldsymbol{X_i}$ sufre de la exclusión del término $\lambda(\alpha_i)$. Para solucionar este problema, se debe incluir dentro del modelo de regresión la "variable explicativa" equivalente al IMR, corrigiendo el sesgo en la estimación cuando la muestra observada esta truncada.

VI.2.2. Aplicación: Estimación de los determinantes de las horas trabajadas de las mujeres

El objetivo de esta aplicación es estudiar los determinantes de las horas trabajadas de las mujeres casadas utilizando como variables explicativas el número de hijos entre 0 y 2 años, entre 2 y 6 años, y entre 6 y 18 años, la edad y los años de escolaridad. El problema es que sólo observamos a las mujeres que trabajan, por lo cual las horas trabajadas están truncadas hasta el cero. Para esta estimación utilizaremos la información de la encuesta CASEN 2003. En realidad en esta encuesta tenemos observaciones para todas las mujeres las que trabajan y no trabajan, pero para efectos de entender la estimación con datos truncados borraremos las observaciones de las mujeres que no trabajan, es decir, nos quedaremos con una sub-muestra de mujeres, aquellas que trabajan.

A continuación se presentan las estadísticas descriptivas de las variables de interés:

```
. sum o19_hrs t_hijos0_2 t_hijos2_6 t_hijos6_18 edad esc if o19_hrs>0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     o19_hrs |     11219    167.5696    71.79072          1        480
  t_hijos0_2 |     11219    .1087441    .3286048          0          2
  t_hijos2_6 |     11219    .2264016    .4719763          0          3
 t_hijos6_18 |     11219    .9598003    1.003509          0          6
       edad  |     11219    42.04065     10.6971         16         85
-------------+--------------------------------------------------------
         esc |     11190    10.52458    4.307975          0         22
```

Para ilustrar las consecuencias de estimar el modelo ignorarnos el problema de truncamiento, primero estimaremos el siguiente modelo por MCO:

$$hrs_i = \alpha + \beta_1 \cdot hijos0\_2_i + \beta_2 \cdot hijos2\_6_i + \beta_3 \cdot hijos6\_18_i + \beta_4 \cdot edad_i + \beta_5 \cdot esc_i + \mu_i$$

Los resultados de esta estimación son los siguientes:

```
. reg o19_hrs t_hijos0_2 t_hijos2_6 t_hijos6_18 edad esc if o19_hrs>0

      Source |       SS       df       MS              Number of obs =   11190
-------------+------------------------------           F(  5, 11184) =   14.80
       Model |  378918.465     5   75783.693           Prob > F      =  0.0000
    Residual |  57279256.7 11184  5121.53583           R-squared     =  0.0066
-------------+------------------------------           Adj R-squared =  0.0061
       Total |  57658175.2 11189  5153.11245           Root MSE      =  71.565

------------------------------------------------------------------------------
     o19_hrs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  t_hijos0_2 |  -.4547121   2.208499    -0.21   0.837    -4.783758    3.874334
  t_hijos2_6 |  -4.296759   1.567057    -2.74   0.006    -7.368466   -1.225052
 t_hijos6_18 |  -4.453983   .7020404    -6.34   0.000    -5.830106    -3.07786
       edad  |   .1955919   .0767236     2.55   0.011     .0452002    .3459836
         esc |   .2089916   .1644957     1.27   0.204    -.1134489    .5314321
       _cons |   162.3907   4.420836    36.73   0.000      153.725    171.0563
------------------------------------------------------------------------------
```

Del modelo anterior podemos concluir que tener hijos entre 2 y 18 años disminuyen las horas trabajadas de las mujeres casadas. La edad aumenta las horas trabajadas,

la escolaridad no resulta ser estadísticamente significativa. Recordemos que cuando tenemos un problema de truncamiento en la distribución de la variable de interés, los coeficientes estimados son sesgados hacia cero, es decir, en valor absoluto son menores a los que se debería obtener. Además, la varianza estimada del error es sesgada hacia abajo, es menor a la que debería, por lo cual los estadísticos t están sobre estimados y se esta rechazando la hipótesis nula de que los coeficientes son iguales a cero más de lo que se debiera, es decir, con esta estimación podemos concluir que parámetros son estadísticamente significativos cuando realmente no lo son.

Ahora, si podemos justificar que los errores del modelo se distribuyen en forma normal, es posible estimar este modelo truncado en STATA utilizando el comando **`truncreg`**. Bajo el supuesto de normalidad, podemos hacer inferencia para toda la población asociada a partir del modelo truncado estimado. En este caso en particular, podemos hacer inferencia sobre todas las mujeres y no sólo las que trabajan. El comando tiene la opción **`ll(#)`** que indica que todas las observaciones con valores de la variable dependiente igual o menor a # han sido truncadas, la opción **`lu(#)`** indica que las observaciones con valores de la variable dependiente iguales o mayores a # han sido truncadas. La distribución puede estar truncada por arriba y abajo, ambas opciones pueden ser utilizadas en forma simultánea.

A continuación se presentan los resultados del modelo estimado anteriormente, pero utilizando el comando **truncreg**:

```
. truncreg o19_hrs t_hijos0_2 t_hijos2_6 t_hijos6_18 edad esc, ll(0)
(note: 30287 obs. truncated)

Fitting full model:

Iteration 0:   log likelihood = -63552.484
Iteration 1:   log likelihood = -63533.426
Iteration 2:   log likelihood = -63533.416
Iteration 3:   log likelihood = -63533.416

Truncated regression
Limit:   lower =          0                     Number of obs =  11190
         upper =       +inf                     Wald chi2(5)  =  73.86
Log likelihood = -63533.416                     Prob > chi2   = 0.0000

------------------------------------------------------------------------------
     o19_hrs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
eq1          |
  t_hijos0_2 |  -.5077218   2.403286    -0.21   0.833    -5.218076    4.202633
  t_hijos2_6 |  -4.708079    1.71185    -2.75   0.006    -8.063243   -1.352916
 t_hijos6_18 |  -4.863076   .7671873    -6.34   0.000    -6.366735   -3.359416
        edad |   .2100532   .0831678     2.53   0.012     .0470473    .3730591
         esc |   .2277216   .1785026     1.28   0.202    -.1221372    .5775804
       _cons |   159.4184   4.803027    33.19   0.000     150.0046    168.8321
-------------+----------------------------------------------------------------
sigma        |
       _cons |   74.57292   .5779114   129.04   0.000     73.44023     75.7056
------------------------------------------------------------------------------
```

La siguiente tabla muestra la comparación de ambas estimaciones:

```
. estimates table mco truncreg, stat(rmse) b(%7.3g) p(%4.3f)

----------------------------------
    Variable |   mco     trunc~g
-------------+--------------------
             |
_
  t_hijos0_2 |   -.455
             |   0.837
  t_hijos2_6 |    -4.3
             |   0.006
 t_hijos6_18 |   -4.45
             |   0.000
        edad |    .196
             |   0.011
         esc |    .209
             |   0.204
       _cons |     162
             |   0.000
-------------+--------------------
eq1          |
  t_hijos0_2 |             -.508
             |             0.833
  t_hijos2_6 |             -4.71
             |             0.006
 t_hijos6_18 |             -4.86
             |             0.000
        edad |               .21
             |             0.012
         esc |              .228
             |             0.202
       _cons |              159
             |             0.000
-------------+--------------------
sigma        |
       _cons |              74.6
             |             0.000
-------------+--------------------
Statistics   |
        rmse |    71.6
----------------------------------
                       legend: b/p
```

Podemos observar que los coeficientes de las variables explicativas efectivamente estaban sesgados hacia cero cuando se utilizó la metodología de MCO. El coeficiente `sigma _cons` representa la estimación de la desviación estándar y es comparable a

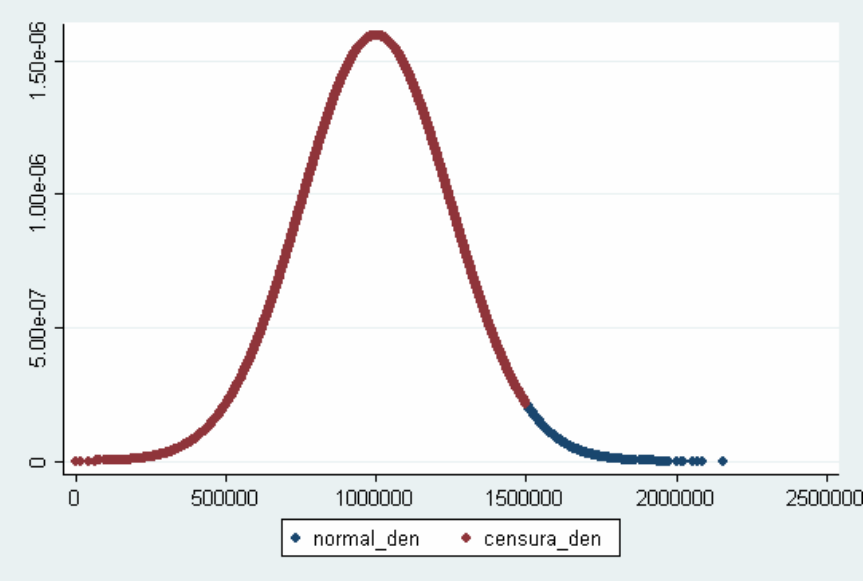
rmse de la estimación MCO, nuevamente podemos apreciar que la desviación estándar de la estimación MCO esta siendo subestimada.
Los resultados de la estimación del modelo truncado, considerando este efecto, pueden ser utilizados para hacer inferencias sobre la población completa.

VI.3. Datos Censurados

VI.3.1. Estimador Tobit para el problema de Censura

La Censura ocurre cuando la variable dependiente es asumida igual a cierto valor cuando la variable esta por debajo (sobre) el punto de censura. Cuando los datos están censurados tenemos acceso al total de observaciones de la muestra, sólo que para cierto grupo de observaciones, aquellas debajo (sobre) del punto de censura, no observamos la variable dependiente sino sólo las variables explicativas. Este problema es típico en las variables de ingresos, donde existe un tope y sobre este todos los valores son codificados con igual valor (máximo).
En presencia de censura el sesgo introducido a los parámetros es más grave mientras mayor es la probabilidad de que las observaciones de la variable dependiente caigan en la zona censurada.

El siguiente gráfico muestra la función de densidad cuando la variable está censurada:



Una solución al problema de censura, cuando el punto de censura es el cero, fue propuesto por primera vez por Tobin (1958), y se denomina ***Cesored Regresión Model***, por lo cual se conoce como modelo Tobit ("Tobin's probit").

El modelo puede ser expresado en función de una variable latente de la siguiente forma:

$$y_i^* = X_i\beta + \mu_i$$

$$y_i = \begin{cases} 0 & si \quad y_i^* \leq 0 \\ y_i^* & si \quad y_i^* > 0 \end{cases}$$

Este modelo combina dos tipos de modelos, un modelo probit binario para distinguir entre y_i=0 e y_i>0, y el modelo de regresión para los valores de y_i mayores a cero. El comando en STATA que realiza la estimación de estos modelos es **`tobit`**, los puntos

de censura son definidos con las opciones **ll()** y **ul()**, el primero indica la censura por la izquierda, el segundo la censura por la derecha o arriba de la distribución, y se pueden especificar ambos al mismo tiempo.

Por supuesto, se podrían agrupar todas las observaciones donde la variable dependiente es mayor a cero en un solo código (igual a uno), y el modelo se transforma en un modelo binario (probit o logit), pero haciendo esto estamos descartando toda la información importante en la variable dependiente. También podríamos olvidarnos de todas las observaciones donde la variable dependiente es igual a cero, pero en este caso tendríamos un problema de truncamiento con las consecuencias que esto genera sobre la estimación, y además al sacar los datos se esta perdiendo información relevante.

VI.3.2. Aplicación: Estimación de los salarios para las mujeres

Para esta aplicación utilizaremos la encuesta CASEN 2003, y estimaremos los determinantes del salario por hora, para lo cual se utilizará como variable dependiente el logaritmo del salario por hora, y como variables explicativas la escolaridad, la edad, una dummy si la mujer esta casada, y las tres variables de número de hijos utilizadas en la sección II.1. El salario por hora de las mujeres que no trabajan ha sido reemplazado por cero.

El modelo a estimar es el siguiente:

$$lyph_i = \alpha + \beta_1 \cdot esc_i + \beta_2 \cdot edad_i + \beta_3 \cdot casada_i + \beta_4 \cdot hijos0\_2_i + \beta_5 \cdot hijos2\_6_i + \beta_6 \cdot hijos6\_18_i + \mu_i$$

Tenemos un problema de censura, ya que para las mujeres que no trabajan el salario por hora es igual a cero.

Para poder comparar, primero estimaremos el modelo por MCO. Estos son los resultados:

```
. reg lyph esc edad casada t_hijos0_2 t_hijos2_6 t_hijos6_18

      Source |       SS       df       MS              Number of obs =   73947
-------------+------------------------------           F(  6, 73940) = 2102.64
       Model |  105449.586     6  17574.9309           Prob > F      =  0.0000
    Residual |  618028.471 73940  8.35851327           R-squared     =  0.1458
-------------+------------------------------           Adj R-squared =  0.1457
       Total |  723478.057 73946  9.78387008           Root MSE      =  2.8911

------------------------------------------------------------------------------
        lyph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         esc |   .2410843   .0027767    86.82   0.000     .2356419    .2465266
        edad |  -.0161103   .0008979   -17.94   0.000    -.0178703   -.0143503
      casada |   -.738852   .0218491   -33.82   0.000    -.7816762   -.6960279
  t_hijos0_2 |  -1.210209   .0320357   -37.78   0.000    -1.272999   -1.147419
  t_hijos2_6 |  -.5096824   .0250836   -20.32   0.000    -.5588461   -.4605187
 t_hijos6_18 |   .0711068   .0119349     5.96   0.000     .0477143    .0944992
       _cons |   1.416301   .0609413    23.24   0.000     1.296856    1.535745
------------------------------------------------------------------------------
```

De este modelo podemos concluir que el retorno a la educación es de un 24%, que la edad influye negativamente sobre el salario por hora, pasar de no estar casada a estar casada disminuye el salario por hora en un 73%, y que tener hijos menores de 6 años disminuye el salario por hora. Sin embargo, este modelo no esta estimado en forma correcta, ya que no considera la correcta distribución de la variable dependiente.

Utilizando la metodología Tobit, que estima la probabilidad de que el salario por hora sea igual a cero versus mayor a cero, y para las personas con salarios mayores a cero estiman los parámetros de interés, se obtienen los siguientes resultados:

```
. tobit lyph esc edad casada t_hijos0_2 t_hijos2_6 t_hijos6_18, ll(0)

Tobit regression                                  Number of obs   =      73947
                                                  LR chi2(6)      =   10900.74
                                                  Prob > chi2     =     0.0000
Log likelihood = -100018.91                       Pseudo R2       =     0.0517
```

```
------------------------------------------------------------------------------
        lyph |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         esc |   .7039213   .0094809    74.25   0.000     .6853389    .7225037
        edad |  -.0774246   .0030792   -25.14   0.000    -.0834597   -.0713894
      casada |  -2.414922   .0724712   -33.32   0.000    -2.556965   -2.272879
  t_hijos0_2 |  -3.982303   .1096702   -36.31   0.000    -4.197256    -3.76735
  t_hijos2_6 |  -1.644371   .0820848   -20.03   0.000    -1.805257   -1.483485
 t_hijos6_18 |   .2869502   .0382811     7.50   0.000     .2119194     .361981
       _cons |  -4.012681   .2010891   -19.95   0.000    -4.406814   -3.618547
-------------+----------------------------------------------------------------
      /sigma |   7.546969   .0430214                      7.462647    7.631291
------------------------------------------------------------------------------
  Obs. summary:      51802  left-censored observations at lyph<=0
                     22145     uncensored observations
                         0 right-censored observations
```

Pero ojo con la interpretación de los coeficientes, al igual que en las estimaciones de modelos probit, la interpretación de los coeficientes se debe hacer con los efectos marginales.

Para ver como se obtienen los efectos marginales, primero debemos notar que este modelo podemos obtener tres tipos de predicciones:

- Predicción de la probabilidad de que el logaritmo del salario por hora sea mayor a cero `Pr(lyh>0)`:

```
predict prob, pr(0,.)
.  sum prob

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        prob |     73947    .3271485    .1649213   .0068406   .8833697
```

- Predicción del valor esperado de la variable dependiente, pero en el intervalo censurado ($E[X_i\hat{\beta} + \mu_i \mid X_i\hat{\beta} + \mu_i > 0]$):

```
predict esp_cen, e(0,.)
. sum esp_cen

    Variable |       Obs        Mean    Std. Dev.       Min        Max
```

```
-------------+--------------------------------------------------------
     esp_cen |     73947    4.987976    1.067534   2.459088   10.67095
```

- Por último, se puede predecir la variable latente y_i^*, en la zona de censura, es decir, hace la predicción censurada, donde le umbral es considerado en la distribución de la variable dependiente:

```
predict esp_ncen, ystar(0,.)

. sum esp_ncen

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    esp_ncen |     73947    1.807126    1.275492   .0168217   9.426397
```

Los coeficientes estimados en el output anterior representan los efectos marginales de las variables explicativas sobre la variable latente:

$$(3) \quad \frac{\partial E[lyph^* \mid X]}{\partial X_j} = \beta_j$$

Pero esta información no es la que nos interesa en realidad, el efecto marginal sobre la variable observada (censurada) es el siguiente:

$$(4) \quad \frac{\partial E[lyph \mid X]}{\partial X_j} = \beta_j \cdot \Pr(lyph > 0)$$

Así, como la probabilidad es como máximo 1, al considerar la variable censurada (y no la variable latente), los efectos marginales se acercan más a cero. Un incremento en una variable explicativa que tiene un coeficiente positivo tiene como efecto para los individuos censurados que para ellos sea menos probable estar en la zona censurada, y para los individuos no censurados, un aumento marginal en la variable explicativa aumenta `E[lyph|lyph>0]`. Los efectos marginales capturan la combinación de ambos efectos.

En STATA los efectos marginales se estiman mediante el comando **mfx**, y se debe especificar que se quiere predecir, la variable observada, la variable latente o la probabilidad.

```
. mfx, predict(e(0,.))

Marginal effects after tobit
      y  = E(lyph|lyph>0) (predict, e(0,.))
         =  4.8324679
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
     esc |   .1886656      .00244   77.31   0.000   .183883  .193449   8.21227
    edad |  -.0207514      .00082  -25.34   0.000  -.022356 -.019146   44.9986
  casada*|  -.6560034      .01981  -33.12   0.000  -.694827  -.61718   .560902
t_hijo~2 |   -1.06734      .02908  -36.70   0.000  -1.12435 -1.01034   .143441
t_hijo~6 |  -.4407259      .02193  -20.10   0.000  -.483711 -.397741   .216885
t_hij~18 |   .0769087      .01025    7.50   0.000   .056812  .097005   .709008
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

La tabla anterior nos muestra que todas las variables explicativas son estadísticamente significativas, el retorno a la educación de las mujeres es de un 18.8%, un año adicional de edad disminuye el salario por hora en un 2.1%, pasar de estar no casada a casada diminuye el salario por hora promedio en 65.6%, los hijos hasta seis años disminuye el salario por hora mientras que tener hijos entre 6 y 18 años aumenta en 7.7% los salarios por hora promedio de las mujeres.

```
. mfx, predict(ystar(0,.))

Marginal effects after tobit
      y  = E(lyph*|lyph>0) (predict, ystar(0,.))
         =  1.4861891
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
     esc |   .2164857      .00284   76.25   0.000   .210921   .22205   8.21227
    edad |  -.0238113      .00094  -25.38   0.000   -.02565 -.021972   44.9986
  casada*|  -.7610779      .02327  -32.70   0.000  -.806693 -.715463   .560902
t_hijo~2 |  -1.224727      .03344  -36.62   0.000  -1.29028 -1.15918   .143441
t_hijo~6 |   -.505714      .02518  -20.08   0.000  -.555072 -.456357   .216885
t_hij~18 |   .0882494      .01176    7.50   0.000   .065191  .111308   .709008
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

La tabla anterior muestral los efectos marginales sobre la variable latente, aquella que no observamos su distribución completa. Primero, primero podemos apreciar que el valor predicho de la variable dependiente (logaritmo del salario por hora) es menor, esto porque la variable latente considera los valores de la distribución que están por debajo de cero. Todos los coeficientes son mayores (en valor absoluto) que los sobre la variable observada. Tal como se mostró en la ecuaciones (3) y (4) al considerar los efectos marginales sobre la variable observada los coeficientes se multiplican por la probabilidad que siempre es menor a 1.

Por último, se muestra el efecto marginal de la probabilidad de que el logaritmo del salario por hora sea mayor cero, efectos marginales del modelo de probabilidad binario.

Vemos que, un año adicional de escolaridad aumenta la probabilidad de que el salario por hora sea positivo en un 3.3%. Un año adicional de edad disminuye la probabilidad de que las mujeres trabajen en un 0.4%. Estar casada disminuye la probabilidad en un 11.3%, tener hijos entre 0 y 2 años disminuye la probabilidad en un 18.6%, tener hijos entre 2 y 6 años disminuye la probabilidad en un 7.7%, y tener hijos entre 6 y 18 años aumenta la probabilidad de trabajar de las mujeres en un 1.3%.

```
. mfx, predict(p(0,.))

Marginal effects after tobit
      y  = Pr(lyph>0) (predict, p(0,.))
         =  .30754248
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
     esc |   .0327913      .00043   76.66   0.000   .031953   .03363   8.21227
    edad |  -.0036067      .00014  -25.40   0.000  -.003885 -.003328   44.9986
  casada*|  -.1132121      .00338  -33.54   0.000  -.119827 -.106597   .560902
t_hijo~2 |  -.1855108      .00505  -36.72   0.000  -.195413 -.175609   .143441
t_hijo~6 |   -.076601      .00381  -20.10   0.000  -.084071 -.069131   .216885
t_hij~18 |   .0133672      .00178    7.50   0.000   .009874   .01686   .709008
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

VI.4. Sesgo de Selección (truncamiento incidental)

Cuando tenemos datos truncados como los vistos en la sección II, los datos se tienen sólo para un sub-conjunto de la población, y los datos ni de la variable dependiente ni de la variable explicativa se tienen para la muestra truncada. Cuando existe truncamiento incidental la muestra es representativa de toda la población, pero las observaciones de la variable dependiente están truncadas de acuerdo a una regla relacionada con la ecuación de interés. No observamos la variable dependiente para una sub-muestra como resultado de otras variables que generan un indicador de selección. Truncamiento incidental significa que observamos Y_i no basada en su valor sino que como resultado de otras variables. Por ejemplo, observamos las horas trabajadas cuando los individuos participan en la fuerza de trabajo, y podemos imaginar que existe una serie de variables que determinan que una mujer participe o no en la fuerza de trabajo, por lo cual podemos estimar esta probabilidad mediante un modelo probit o logit.

Para el ejemplo de las horas trabajadas donde la variable estaba truncada, podemos estimar conjuntamente la probabilidad de trabajar o no, es decir, de observar o no

horas trabajadas. La metodología para esta estimación consiste en el método de Heckman.

La ecuación que se estima y que corrige el problema de sesgo de selección es la siguiente:

$$E[Y_i \mid X_i, Z_i, S = 1] = X_i\beta + \rho\lambda(Z_i\gamma)$$

Donde S es la variable indicador de selección, cuya probabilidad de que sea igual a uno es estimada utilizando las variables Z, y γ son los coeficientes de esta estimación. La variable $\lambda()$ es el IMR que corrige el problema de truncamiento pero no basado en un umbral fijo (como en la sección II), sino con las variables que determinan la probabilidad de selección.

La siguiente tabla muestra la estimación de la ecuación de salarios, cuando no observamos salarios para el grupo de mujeres que no participa en el mercado del trabajo, para lo cual se ha considerado como determinantes de la probabilidad de participar en el mercado del trabajo: años de escolaridad, edad, estatus marital, e hijos.

```
. heckman lyph esc edad t_hijos0_2 t_hijos2_6 t_hijos6_18, select(esc edad
casada t_hijos0_2 t_hijos2_6 t_hijos6_18)

Heckman selection model                         Number of obs      =     73947
(regression model with sample selection)        Censored obs       =     51802
                                                Uncensored obs     =     22145

                                                Wald chi2(5)       =   9254.96
Log likelihood = -63620.79                      Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lyph         |
         esc |   .1563621   .0016359    95.58   0.000     .1531558    .1595685
        edad |   .0085225   .0005689    14.98   0.000     .0074074    .0096376
  t_hijos0_2 |  -.2558342    .018441   -13.87   0.000    -.2919779   -.2196905
  t_hijos2_6 |  -.1166974    .013144    -8.88   0.000    -.1424591   -.0909357
 t_hijos6_18 |   .0065165   .0058653     1.11   0.267    -.0049792    .0180122
       _cons |   4.033006   .0346434   116.41   0.000     3.965106    4.100906
-------------+----------------------------------------------------------------
select       |
         esc |     .09712   .0013709    70.85   0.000     .0944332    .0998069
        edad |  -.0106292   .0004388   -24.22   0.000    -.0114893   -.0097691
      casada |  -.2971806   .0092305   -32.20   0.000     -.315272   -.2790892
  t_hijos0_2 |  -.5684475   .0156484   -36.33   0.000    -.5991177   -.5377773
  t_hijos2_6 |  -.2397239   .0118383   -20.25   0.000    -.2629264   -.2165213
 t_hijos6_18 |   .0365936   .0055272     6.62   0.000     .0257605    .0474267
       _cons |  -.6399167    .028931   -22.12   0.000    -.6966205   -.5832129
-------------+----------------------------------------------------------------
     /athrho |    1.19052   .0216014    55.11   0.000     1.148182    1.232858
    /lnsigma |  -.0288984   .0094734    -3.05   0.002    -.0474658   -.0103309
-------------+----------------------------------------------------------------
         rho |     .83074   .0066936                      .8171508    .8434062
       sigma |   .9715152   .0092035                      .9536431    .9897222
      lambda |   .8070765   .0136363                      .7803498    .8338032
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):   chi2(1) =   766.02   Prob > chi2 = 0.0000
------------------------------------------------------------------------------
```

De los resultados, lo primero que debemos notar es que se rechaza la probabilidad de que ρ sea igual a cero, este ρ mide el grado en que el sesgo de selección es importante, si es igual a cero significa que no hay sesgo de selección.

Con respecto a la ecuación de selección se concluye que un año adicional de escolaridad aumenta la probabilidad de participar en el mercado laboral en un 9.7%, la edad disminuye la probabilidad en un 1%, estar casada disminuye la probabilidad

en un 29.7%, tener hijos entre 0 y 2 años la disminuye en un 56.8%, entre 2 y 6 años en un 24%, y tener hijos entre 6 y 18 años aumenta la probabilidad de participar en un 3.7%.

Ahora, con respecto a la ecuación de salarios se estima un retorno a la ecuación de la mujeres de un 15.6%, un año más de edad aumenta, en promedio, el salario por hora en 0.85%, el número de hijos sólo son significativos los entre 0 y 2 años y los entre 2 y 6 años, ambos afectan negativamente el salario promedio de las mujeres, en un 25.6% y 11.7%, respectivamente.

Capítulo VII. Modelos para Datos Longitudinales o Datos de Panel

VII.1. Introducción

El modelo de regresión lineal múltiple es una herramienta sumamente poderosa para estimar los efectos de variables que observamos sobre la variable de interés, sin embargo, en algunos casos no observamos todas las variables relevantes y el estimador MCO es sesgado por la omisión de variables relevantes.
En esta clase estudiaremos una metodología que nos permite controlar por algunas de estas variables aún cuando no las observamos. Esta metodología requiere un tipo particular de datos, donde cada unidad de análisis es observada dos o más periodos de tiempo. Estos datos se denominan **Datos de Panel (o Datos Longitudinales)**. La idea es que si las variables omitidas son constantes en el tiempo tomando diferencias en el tiempo de la variable dependiente, es posible eliminar estas variables omitidas que no varían en el tiempo.

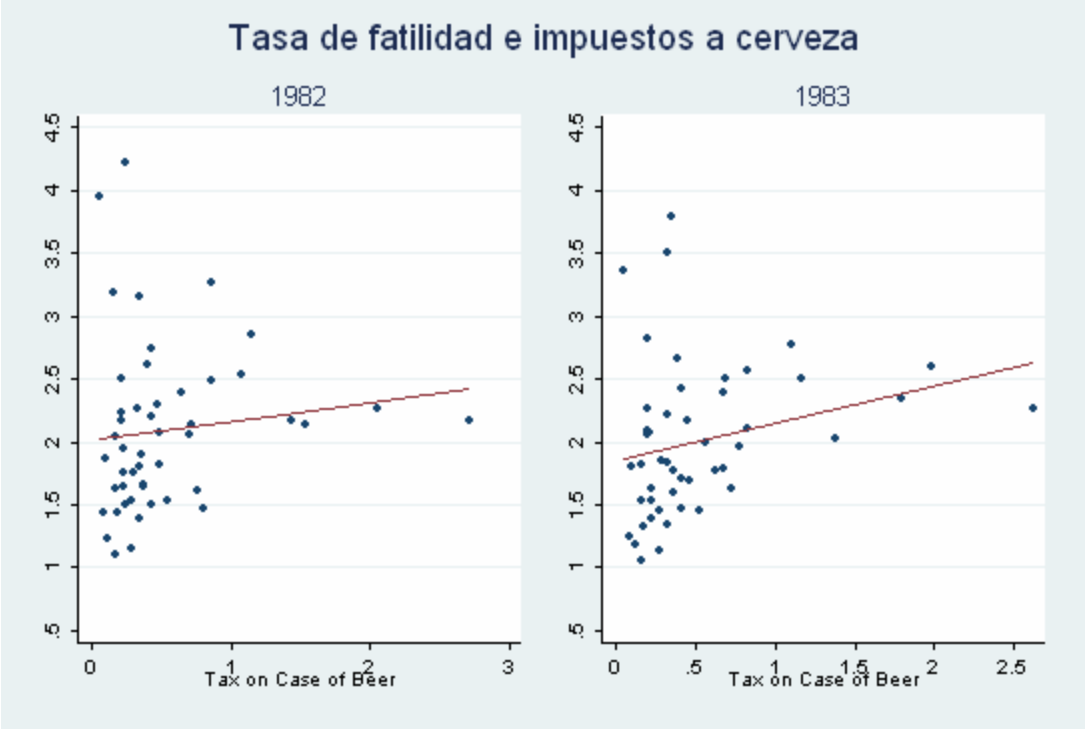
Con respecto a las base de datos panel estas pueden ser **balanceados**, es decir, se tiene para cada individuo a la misma cantidad de años, o **no balanceados** cuando algunos individuos no se observan en todo momento del tiempo. Para las metodologías de estimación presentadas en esta clase, el tener un panel balanceado o no, no afectar los resultados ni la forma de utilizar estos métodos.

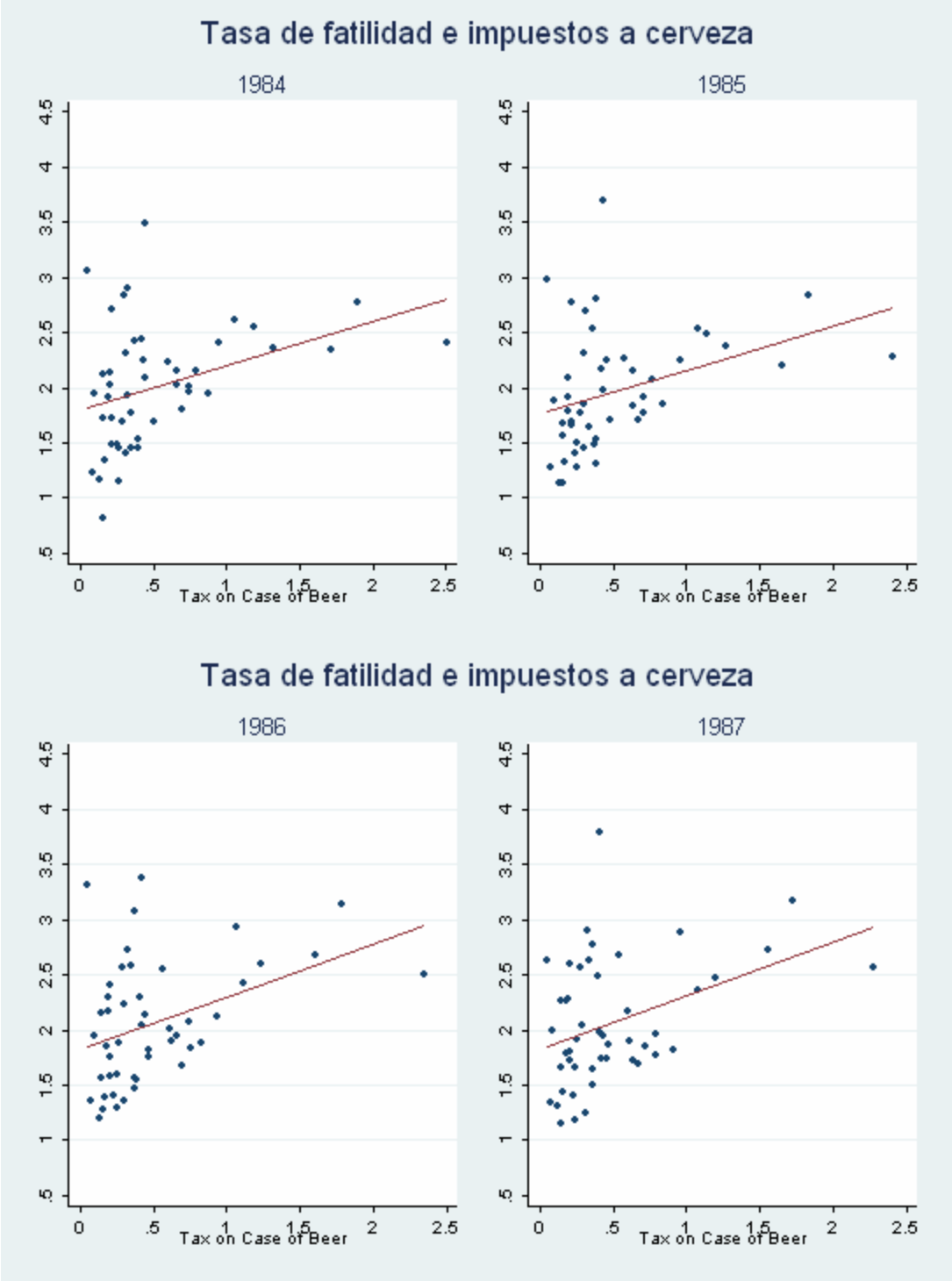
En esta clase se utilizará una base de datos panel de Estados Unidos que tiene información desde 1982 a 1988 para 48 estados, para estimar la efectividad de distintas políticas de gobierno que buscan desincentivar conducir bajo efectos del alcohol, en reducir las muertes por accidentes de tránsito. La base de datos contiene información sobre el número de accidentes fatales en cada estado en cada año, los tipos de leyes de alcohol de cada estado en cada año, y el impuesto a la cerveza en cada estado. La medida utilizada para medir las muertes por accidentes de autos se

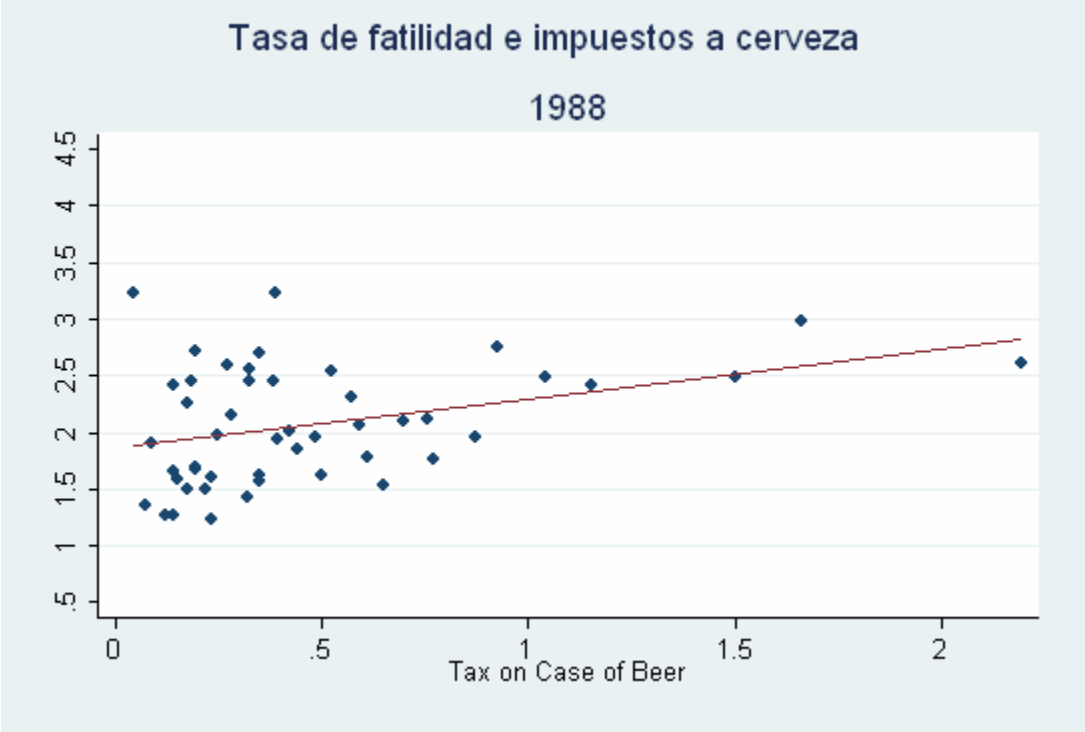
denomina tasa de fatalidad que equivale a la cantidad de personas muertas en accidentes de tránsito por cada 10.000 habitantes.

Concentrémonos primero en un modelo sencillo donde queremos explicar la tasa de fatalidad como función de los impuestos a la cerveza. El siguiente gráfico nos muestra separado para cada año el ajuste lineal entre la tasa de fatalidad y los impuestos a la cerveza.

```
use fatality.dta, clear

g fatality=allmort*10000/pop

graph drop g_1982 g_1983 g_1984 g_1985 g_1986 g_1987 g_1988

twoway (scatter fatality beertax) || (lfit fatality beertax) if year==1982,
name(g_1982) legend(off) title(1982) ylabel(0.5 (0.5) 4.5)
twoway (scatter fatality beertax) || (lfit fatality beertax) if year==1983,
name(g_1983) legend(off) title(1983) ylabel(0.5 (0.5) 4.5)
twoway (scatter fatality beertax) || (lfit fatality beertax) if year==1984,
name(g_1984) legend(off) title(1984) ylabel(0.5 (0.5) 4.5)
twoway (scatter fatality beertax) || (lfit fatality beertax) if year==1985,
name(g_1985) legend(off) title(1985) ylabel(0.5 (0.5) 4.5)
twoway (scatter fatality beertax) || (lfit fatality beertax) if year==1986,
name(g_1986) legend(off) title(1986) ylabel(0.5 (0.5) 4.5)
twoway (scatter fatality beertax) || (lfit fatality beertax) if year==1987,
name(g_1987) legend(off) title(1987) ylabel(0.5 (0.5) 4.5)
twoway (scatter fatality beertax) || (lfit fatality beertax) if year==1988,
name(g_1988) legend(off) title(1988) ylabel(0.5 (0.5) 4.5)

graph combine g_1982 g_1983, title(Tasa de fatilidad e impuestos a cerveza)
graph export g1.tif, replace
graph combine g_1984 g_1985, title(Tasa de fatilidad e impuestos a cerveza)
graph export g2.tif, replace
graph combine g_1986 g_1987, title(Tasa de fatilidad e impuestos a cerveza)
graph export g3.tif, replace
graph combine g_1988, title(Tasa de fatilidad e impuestos a cerveza)
graph export g4.tif, replace
```


Tasa de fatilidad e impuestos a cerveza
1982
1983
.5
1
1.5
2
2.5
3
3.5
4
4.5
0
1
2
3
.5
1.5
2.5
Tax on Case of Beer
Tax on Case of Beer


Tasa de fatilidad e impuestos a cerveza
1984
1985
Tax on Case of Beer
Tasa de fatilidad e impuestos a cerveza
1986
1987
Tax on Case of Beer



Las estadísticas de ambas variables por año se pueden apreciar el la siguiente tabla:

```
. tabstat  fatality, stats(mean p50 min max) by(year)

Summary for variables: fatality
     by categories of: year (Year)

    year |      mean       p50       min       max
---------+--------------------------------------------
    1982 |  2.089106   2.04551   1.10063   4.21784
    1983 |  2.007846   1.90619   1.04603   3.78745
    1984 |  2.017123    2.0169    .82121   3.48527
    1985 |  1.973671   1.86627   1.12603   3.68966
    1986 |  2.065071  1.980135   1.19247    3.3739
    1987 |  2.060696  1.924885   1.14604   3.78667
    1988 |  2.069594  1.998185   1.23111   3.23591
---------+--------------------------------------------
   Total |  2.040444  1.955955    .82121   4.21784
-------------------------------------------------

. tabstat  beertax, stats(mean p50 min max) by(year)

Summary for variables: beertax
     by categories of: year (Year)
```

year	mean	p50	min	max
1982	.5302734	.3517261	.0536993	2.720764
1983	.532393	.3547879	.0516055	2.614679
1984	.5295902	.3461539	.0494506	2.505494
1985	.5169272	.3634799	.0476695	2.415254
1986	.5086639	.371517	.0464396	2.352941
1987	.4951288	.36	.045	2.28
1988	.4798154	.346487	.0433109	2.194418
Total	.513256	.3525886	.0433109	2.720764

A continuación se muestral la estimación por MCO de los modelos separados para cada año:

Variable	M1982	M1983	M1984	M1985	M1986	M1987	M1988
beertax	.148	.299	.4	.392	.48	.483	.439
	0.435	0.082	0.011	0.015	0.005	0.007	0.011
_cons	2.01	1.85	1.81	1.77	1.82	1.82	1.86
	0.000	0.000	0.000	0.000	0.000	0.000	0.000
r2_a	-.00813	.0439	.112	.103	.143	.131	.115
rmse	.67	.592	.517	.51	.518	.526	.49

legend: b/p

Tabla que se obtuvo de ejecutar los siguientes comandos:

```
quietly reg fatality beertax if year==1982
estimates store M1982
quietly reg fatality beertax if year==1983
estimates store M1983
quietly reg fatality beertax if year==1984
estimates store M1984
quietly reg fatality beertax if year==1985
estimates store M1985
quietly reg fatality beertax if year==1986
estimates store M1986
quietly reg fatality beertax if year==1987
estimates store M1987
quietly reg fatality beertax if year==1988
estimates store M1988

estimates table M1982 M1983 M1984 M1985 M1986 M1987 M1988, stat(r2_a, rmse)
b(%7.3g) p(%4.3f)
```

Se puede apreciar que en el año 1982 los impuestos a la cerveza no son significativos en explicar la tasa de accidentes automovilísticos. Sin embargo, para todos los demás años existe un impacto significativo (a distintos niveles de significancia), los valores del parámetro estimado tiene un rango entre 0.3 y 0.48, lo que significa que US$1 adicional de impuesto a la cerveza por caja, aumenta la mortalidad en accidentes de autos en 0.3/0.48 personas por cada 10.000 habitantes.

¿Podemos concluir entonces que un mayor impuesto a la cerveza genera un aumento en las muertes por accidentes de autos?

No necesariamente, ya que la especificación anterior probablemente sufra de un problema de omisión de variables relevantes. Existen muchos otros factores que afectan la tasa de accidentes mortales, incluyendo la calidad de los automóviles que son manejados en cada estado, la calidad de las autopistas o calles de cada estado, la densidad de autos en las calles, si es socialmente aceptado el beber o no alcohol, etc. Cualquiera de estos factores puede estar correlacionado con el impuesto a la cerveza, y generar un problema de estimación en los modelos anterior, por la omisión de variables relevantes.

Una solución para este problema podría ser recolectar la información faltante, pero alguna de esta información como si es aceptable o no beber, no se puede medir. Si estos factores se mantienen constantes en el tiempo en un estado dado, existe otra posibilidad de estimación si contamos con información de datos de panel, esto consiste en hacer una regresión MCO con efectos individuales.

VII.2. Datos de panel con dos periodos: comparación antes y después

Cuando los datos para cada estado son obtenidos para dos periodos de tiempo, es posible computar el cambio en la variable dependiente para el análisis, esta comparación "antes-después" mantiene constante todos los factores no observables que difieren entre estados pero no en el tiempo para un mismo estado.

Supongamos que sólo tenemos los datos para los años 1982 y 1988, y que existe una variable *W*, por ejemplo, aceptación social por beber alcohol que deberíamos incluir en el modelo antes estimado:

$$fatality_{i,1982} = \beta_0 + \beta_1 BeerTax_{i,1982} + \gamma W_i + \mu_{i,1982}$$
$$fatality_{i,1988} = \beta_0 + \beta_1 BeerTax_{i,1988} + \gamma W_i + \mu_{i,1988}$$

Pero en la práctica no observamos la variable *W*, pero como este es constante entre 1982 y 1988, no varia en el tiempo para un mismo estado, restando ambas ecuaciones anteriores se puede eliminar el efecto de *W*.

El modelo a estimar en este caso es:

$$fatality_{i,1988} - fatality_{i,1982} = \beta_0 + \beta_1 (BeerTax_{i,1988} - BeerTax_{i,1982}) + \mu_{i,1988} - \mu_{i,1982}$$

Para lo cual se hace la siguiente transformación en la base de datos que permitan generar las versiones en diferencias de la variable dependiente y de las variables explicativas.

```
keep fatality beertax year state
keep if year==1982 | year==1988

reshape wide fatality beertax, i(state) j(year)

g dfatality= fatality1988- fatality1982
g dbeertax=beertax1988-beertax1982
```

```
reg  dfatality dbeertax, nocon

      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  1,    47) =    4.89
       Model |  .765652898     1  .765652898           Prob > F      =  0.0319
    Residual |  7.36082332    47  .156613262           R-squared     =  0.0942
-------------+------------------------------           Adj R-squared =  0.0749
       Total |  8.12647622    48  .169301588           Root MSE      =  .39574

------------------------------------------------------------------------------
   dfatality |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    dbeertax |  -.8689218   .3929877    -2.21   0.032    -1.659511   -.0783325
------------------------------------------------------------------------------
```

En contraste con la estimación de corte transversal vista en un comienzo, el efecto estimado de un cambio en el impuesto a la cerveza sobre la tasa de fatalidad es negativo, tal como se esperaba que fuese. Un aumento del impuesto por caja de la cerveza en US$1 reduce la tasa de fatalidad en accidentes de autos en 0.87 muertes por cada 10.000 habitantes. El efecto estimado es bastante grande, el promedio de la tasa de fatalidad es de 2 muertes por cada 10.000 habitantes, esta estimación nos indica que la tasa de fatalidad se puede reducir a la mitad aumenta el impuesto a la cerveza en US$1 por caja.

La estimación antes después es una de las técnicas utilizadas para la evaluación de programas, este estimador se llama estimador en diferencias, donde el impacto de recibir cierto tratamiento **T** se mide por el cambio en la variable de interés. Por ejemplo, si el tratamiento es la disminución en los alumnos por cursos. Observamos los alumnos por cursos antes del "tratamiento" y después, y observamos el rendimiento antes y después, la estimación del modelo simple de estas dos variables en diferencias mide el impacto sobre el rendimiento de disminuir los alumnos por cursos, todos los factores que son diferentes entre colegios, pero que no varían entre un año y otro están siendo eliminados al considerar la diferencia de las variables.

La estructura general de los modelos de regresión con datos de panel, ya sea efectos fijos o efectos aleatorios es la siguiente:

$$(1) \quad Y_{it} = \beta_1 X_{it} + \gamma Z_i + u_i + \varepsilon_{it}$$

Donde X_{it} representan las variables explicativas que varían entre individuos y en el tiempo, Z_i representan variables explicativas que sólo varían entre individuos pero no en el tiempo, u_i representa un efecto individual de nivel, y ε_{it} corresponde al error del modelo.

El estimador de **efectos aleatorios** asume que u_i no esta correlacionado con las variables explicativas (X_{it} y Z_i) del modelo, de esta forma el efecto individual de nivel es simplemente otro componente del término de error. Es decir, en modelos de efectos aleatorios el término de error es $u_i + \varepsilon_{it}$.

Deberíamos ocupar efectos aleatorios cuando, por una parte tenemos acceso a las variables explicativas Z_i que son constantes en el tiempo pero que varían entre individuos, y todo lo que es efecto individual no observado es completamente aleatorio.

El estimador de **efectos fijos** se utiliza cuando u_i esta correlacionado con las variables explicativas del modelo, por lo que tiene que se modelado de alguna forma y se tratan como parámetros a estimar del modelo, en este caso denotaremos a estos efectos individuales como α_i, para identificar en términos de notación que son variables fijas y no aleatorias. El problema del estimador de efectos fijos es que no permite incluir variables explicativas que no varían en el para las cuales tenemos observaciones. Todo lo que sea constante en el tiempo para los individuos, quedará

obligatoriamente capturado por este parámetro constante individual (α_i). De esta forma, el modelo de efectos fijos debe quedar especificado de la siguiente forma:

$$(2) \quad Y_{it} = \alpha_i + \beta_1 X_{it} + \varepsilon_{it}$$

VII.3.1. Modelo de regresión de efectos fijos

La regresión por efectos fijos es un método que permite controlar por variables omitidas cuando estas variables no varían en el tiempo pero si entre la unidad de observación. La estimación por efectos fijos es equivalente a la estimación antes-después de la sección II, pero puede ser utilizando cuando tenemos dos o más periodos de tiempo.

La regresión de efectos fijos estima n diferentes interceptos, uno para cada unidad de observación. Estos interceptos pueden ser representados por una serie de variables binarias (dummies), las que absorberán la influencia de TODAS las que son constantes en el tiempo.

Consideremos el siguiente modelo de regresión donde tenemos una variable explicativa que varían tanto entre individuos como en el tiempo X_{it} y otra variable explicativas que es constante en el tiempo W_i, pero que no observamos. La estimación por efectos fijos nos permitirá controlar por la presencia de estas variables fijas aunque en la práctica no las observamos. De esta forma, el modelo se puede especificar de la siguiente forma:

$$(3) \quad Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 W_i + \varepsilon_{it} \quad i = 1,...,n \; y \; t = 1,...,T$$

La ecuación (3) puede ser interpretada como que tiene *n* interceptos distintos:

$$\alpha_i = \beta_0 + \beta_2 W_i$$

Es decir,

$$(4) \quad Y_{it} = \alpha_i + \beta_1 X_{it} + \varepsilon_{it}$$

La ecuación (4) representa el **modelo de regresión por efectos fijos**, en donde α_1, α_2,…,α_n son tratados como parámetros desconocidos que deben ser estimados, uno para cada estado, y β_1 representa el efecto de la variable explicativa sobre la variable de interés pero que esta libre de sesgo de omisión de variables relevantes que son constantes en el tiempo para cada estado. El problema es que observamos algunas variables explicativas que son constantes en el tiempo, no se podrá estimar separadamente un coeficiente para esta variable, todo esto queda capturado por el efecto fijo.

El modelo (4) también puede ser expresando utilizando variables binarias, sea D1 igual a 1 cuando i=1 y cero en otro caso, sea D2 igual a uno cuando i=2 y cero en otro casos, etc. Recuerde que no podemos incluir todas las dummies al mismo tiempo más la constante porque existirá multicolineallidad perfecta, omitamos D1:

$$(3) \quad Y_{it} = \beta_0 + \beta_1 X_{it} + \varphi_2 D2_i + \varphi_3 D3_i + .. + \varphi_n Dn_i + \varepsilon_{it}$$

Podemos establecer una relación entre los coeficientes del modelo (2) y el (3):

$$\begin{aligned} \alpha_1 &= \beta_0 \\ \alpha_2 &= \beta_0 + \varphi_2 \\ &\vdots \\ \alpha_n &= \beta_0 + \varphi_n \end{aligned}$$

Si uno no esta interesado explícitamente en estimar estos efectos individuales (α_i), los que además pueden aumentar significativamente el número de parámetros a estimar cuando las unidades o individuos son un número elevado, se puede

transformar el modelo en (4) de forma tal de eliminar este efecto individual invariante en el tiempo. Esta transformación se conoce como la transformación **within**, la que consiste en tomar el promedio de las variables en el tiempo para cada individuo i y luego restar a cada indiviudo este promedio. Como el efecto individual no varia en el tiempo, esta transformación lo elimina.

Tomando el promedio de (4) en t al interior de cada unidad i:

$$\bar{Y}_i = \alpha_i + \beta_1 \bar{X}_i + \underbrace{\bar{\varepsilon}_i}_{0}$$

Luego restando a (4) la expresión anterior:

$$Y_{it} - \bar{Y}_i = \beta_1 (X_{it} - \bar{X}_i) + \varepsilon_{it}$$

Así, el estimador within de β_1 es:

$$\beta_1^W = \frac{\sum_{i=1}^{N} \sum_{t=}^{T} (Y_{it} - \bar{Y}_i)(X_{it} - \bar{X}_i)}{\sum_{i=1}^{N} \sum_{t=}^{T} (X_{it} - \bar{X}_i)^2}$$

El poder explicativo del estimador de efectos fijos (within) depende de la variación de la variable dependiente y las explicativas al interior de la unidad de observación.

Para estimar modelos de efecto fijos en STATA se debe utilizar el comando **xtreg**, el cual requiere que previamente se defina la variable que representa al individuo y la variable que representa el tiempo, a través de la siguiente indicación:

```
iis state
tis year
```

Utilicemos los datos para el año 1982 y 1988 con los que estimamos el modelo en diferencias, pero utilicemos la estimación de efectos fijos, podrán apreciar que se obtienen exactamente los mismos resultados, un US$1 adicional de impuesto a la caja de cervezas disminuye en 0.86 los muertos cada 10.000 habitantes por accidentes de autos.

```
. xtreg fatality beertax, fe

Fixed-effects (within) regression               Number of obs      =        96
Group variable (i): state                       Number of groups   =        48

R-sq:  within  = 0.0942                         Obs per group: min =         2
       between = 0.0622                                        avg =       2.0
       overall = 0.0462                                        max =         2

                                                F(1,47)            =      4.89
corr(u_i, Xb)  = -0.7038                        Prob > F           =    0.0319

------------------------------------------------------------------------------
    fatality |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |  -.8689218   .3929877    -2.21   0.032    -1.659511   -.0783325
       _cons |   2.518194   .2005206    12.56   0.000     2.114799     2.92159
-------------+----------------------------------------------------------------
     sigma_u |  .77501684
     sigma_e |  .27983322
         rho |  .88466641   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(47, 47) =     7.74                 Prob > F = 0.0000
```

El mismo resultado se puede obtener a partir de la estimación del modelo con variables binarias, mediante la utilización del siguiente comando:

```
xi: reg fatality beertax i.state
```

```
i.state           _Istate_1-56        (naturally coded; _Istate_1 omitted)

      Source |       SS       df       MS              Number of obs =      96
-------------+------------------------------           F( 48,    47) =    8.00
       Model |  30.0536336    48  .626117366           Prob > F      =  0.0000
    Residual |  3.68041166    47  .078306631           R-squared     =  0.8909
-------------+------------------------------           Adj R-squared =  0.7795
       Total |  33.7340452    95  .355095213           Root MSE      =  .27983

------------------------------------------------------------------------------
    fatality |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |  -.8689218   .3929877    -2.21   0.032    -1.659511   -.0783325
   _Istate_4 |  -.7860027   .5618579    -1.40   0.168    -1.916315    .3443096
   _Istate_5 |  -.6562953   .4612301    -1.42   0.161    -1.584171      .27158

.....
```

Ahora que hemos comprobado que las tres tipos de estimaciones nos entregan la misma estimación del parámetro de interés β, utilicemos todos los años (1982 a 1988) para la estimación del modelo por efectos fijos.

```
Fixed-effects (within) regression               Number of obs      =       336
Group variable (i): state                       Number of groups   =        48

R-sq:  within  = 0.0407                         Obs per group: min =         7
       between = 0.1101                                        avg =       7.0
       overall = 0.0934                                        max =         7

                                                F(1,287)           =     12.19
corr(u_i, Xb)  = -0.6885                        Prob > F           =    0.0006

------------------------------------------------------------------------------
    fatality |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |  -.6558737     .18785    -3.49   0.001    -1.025612   -.2861353
       _cons |   2.377075   .0969699    24.51   0.000     2.186213    2.567937
-------------+----------------------------------------------------------------
     sigma_u |  .71471463
     sigma_e |  .18985942
         rho |  .93408484   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(47, 287) =    52.18             Prob > F = 0.0000
```

La estimación anterior se realiza para los 48 estados, cada estado tiene 7 observaciones, lo que suma un total de 336 observaciones. Se encuentra un impacto negativo y significativo, que indica que US$1 adicional de impuesto por caja de

cerveza disminuye los muertos en accidentes de autos en 0.65 personas por cada 10.000 habitantes.

En el output además aparece σ_u^2 (sigma_u), σ_ε^2 (sigma_e) y rho; donde rho indica la parte de la varianza total que corresponde a la varianza de u. Además el output entrega la correlación entre u y las variables explicativas del modelo, en este caso la correlación es muy distinta de cero lo que va a favor del modelo de efectos fijos, ya que efectos aleatorios asume que esta correlación es cero. Por último, al final se entrega el test de la hipótesis conjunta de que todos los coeficientes de efecto fijo son iguales a cero, lo que validaría la simple estimación por MCO tomando una sola constante común para todos (pooled MCO), sin embargo, en este caso se rechaza la hipótesis nula de que todos los coeficientes son iguales a cero.

VII.3.2. Efectos fijos de tiempo

Tal como los efectos fijos vistos hasta ahora controlan por efectos que son constantes en el tiempo pero varían entre unidades, podemos tratar de controlar por efectos fijos que varían en el tiempo pero son constantes entre individuos. Para controlar por estos efectos fijos de tiempo se deben introducir variables dummies para cada año:

```
. xtreg fatality beertax DY_2 DY_3 DY_4 DY_5 DY_6 DY_7, fe

Fixed-effects (within) regression               Number of obs      =       336
Group variable (i): state                       Number of groups   =        48

R-sq:  within  = 0.0803                         Obs per group: min =         7
       between = 0.1101                                        avg =       7.0
       overall = 0.0876                                        max =         7

                                                F(7,281)           =      3.50
corr(u_i, Xb)  = -0.6781                        Prob > F           =    0.0013

------------------------------------------------------------------------------
    fatality |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |    -.63998   .1973768    -3.24   0.001    -1.028505   -.2514552
        DY_2 |  -.0799029   .0383537    -2.08   0.038       -.1554   -.0044058
        DY_3 |  -.0724205   .0383517    -1.89   0.060    -.1479136    .0030725
        DY_4 |  -.1239763   .0384418    -3.23   0.001    -.1996468   -.0483058
        DY_5 |  -.0378645   .0385879    -0.98   0.327    -.1138225    .0380936
        DY_6 |  -.0509021   .0389737    -1.31   0.193    -.1276196    .0258155
        DY_7 |  -.0518038   .0396235    -1.31   0.192    -.1298003    .0261927
       _cons |    2.42847   .1081198    22.46   0.000     2.215643    2.641298
-------------+----------------------------------------------------------------
     sigma_u |  .70945969
     sigma_e |  .18788295
         rho |  .93446373   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(47, 281) =    53.19             Prob > F = 0.0000

. test DY_2 DY_3 DY_4 DY_5 DY_6 DY_7

 ( 1)  DY_2 = 0
 ( 2)  DY_3 = 0
 ( 3)  DY_4 = 0
 ( 4)  DY_5 = 0
 ( 5)  DY_6 = 0
 ( 6)  DY_7 = 0

       F(  6,   281) =    2.01
            Prob > F =    0.0642
```

Los resultados obtenidos en este caso son similares a los del modelo anterior, un aumento en US$1 el impuesto por caja de cerveza disminuye las muertes en accidentes de autos en 0.64 personas por cada 100.000 habitantes. Las dummies de los años 83 y 85 resultan significativas a un 5%, la dummy del año 84 es significativa al 6%. Los otros años no tienen efectos fijos significativos.

VII.3.3. Modelo de regresión de efectos aleatorios

El estimador de efectos aleatorios considera que el componente individual en la estimación no es fijo sino aleatorio, por lo tanto que no esta correlacionado con las variables explicativas del modelo. Este supuesto permite incorporar variables explicativas que observamos y son constantes en el tiempo, esto no lo podemos hacer cuando estimamos por efectos fijos.

El modelo de efectos fijos asume un error compuesto:

$$Y_{it} = \beta_1 X_{it} + \gamma Z_i + u_i + \varepsilon_{it}$$

Donde u_i son los efectos individuales.

Como este efecto individual no esta correlacionado con las variables explicativas, se podría estimar el modelo por MCO, el problema es que esta estimación no es eficiente porque no considera la información que parte del término de error esta compuesto por un efecto individual que se puede estimar para computar correctamente la matriz de varianzas y covarianzas. La ventaja del modelo de efectos aleatorios es que incorpora dentro de la estimación este efecto como parte del término de error, obteniendo una estimación eficiente pero que será consistente sólo bajo el supuesto de que este componente del error (el componente de efecto individual) no esta correlacionado con las variables explicativas.

La siguiente tabla muestra el resultado de la estimación del modelo anterior por efectos aleatorios:

```
. xtreg fatality beertax, re

Random-effects GLS regression                   Number of obs      =       336
Group variable (i): state                       Number of groups   =        48

R-sq:  within  = 0.0407                         Obs per group: min =         7
       between = 0.1101                                        avg =       7.0
       overall = 0.0934                                        max =         7

Random effects u_i ~ Gaussian                   Wald chi2(1)       =      0.18
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.6753

------------------------------------------------------------------------------
    fatality |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |  -.0520158   .1241758    -0.42   0.675    -.2953959    .1913643
       _cons |   2.067141   .0999715    20.68   0.000     1.871201    2.263082
-------------+----------------------------------------------------------------
     sigma_u |   .5157915
     sigma_e |  .18985942
         rho |  .88067496   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Los resultados cambian radicalmente con respecto a los obtenidos por efectos fijos, se obtiene un efecto muy pequeño y no significativa, ¿Por qué?.

Notemos que el estimador de efectos de aleatorios impone que la correlación entre el componente individual aleatorio y las variables explicativas del modelo es igual a cero. Pero en realidad, de la estimación de efectos aleatorios teníamos que existía una correlación importante entre el componente individual y el efecto aleatorio, lo que inválida este supuesto y hace que la estimación por este método sea inconsistente.

VII.3.4. Testeando la validez de efectos aleatorios

La hipótesis nula que se quiere testear es si el supuesto de no correlación entre el componente individual y las variables explicativas. Cuando se cumple el supuesto de

no correlación el estimador de efectos aleatorios es consistente y eficiente, sin embargo, cuando no se cumple este supuesto el estimador de efectos aleatorios es inconsistente. El estimador de efectos fijos siempre es consistente pero menos eficiente. Este antecedente nos permite ocupar el test de hausman para ver la validez del estimador de efectos aleatorio, la hipótesis nula es que no existe correlación entre el componente individual y las variables explicativas del modelo, bajo la hipótesis nula el coeficiente estimado por efectos fijos y el estimado por efectos aleatorios no debería diferir significativamente.

A continuación se muestra el procedimiento para realizar este test:

```
xtreg fatality beertax, fe
estimates store fix

xtreg fatality beertax, re
estimates store ran

hausman fix ran
```

```
                ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |      fix          ran         Difference          S.E.
-------------+----------------------------------------------------------------
     beertax |   -.6558737    -.0520158       -.6038579        .1409539
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       18.35
                Prob>chi2 =      0.0000
```

Se rechaza la hipótesis nula de que los coeficientes son iguales, con lo cual se concluye que el estimador de efectos aleatorios no es apropiado.

Capítulo VIII. Modelos de Duración

VIII.1. Introducción

Cuando una persona esta desempleada, o cuando los trabajadores de una empresa están en huelga, podríamos esperar que mientras más tiempo se ha permanecido en ese "estado" mayor es la probabilidad de que la persona encuentre un trabajo o de que la huelga termine en las próximas semanas. Pero también podríamos pensar que mientras más tiempo ha durado este estado las características que provocaron este estado son más fuertes y por lo tanto es poco probable salir de este estado. En este tipo de problemas no sólo interesa el tiempo transcurrido en cierto estado, sino además interesa la probabilidad de transición a otro estado.

En la clase de hoy se estudiarán los **Modelos de duración**, en estos modelos la variable de interés es el tiempo transcurrido desde el inicio de cierto evento hasta el término de este o hasta que donde se tiene acceso a la información (fecha de medición). En los modelos de duración la censura en los datos es un problema inherente a la forma de obtener los datos, la estimación de estos modelos debe considerar el problema de censura presente en los datos.

Para efectos del análisis de los modelos de duración, se definirá **estado** como la clasificación de un individuo o unidad en un momento del tiempo, **transición** es el movimiento de un estado a otro, y **duración (spell)** corresponde al tiempo transcurrido en cierto estado.

La variable de interés en los modelos de duración corresponde al tiempo transcurrido entre el inicio de cierto estado hasta que termina o hasta cuando la medición fue realizada. Los datos que debemos poseer para hacer una análisis de duración consiste en tiempos de duración de un estado: t_1, t_2,…,t_N.

La Encuesta de Protección Social 2002 y 2004 reporta la historia laboral completa desde Enero de 1980 hasta Diciembre de 2004 para los individuos entrevistados. En la clase de hoy se utilizará esta base de datos, donde se han identificado los periodos de cesantía de cada uno de los individuos. Sólo se consideraron los periodos entre los 18 y 60 años de edad. Se ha determinado, entre otras variables, la duración de la cesantía y la observación esta censurada, es decir, corresponde a la situación laboral a la fecha de la ultima medición de la encuesta (Diciembre de 2004), edad de inicio de la cesantía, entre otras variables.

En los modelos de duración un aspecto fundamental es contar con la base de datos organizada de forma apropiada. A continuación se muestra la forma en que una base de datos para modelos de duración debería estar estructurada, donde spell indica los meses en que la persona ha estado desempleado, y censura indica si el evento término o no, específicamente esta variable debe tomar valor 1 si la observación ha "fallado" es decir, el estado ha terminado (no esta censurada), y cero si no ha terminado y aún se encuentra en este estado cuando se levantaron los datos.

```
. list  folio  fecha_ini fecha_ter spell censura

      +-------------------------------------------------+
      |   folio   fecha_~i   fecha_~r   spell   censura |
      |-------------------------------------------------|
   1. |     179     198001     198106      18         1 |
   2. |     179     198107     198206      12         1 |
   3. |     179     198207     198212       6         1 |
   4. |     179     198301     198501      25         1 |
   5. |     179     198502     199212      95         0 |
      |-------------------------------------------------|
   6. |     785     198001     198712      96         0 |
   7. |     882     198001     198506      66         1 |
   8. |     882     198507     200412     234         0 |
   9. |     948     198001     198204      28         1 |
  10. |     948     198205     198208       4         1 |
      |-------------------------------------------------|
  11. |     948     198209     200412     268         0 |
  12. |    1267     199201     199912      96         1 |
  13. |    1267     200001     200104      16         1 |
  14. |    1267     200105     200412      44         0 |
  15. |    1401     198501     198812      48         1 |
```

¿Por qué no se estima un modelo por MCO donde la variable dependiente es el tiempo de duración del estado contra las variables explicativas relevantes?

En efecto, se podría estimar el siguiente modelo por MCO utilizando el comando **regress** de STATA:

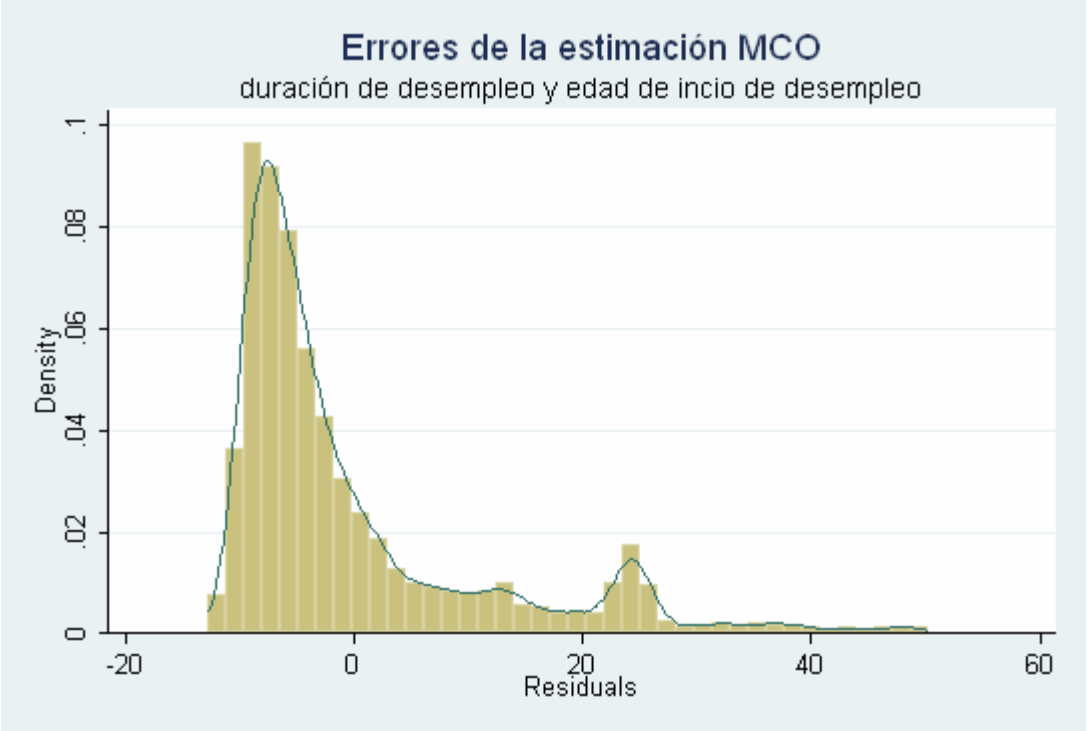
$$spell_i = \alpha + \beta \cdot edad_i + \mu_i$$

Los resultados de la estimación de este modelo son los siguientes:

```
. reg spell edad_ini, robust

Linear regression                                      Number of obs =   10473
                                                       F(  1, 10471) =   71.90
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0074
                                                       Root MSE      =  11.856

------------------------------------------------------------------------------
             |               Robust
       spell |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    edad_ini |   .0968173   .0114179     8.48   0.000      .074436    .1191987
       _cons |   7.941263    .373719    21.25   0.000     7.208703    8.673824
------------------------------------------------------------------------------
```

El problema en la estimación por MCO es que se asume normalidad del término de error, sin embargo, al obtener los errores de la estimación anterior podemos apreciar un claro desvío de la normalidad.

```
. predict u_mco, resid

. sum  u_mco, detail

                        Residuals
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -11.29804       -12.7503
 5%    -9.975764       -12.7503
10%    -9.201226       -12.7503       Obs                10473
25%    -7.780793       -12.7503       Sum of Wgt.        10473

50%    -4.717139                      Mean           2.12e-09
                        Largest       Std. Dev.      11.85588
75%     2.219208       50.31602
90%     20.73512       50.31602       Variance       140.5618
95%     25.21921       50.31602       Skewness       1.774382
99%     41.73512       50.31602       Kurtosis       5.700185


. histogram u_mco, kdensity title(Errores de la estimación MCO)
subtitle(duración de desempleo y edad de incio de desempleo)
```



En la mayoría de los problemas involucrados en los modelos de duración la distribución del evento no es normal. Adicionalmente, la distribución normal toma el rango completo de valores, negativos y positivos, sin embargo, el tiempo de duración de cierto evento siempre es positivo.

Así, los modelos de duración lo que hacen es sustituir el supuesto de normalidad que caracteriza la estimación por MCO, por algo más apropiado para estructura particular de estos modelos.

VIII.3. Función de Sobrevivencia (survivor) y Hazard rate

Sea T una variable aleatoria no negativa que denota el tiempo en que cierto evento termina (falla), como T es una variable aleatoria podemos definir su función de densidad y función de distribución de probabilidad acumulada como:

$$f(t)$$
$$F(t) = \Pr[T \leq t]$$

Se define la **función de sobrevivencia** (survivor function), la que representa la probabilidad de sobrevivir más allá de t, o la probabilidad de que el evento dure al menos t:

$$S(t) = 1 - F(t) = \Pr[T \geq t]$$

En los modelos de duración la función de densidad *f(t)* se reemplaza por la **función hazard**, que se denota por *h(t)*, y se denomina la tasa de falla condicional o tasa instantánea de falla:

$$h(t) = \frac{f(t)}{S(t)}$$

La función hazard puede tomar valores desde cero, significando que no hay riesgo de falla, hasta infinito significando que falla cierta en este instante.
Existe una relación uno a uno entre la probabilidad de sobrevivir pasado cierto tiempo y la cantidad de riesgo que ha sido acumulada hasta esa cantidad de tiempo, la hazard rate mide la *tasa a la cual el riesgo es acumulado*.

Adicionalmente se puede definir la función hazard acumulada, la que mide la cantidad de riesgo acumulado hasta el tiempo t.

Los dos ejemplos más comunes de función hazard son la exponencial y weibull. La función exponencial tiene una hazard rate constante, es decir, la tasa a la cual se van completando los eventos es constante en la duración de los eventos, y así la función de sobreviencia disminuye en forma exponencial:

$$h(t) = c$$
$$S(t) = \exp(-ct)$$

La función Weibull tiene la siguiente hazard rate y función de sobrevivencia:

$$h(t) = pt^{p-1}$$
$$S(t) = \exp(-t^{p})$$

Donde *p* es un parámetro que debe ser estimado a partir de los datos.

Para poder estimar modelos de duración en STATA lo primero que debemos hacer es indicarle al programa que la estructura de la base de datos tiene estas características. Esto se hace a través del comando **stset**, en el cual debemos indicar la variable que mide el tiempo de duración del evento, la variable que indica si el evento esta censurado o no, y la variable que identifica a cada individuo si es que hay más de una observación por individuo.

En la base de datos la variable spell indica la duración del estado desempleado, y la variable censura indica si el estado desempleado ha terminado o esta censurado, es decir, en el momento de observación o de levantamiento de datos la persona de encontraba desempleada por lo cual aún no ha finalizado este estado. Para indicar que estamos trabajando con datos de duración debemos ejecutar el siguiente comando:

```
stset spell, failure(censura)

     failure event:  censura != 0 & censura < .
obs. time interval:  (0, spell]
 exit on or before:  failure
```

```
------------------------------------------------------------------------------
    10473  total obs.
        0  exclusions
------------------------------------------------------------------------------
    10473  obs. remaining, representing
     9145  failures in single record/single failure data
   115866  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =        60
```

Luego podemos estimar a través de los datos de duración de desempleo que observamos la funciones hazard rate y función de sobrevivencia que nos muestran la tasa a la cual se acumula el riesgo de terminar el desempleo, y la probabilidad de que la persona siga desempleada condicional que se ha cumplido t meses de desempleo, respectivamente.

Con el siguiente comando estimamos el parámetro de la función exponencial:

```
streg, dist(exponential)
```

Del cual obtenemos el siguiente resultado:

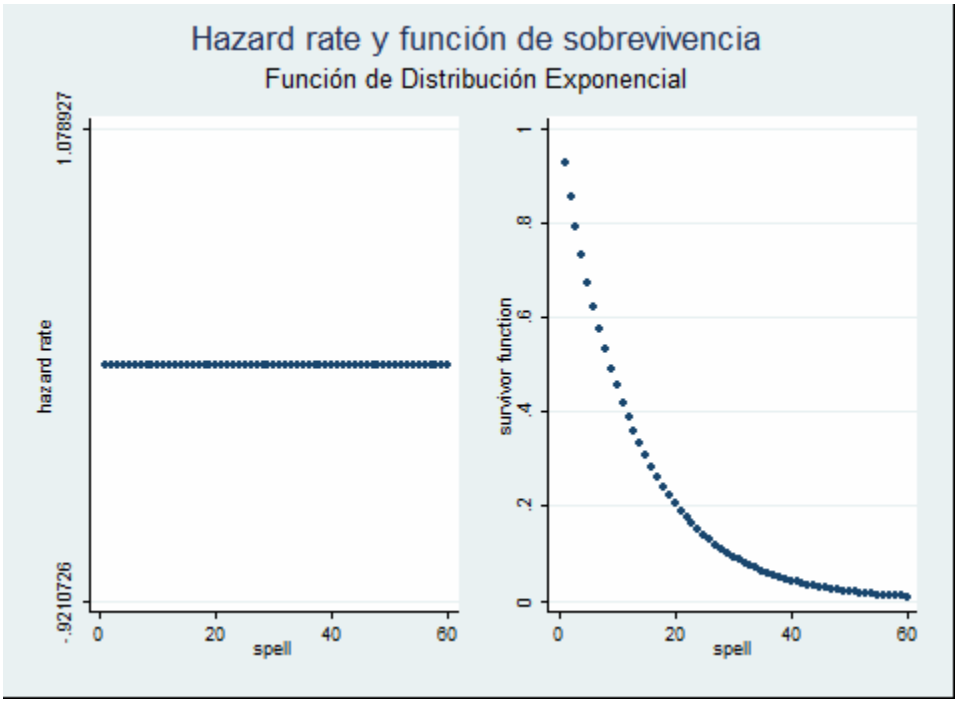
```
        failure _d:  censura
  analysis time _t:  spell

Iteration 0:   log likelihood =  -15834.87
Iteration 1:   log likelihood =  -15834.87

Exponential regression -- log relative-hazard form

No. of subjects =        10473                     Number of obs   =     10473
No. of failures =         9145
Time at risk    =       115866
                                                   LR chi2(0)      =      0.00
Log likelihood  =    -15834.87                     Prob > chi2     =         .

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
------------------------------------------------------------------------------
```

Luego para obtener la variable que contenga hazard rate y otra que contenga la función de sobrevivencia se debe utilizar el comando predict:

```
predict h_e, hazard
predict S_e, surv
```

Y luego se puede graficar cada una de estas funciones con relación al tiempo utilizando los siguientes comandos.

```
twoway (scatter h_e spell), name(exponential_h) ytitle(hazard rate)
xtitle(spell) legend(title(exponential))
twoway (scatter S_e spell), name(exponential_s) ytitle(survivor function)
xtitle(spell)
graph combine exponential_h exponential_s, title(Hazard rate y función de
sobrevivencia) subtitle(Función de Distribución Exponencial)
```



Luego con el siguiente comando estimamos el parámetro de la función weibull:

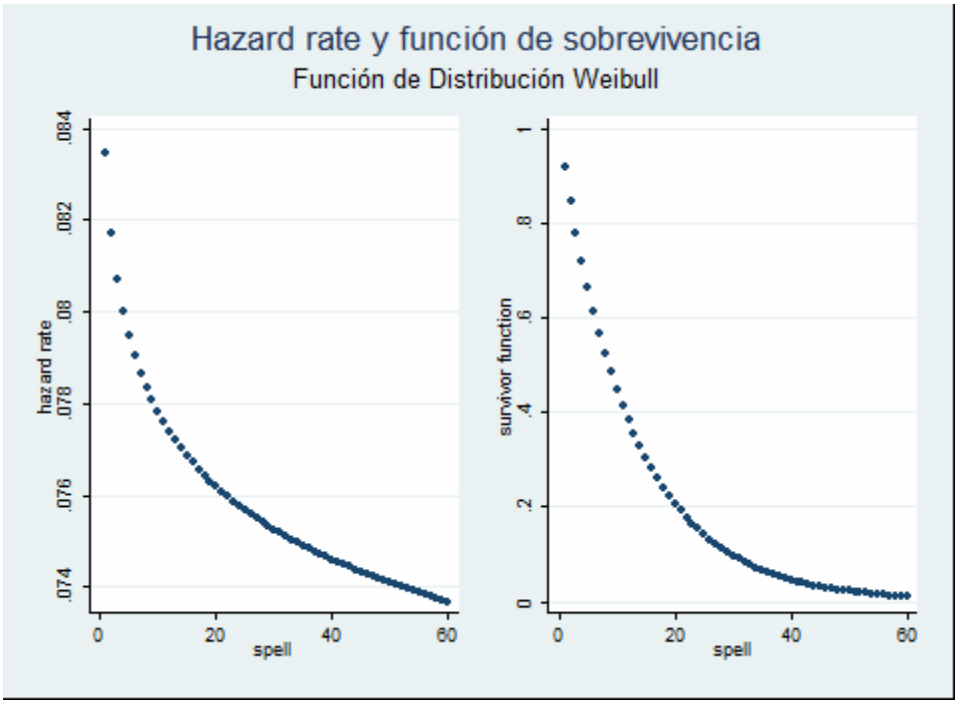
```
streg, dist(weibull)

         failure _d:  censura
   analysis time _t:  spell

Fitting constant-only model:

Iteration 0:   log likelihood =  -15834.87
Iteration 1:   log likelihood = -15827.071
Iteration 2:   log likelihood = -15827.071

Fitting full model:
Iteration 0:   log likelihood = -15827.071

Weibull regression -- log relative-hazard form

No. of subjects =        10473                     Number of obs   =     10473
No. of failures =         9145
Time at risk    =       115866
                                                   LR chi2(0)      =     -0.00
Log likelihood  =   -15827.071                     Prob > chi2     =         .

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
-------------+----------------------------------------------------------------
       /ln_p |  -.0309431   .0078996    -3.92   0.000    -.0464261   -.0154601
-------------+----------------------------------------------------------------
           p |   .9695307   .0076589                      .9546351    .9846588
         1/p |   1.031427   .0081479                       1.01558    1.047521
------------------------------------------------------------------------------
```

de igual forma que para la función exponencial podemos obtener la función de sobrevivencia y hazard rate a través del comando predict:

```
predict h, hazard
predict S, surv
```

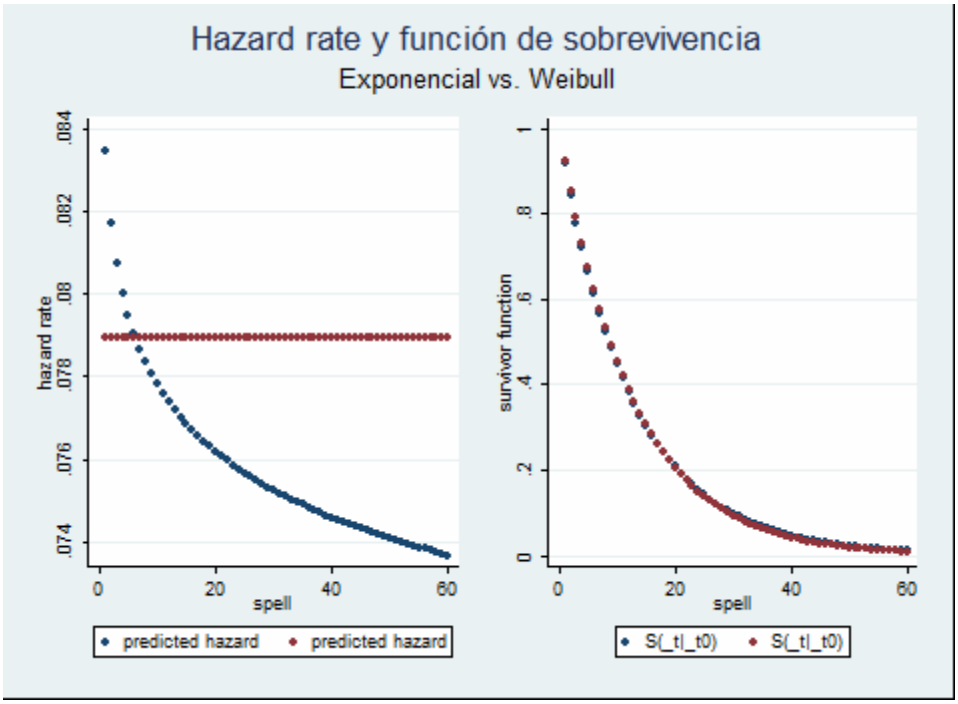
Y luego graficamos ambas funciones:

```
twoway (scatter h spell), name(weibull_h) ytitle(hazard rate) xtitle(spell)
legend(title(weibull))
twoway (scatter S spell), name(weibull_s) ytitle(survivor function)
xtitle(spell) legend(title(weibull))
graph combine weibull_h weibull_s, title(Hazard rate y función de
sobrevivencia) subtitle(Función de Distribución Weibull)
```



El siguiente gráfico permite comparar ambas funciones:

```
twoway (scatter h spell) (scatter h_e spell), name(ambos_h) ytitle(hazard rate)
xtitle(spell)

twoway (scatter S spell) (scatter S_e spell), name(ambos_s) ytitle(survivor
function) xtitle(spell)

graph combine ambos_h ambos_s, title(Hazard rate y función de sobrevivencia)
subtitle(Exponencial vs. Weibull)
```



VIII.4. Hazard models

Hasta el momento hemos estimado los parámetros de la función hazard que nos permiten ver como va disminuyendo la probabilidad de permanecer en cierto estado a medida que pasa el tiempo, pero en muchos de los problemas económicos que estamos interesados de modelar con este tipo de modelos, lo único relevante no es el tiempo. Por ejemplo, la probabilidad de dejar de estar desempleado cambia con el tiempo, pero el tiempo en sí mimo no es una variable fundamental para explicar del desempleo. A las estimaciones de la sección anterior debemos incorporar otras variables explicativas que sean relevantes en determinar la probabilidad de permanecer desempleado, o que expliquen la tasa a la cual se puede dejar el desempleo.

Para incorporar otras variables explicativas, a parte del tiempo, se utilizan los **modelos de hazard proporcionales** (proporcional hazard models). Estos modelos tienen como "variable dependiente" la hazard rate (tasa a la cual el riesgo de dejar cierto estado es acumulado), y como variables explicativas el tiempo y un conjunto de variables de control. El efecto del tiempo es capturado por un componente común para todos los individuos que se llama **baseline hazard**, y se denota por $h_0(t)$. Así, un individuo i en frente el riesgo que todos enfrentan pero modificado por las variables explicativas X_i. Estos modelos entonces tienen la siguiente forma funcional:

$$h_i(t) = h_0(t) \cdot \exp(\beta_0 + \beta_1 X_{1,i})$$

Entonces este modelo es proporcional en el sentido que el riesgo enfrentado cada individuo es multiplicado proporcionalmente por el factor baseline hazard, la función exponencial en la que esta evaluado $X\beta$ se ha escogido arbitrariamente para evitar el problema de que den número negativos.

Entonces al igual que cuando no incluíamos tras variables explicativas, la función $h_0(t)$ debe ser parametrizada a través de alguna forma funcional: exponencial, weibull, entre otras. En este caso el modelo se estima de la siguiente forma:

```
. streg edad_ini sexo esc04, dist(weibull)

         failure _d:  censura
   analysis time _t:  spell

Fitting constant-only model:

Iteration 0:   log likelihood = -15746.448
Iteration 1:   log likelihood = -15739.268
Iteration 2:   log likelihood = -15739.268

Fitting full model:

Iteration 0:   log likelihood = -15739.268
Iteration 1:   log likelihood = -15465.357
Iteration 2:   log likelihood = -15461.097
Iteration 3:   log likelihood = -15461.096

Weibull regression -- log relative-hazard form

No. of subjects =        10422                     Number of obs   =     10422
No. of failures =         9106
Time at risk    =       115183
                                                   LR chi2(3)      =    556.34
Log likelihood  =   -15461.096                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    edad_ini |   .9825537   .0010417   -16.60   0.000     .9805141    .9845975
        sexo |   1.414618   .0305577    16.06   0.000     1.355976    1.475796
       esc04 |   .9920643   .0029087    -2.72   0.007     .9863798    .9977816
-------------+----------------------------------------------------------------
       /ln_p |   -.009674   .0078789    -1.23   0.220    -.0251164    .0057684
-------------+----------------------------------------------------------------
           p |    .9903726   .0078031                     .9751964    1.005785
         1/p |   1.009721   .0079555                     .9942482    1.025434
------------------------------------------------------------------------------
```

Este output nos entrega como resultados los Hazard ratios, pero ¿cómo se interpretan estos resultados?

Los valores mostrados en este output corresponden a los Hazard ratios o a la exponencial del coeficiente asociado a la variable explicativa. Por ejemplo, para la variable esc04 nos dice que con un año adicional de escolaridad el Hazard o o como se acumula el riesgo de falla es 0.99 veces el Hazard de una persona con un año de escolaridad menos. Así, lo que muestran estos valores es lo siguiente:

$$h(t \mid esc04 = e+1, edad\_ini, sexo) = h_0(t) \cdot \exp(\beta_0) \cdot \exp(\beta_1 edad\_ini + \beta_2 sexo + \beta_3 (e+1))$$
$$h(t \mid esc04 = e, edad\_ini, sexo) = h_0(t) \cdot \exp(\beta_0) \cdot \exp(\beta_1 edad\_ini + \beta_2 sexo + \beta_3 e)$$

Luego tomando el ratio de las funciones Hazard:

$$\frac{h(t \mid esc04 = e+1, edad\_ini, sexo)}{h(t \mid esc04 = e, edad\_ini, sexo)} = \frac{h_0(t) \cdot \exp(\beta_0) \cdot \exp(\beta_1 edad\_ini + \beta_2 sexo) \cdot \exp(\beta_3 (e+1))}{h_0(t) \cdot \exp(\beta_0) \cdot \exp(\beta_1 edad\_ini + \beta_2 sexo) \cdot \exp(\beta_3 e)}$$
$$\frac{h(t \mid esc04 = e+1, edad\_ini, sexo)}{h(t \mid esc04 = e, edad\_ini, sexo)} = \exp(\beta_3)$$
$$h(t \mid esc04 = e+1, edad\_ini, sexo) = \exp(\beta_3) \cdot h(t \mid esc04 = e, edad\_ini, sexo)$$

Utilizando los valores mostrados en el output anterior:

$$h(t \mid esc04 = e+1, edad\_ini, sexo) = \underbrace{\exp(\beta_3)}_{0.99} \cdot h(t \mid esc04 = e, edad\_ini, sexo)$$

Nos dice que el Hazard, por ejemplo, de una persona con 10 años de escolaridad es 0.99 veces el Hazard de una persona con 9 años de escolaridad, es decir, cada año adicional de escolaridad disminuye la tasa a la cual se sale del estado desempleo. Por otra parte, los hombres tienen una Hazard rate equivalente a 1.4 veces la Hazard rate los las mujeres, es decir, la tasa a la cual los hombres acumulan riesgo de falla (posibilidad de salir del desempleo) es mayor (en un 40%) a la de las mujeres. Finalmente, mientras mayor es la edad a la cual la persona inicia el estado de desempleo la Hazard rate es menor, un año adicional de edad tiene un Hazard rate 0.98 a la Hazard rate de una persona que inicio el desempleo con un año menos, significando que a mayor edad a la cual la persona inicia el desempleo menor es la tasa a la cual puede salir de ese estado.

Por otra parte, si queremos que el output muestre los coeficientes estimados, se puede utilizar la siguiente opción:

```
. streg edad_ini sexo esc04, dist(weibull) nohr

        failure _d:  censura
  analysis time _t:  spell

Fitting constant-only model:

Iteration 0:   log likelihood = -15746.448
Iteration 1:   log likelihood = -15739.268
Iteration 2:   log likelihood = -15739.268

Fitting full model:

Iteration 0:   log likelihood = -15739.268
Iteration 1:   log likelihood = -15465.357
Iteration 2:   log likelihood = -15461.097
Iteration 3:   log likelihood = -15461.096

Weibull regression -- log relative-hazard form

No. of subjects =        10422                 Number of obs   =     10422
No. of failures =         9106
Time at risk    =       115183
                                               LR chi2(3)      =    556.34
Log likelihood  =   -15461.096                 Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    edad_ini |  -.0176003   .0010602   -16.60   0.000    -.0196782   -.0155224
        sexo |   .3468595   .0216014    16.06   0.000     .3045215    .3891975
       esc04 |  -.0079673   .0029319    -2.72   0.007    -.0137138   -.0022208
       _cons |  -2.058246   .0577973   -35.61   0.000    -2.171526   -1.944965
-------------+----------------------------------------------------------------
       /ln_p |   -.009674   .0078789    -1.23   0.220    -.0251164    .0057684
-------------+----------------------------------------------------------------
           p |   .9903726   .0078031                      .9751964    1.005785
         1/p |   1.009721   .0079555                      .9942482    1.025434
------------------------------------------------------------------------------
```

Con esto podemos calcular el componente proporcional en la Hazard rate, por ejemplo después de 10 meses desempleado:

```
. di e(aux_p)*10^(e(aux_p)-1)*exp(_b[_cons])
.12367632
```

```
. di e(aux_p)*20^(e(aux_p)-1)*exp(_b[_cons])
.12285376

. di e(aux_p)*30^(e(aux_p)-1)*exp(_b[_cons])
.12237512

. di e(aux_p)*40^(e(aux_p)-1)*exp(_b[_cons])
.12203666
```

Por otra parte, utilizando el comando mfx podemos obtener los efectos marginales de las variables explicativas sobre la duración mediana o la duración promedio.

```
. mfx

Marginal effects after streg
      y  = predicted median _t (predict)
         =  8.6414658
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
edad_ini |   .1535708      .00953   16.12   0.000   .134895  .172247    32.211
    sexo*|  -3.123529      .20494  -15.24   0.000  -3.52521 -2.72185   .575513
   esc04 |   .0695187      .02561    2.71   0.007   .019316  .119722   9.63644
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1

. mfx, predict(mean)

Marginal effects after streg
      y  = predicted mean _t (predict, mean)
         =  12.563406
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
edad_ini |   .2232691      .01391   16.06   0.000   .196014  .250525    32.211
    sexo*|  -4.541147      .29799  -15.24   0.000  -5.12519  -3.9571   .575513
   esc04 |   .1010698      .03725    2.71   0.007   .028063  .174077   9.63644
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Luego, de los efectos marginales podemos concluir que la edad que la persona tenía al inicial la etapa de desempleo aumenta el tiempo mediano y medio de sobrevivencia, que en este caso es el tiempo mediano antes de dejar el desempleo, disminuye al pasar de ser mujer a ser hombre y aumenta con la escolaridad.

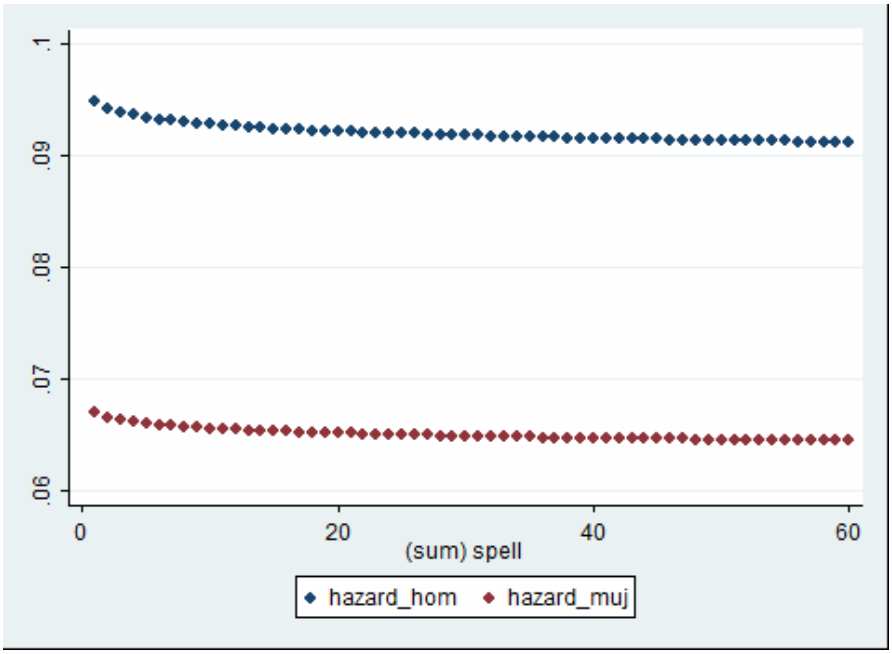
Stata no permite obtener efectos marginales de la hazard rate, estos pueden ser computados en forma manual:

Como cambia el riesgo (hazard rate) cuando se pasa de ser mujer (sexo=0) a ser hombre (sexo=1), podemos ver que disminuye la tasa a la cual se acumula riesgo.

```
di                                                          e(aux_p)*20^(e(aux_p)-
1)*exp(_b[_cons]+_b[edad_ini]*32+_b[esc04]*9+_b[sexo])-e(aux_p)*20^(e(aux_p)-
1)*exp(_b[_cons]+_b[edad_ini]*32+_b[esc04]*9)
.02699581
```

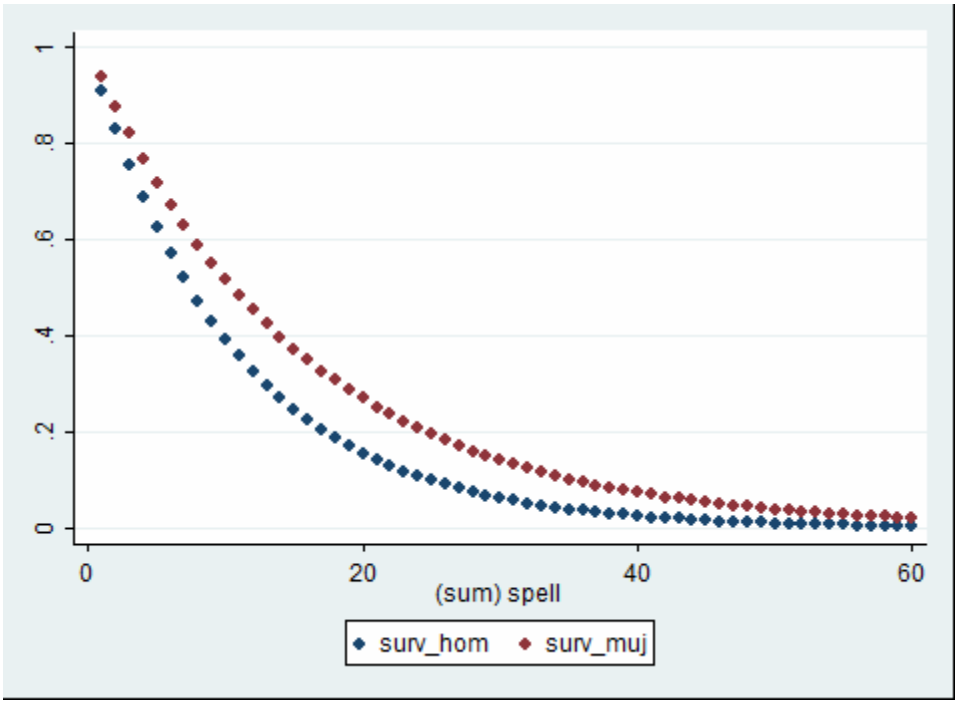
Podemos graficar ambas hazar rate, para hombres y mujeres, asumiendo un valor fijo de edad de inicio del desempleo y años de escolaridad.

```
g hazard_hom=e(aux_p)*spell^(e(aux_p)-
1)*exp(_b[_cons]+_b[edad_ini]*32+_b[esc04]*9+_b[sexo])

g hazard_muj=e(aux_p)*spell^(e(aux_p)-
1)*exp(_b[_cons]+_b[edad_ini]*32+_b[esc04]*9)

twoway (scatter hazard_hom spell) (scatter hazard_muj spell)
```



También se puede obtener la la función de sobrevivencia para cada sexo:

```
g surv_hom=exp(-
exp(_b[_cons]+_b[edad_ini]*32+_b[esc04]*9+_b[sexo])*spell^e(aux_p))

g surv_muj=exp(-exp(_b[_cons]+_b[edad_ini]*32+_b[esc04]*9)*spell^e(aux_p))
twoway (scatter surv_hom spell) (scatter surv_muj spell)
```



Así, viendo ambos gráficos podemos ver que las mujeres acumulan riesgo de fallar a una tasa más baja, su hazard rate esta por debajo de la de los hombres, lo que se refleja en una función de sobrevivencia por arriba de la de los hombres. Por ejemplo, la probabilidad de llegar a 20 meses de desempleo es mayor en las mujeres (aprox. 30%) que en los hombres (cerca 20%), mientras más meses pasan la diferencia va disminuyendo hasta que ambas convergen a cerp.

También podemos computar el efecto marginal sobre hazard rate de los años de escolaridad:

```
di                                                    e(aux_p)*20^(e(aux_p)-
1)*exp(_b[_cons]+_b[edad_ini]*32+_b[esc04]*9+_b[sexo])-e(aux_p)*20^(e(aux_p)-
1)*exp(_b[_cons]+_b[edad_ini]*32+_b[esc04]*8+_b[sexo])
-.00073677
```

Entonces un año adicional de escolaridad disminuye la tasa a la cual se acumula riesgo de fallar, es decir, disminuye la tasa a la cual se sale del estado de desempleo.

Lo mismo podemos hacer con la edad de inicio del desempleo:

```
di                                                    e(aux_p)*20^(e(aux_p)-
1)*exp(_b[_cons]+_b[edad_ini]*32+_b[esc04]*9+_b[sexo])-e(aux_p)*20^(e(aux_p)-
1)*exp(_b[_cons]+_b[edad_ini]*31+_b[esc04]*9+_b[sexo])
-.00163544
```

Mientras mayor es la edad de inicio de desempleo se acumula riesgo de fallar o probabilidad de dejar el desempleo a una menor tasa.

Este modelos hazard proporcional, ya se utilizando la función weibull o cualquiera otra, asume que la hazard rate tiene esta forma funcional rígida. Sin embargo, si desconocemos esta función o simplemente no queremos asumir una forma funcional para la baseline hazard, podemos utilizar el modelo semiparamétrico de Cox. Esta metodología no impone una forma funcional para la baseline hazard ya que no la estima, simplemente asume que esta baseline hazard es la misma para cada individuo. Esto simplemente porque se toma como variable de interés la razón de hazard rate o relative hazard:

$$\frac{h(t \mid x_j)}{h(t \mid x_m)} = \frac{\exp(X_j \beta_j)}{\exp(X_m \beta_m)}$$

```
. stcox edad_ini sexo esc04

         failure _d:  censura
   analysis time _t:  spell

Iteration 0:   log likelihood = -76670.416
Iteration 1:   log likelihood = -76458.443
Iteration 2:   log likelihood = -76458.346
Refining estimates:
Iteration 0:   log likelihood = -76458.346

Cox regression -- Breslow method for ties

No. of subjects =        10422                     Number of obs   =     10422
No. of failures =         9106
Time at risk    =       115183
                                                   LR chi2(3)      =    424.14
Log likelihood  =   -76458.346                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    edad_ini |   .9846358   .0010489   -14.53   0.000     .9825821    .9866938
        sexo |   1.350909   .0291383    13.94   0.000     1.294989    1.409243
       esc04 |   .9938347   .0029107    -2.11   0.035     .9881461     .999556
------------------------------------------------------------------------------
```

Los coeficientes Hazard ratio estimados indican que un año más de edad de inicio del desempleo hace que la persona que tiene más edad tenga un hazard rate o tasa de fallo 0.98 veces la de una persona con una año menos de edad. Los hombres tienen un hazard rate igual a 1.35 el hazard rate de las mujeres, los hombres tienen una tasa a la cual acumulan riesgo o a la cual la probabilidad de dejar el desempleo aumenta mayor que las mujeres. Por último, un año adicional de escolaridad hace que la persona de más escolaridad tenga un hazard rate equivalente a 0.99 veces el hazard rate de una persona con un año menos de escolaridad, es decir, las personas con más escolaridad tienen una tasa de riesgo de fallo o probabilidad de dejar el desempleo menor.

Capítulo IX. Regresión de mediana y cuantiles

IX.1. Definición de la estimación de cuantiles

Cuando estimamos la relación entre una variable de interés, la que hemos llamado variable dependiente, y una o más variables explicativas, por el método de MCO, lo que estamos estimando es la media condicional de la variable dependiente:

$$\hat{E}[Y_i \mid X_i] = \hat{\alpha} + \hat{\beta} X_i$$

Sin embargo, en muchos casos puede que nuestro interés no sea solamente la media de la variable dependiente, sino por ejemplo la mediana o cuantiles de la misma.

En MCO la función que se minimiza es la suma de los errores al cuadrado. En la **regresión de mediana** lo que se minimiza es la suma de los valores absolutos del error:

$$\hat{Med}[Y_i \mid X_i] = \hat{\alpha}_{MED} + \hat{\beta}_{MED} X_i$$

$$\min_{\alpha,\beta} \sum_{i=1}^{N} |\mu_i| \quad \Leftrightarrow \min_{\alpha,\beta} \sum_{i=1}^{N} |Y_i - \alpha - \beta X_i|$$

En la **regresión de cuantiles** se minimiza la siguiente función objetivo:

$$\hat{q}_\tau[Y_i \mid X_i] = \hat{\alpha}_\tau + \hat{\beta}_\tau X_i$$

$$\min_{\alpha_\tau,\beta_\tau} \sum_{i:Y_i \geq \alpha_\tau + \beta_\tau X_i}^{N} \tau \,|Y_i - \alpha_\tau - \beta_\tau X_i| \; + \sum_{i:Y_i < \alpha_\tau + \beta_\tau X_i}^{N} (1-\tau)\,|Y_i - \alpha_\tau - \beta_\tau X_i|$$

Notar que la regresión de mediana es un caso especial de la regresión de cuantiles cuando τ es 0.5.

La ventaja de la regresión de cuantiles es que permite caracterizar de mejor forma los datos, y la regresión de mediana, comparado con de la media, es más robusta frente a la presencia de outliers.

IX.2. Aplicación: Gastos médicos en relación a los gastos totales del hogar

Para la aplicación de los modelos de regresión de cuantiles se utilizarán datos del logaritmo del gasto médicos y el logaritmo de gastos totales del hogar, los datos fueron obtenidos de la encuesta Vietnam Living Standards del Banco Mundial (1997), y consiste en una muestra de 5.006 hogares.

Cuando realizamos la estimación por mínimos cuadrados ordinarios de un modelo de regresión simple entre el logaritmo del gasto médico y el logaritmo del gasto total del hogar, obtenemos el siguiente resultado:

```
. reg  lnmed lntotal, robust

Linear regression                                      Number of obs =    5006
                                                       F(  1,  5004) =  318.05
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0587
                                                       Root MSE      =  1.5458

------------------------------------------------------------------------------
             |               Robust
       lnmed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lntotal |   .5736545   .0321665    17.83   0.000      .510594     .636715
       _cons |   .9352117    .298119     3.14   0.002     .3507677    1.519656
------------------------------------------------------------------------------

predict predmco
```

Podemos apreciar que la estimación MCO de este modelo entrega una elasticidad del gasto médico con respecto al gasto total del hogar de un 0.57. Es decir, un aumento de un 1% en el gasto total del hogar aumenta en un 0.57% el gasto en medicamentos del hogar.

La estimación anterior no considera la heterogeneidad en estas elasticidades que pueden existir en diferentes niveles de ingresos o de gasto total del hogar.

El comando **qreg** de STATA nos permite realizar estimaciones por cuantiles, por ejemplo, a través del siguiente comando podemos estimar una regresión de mediana:

```
. qreg lnmed lntotal, quantile(50)

Median regression                                       Number of obs =      5006
  Raw sum of deviations 6324.265 (about 6.3716121)
  Min sum of deviations 6097.156                        Pseudo R2     =    0.0359

------------------------------------------------------------------------------
       lnmed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lntotal |   .6210917   .0388194    16.00   0.000     .5449886    .6971948
       _cons |   .5921626   .3646869     1.62   0.104    -.1227836    1.307109
------------------------------------------------------------------------------

predict predp50
```

Los intervalos de confianza deben ser obtenidos mediante bootstrap:

```
. bs "qreg lnmed lntotal, quantile(50)" "_b[lntotal]", reps(100)

command:      qreg lnmed lntotal , quantile(50)
statistic:    _bs_1      = _b[lntotal]

Bootstrap statistics                              Number of obs    =      5006
                                                  Replications     =       100

------------------------------------------------------------------------------
Variable     |  Reps  Observed      Bias  Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _bs_1 |   100  .6210917  .0030688  .0451547    .531495   .7106884   (N)
             |                                      .5483162   .7212729   (P)
             |                                      .5522696    .732618  (BC)
------------------------------------------------------------------------------
Note:  N   = normal
       P   = percentile
       BC  = bias-corrected
```

Podemos apreciar que la elasticidad el gasto en médico con respecto al gasto total del hogar es de 0.62 mayor que para el promedio, y es estadísticamente significativo.

También podemos estimar un coeficiente de la elasticidad para el quantil 0.25, lo que ser haría de la siguiente forma:

```
. qreg lnmed lntotal, quantile(25)

.25 Quantile regression                           Number of obs =      5006
  Raw sum of deviations 5162.186 (about 5.2729998)
  Min sum of deviations  5085.31                  Pseudo R2     =    0.0149

------------------------------------------------------------------------------
       lnmed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lntotal |   .4034632   .0454632     8.87   0.000     .3143353     .492591
       _cons |   1.567055   .4271577     3.67   0.000     .7296387    2.404471
------------------------------------------------------------------------------

predict predp25
```

```
. bs "qreg lnmed lntotal, quantile(25)" "_b[lntotal]", reps(100)

command:      qreg lnmed lntotal , quantile(25)
statistic:    _bs_1      = _b[lntotal]

Bootstrap statistics                                    Number of obs    =      5006
                                                        Replications     =       100

------------------------------------------------------------------------------
Variable     |  Reps  Observed      Bias  Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _bs_1 |   100  .4034632  .0057385  .0397641   .3245625   .4823639   (N)
             |                                      .3165871   .4944938   (P)
             |                                      .3165871   .4944938  (BC)
------------------------------------------------------------------------------
Note:  N   = normal
       P   = percentile
       BC  = bias-corrected
```

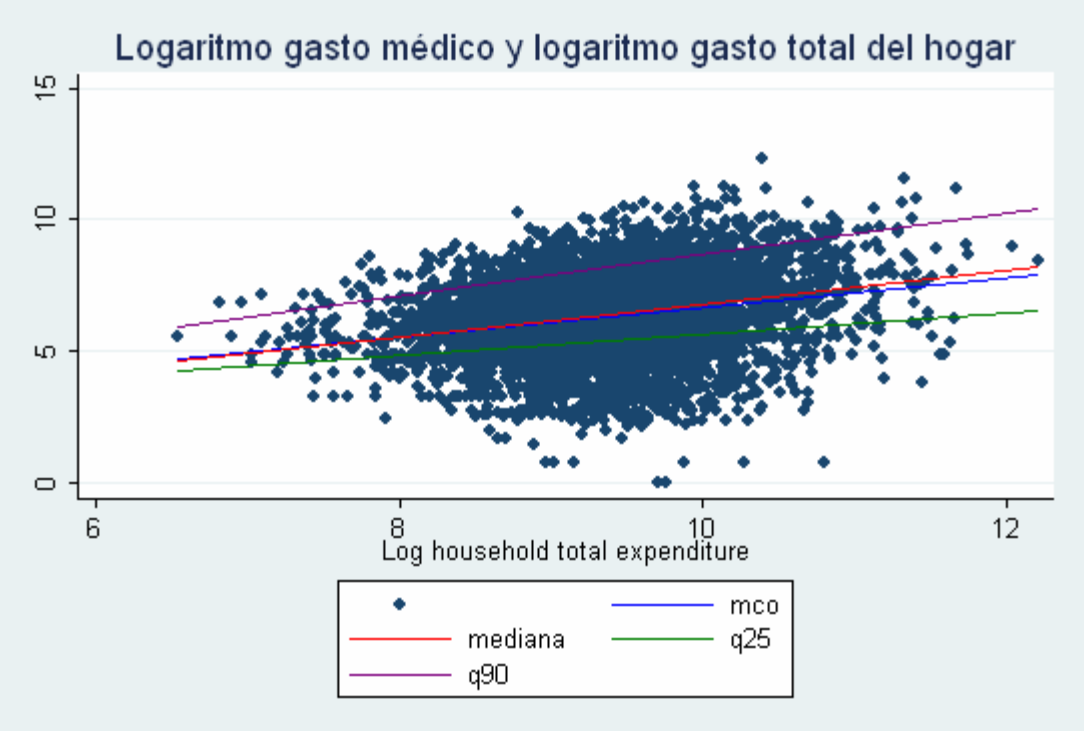
Y para el percentile 90:

```
. qreg lnmed lntotal, quantile(90)

.9 Quantile regression                                  Number of obs =      5006
  Raw sum of deviations 2687.692 (about 8.2789364)
  Min sum of deviations 2505.131                        Pseudo R2     =    0.0679

------------------------------------------------------------------------------
       lnmed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lntotal |   .8003569   .0517226    15.47   0.000      .698958    .9017558
       _cons |   .6750967   .4857565     1.39   0.165     -.277199    1.627392
------------------------------------------------------------------------------

predict predp90
```

```
. bs "qreg lnmed lntotal, quantile(90)" "_b[lntotal]", reps(100)

command:      qreg lnmed lntotal , quantile(90)
statistic:    _bs_1      = _b[lntotal]

Bootstrap statistics                                    Number of obs    =      5006
                                                        Replications     =       100

------------------------------------------------------------------------------
Variable     |  Reps  Observed      Bias  Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _bs_1 |   100  .8003569 -.0076573  .0418586   .7173004   .8834134   (N)
             |                                      .6942345   .8574172   (P)
             |                                      .7027756   .8593536  (BC)
------------------------------------------------------------------------------
Note:  N   = normal
       P   = percentile
       BC  = bias-corrected
```

El siguiente gráfico muestra la relación lineal estimada entre el logaritmo de gasto médico y el logaritmo del gasto total del hogar, para la media, mediana, quantil 25 y quantil 90:

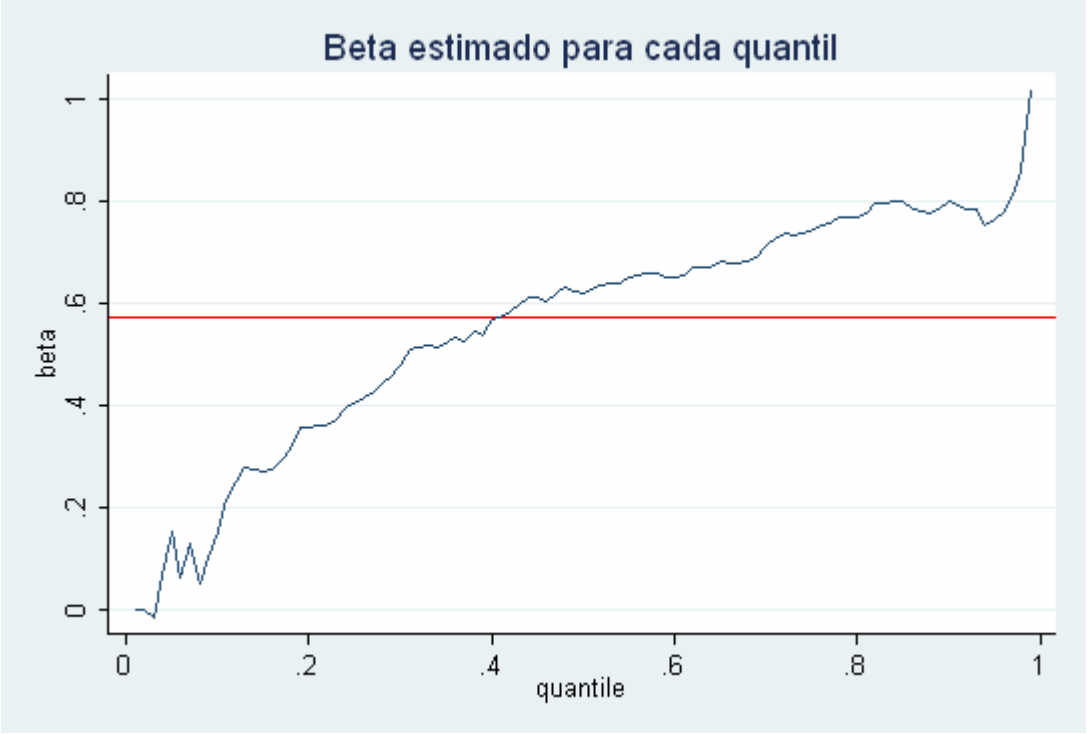
```
twoway (scatter   lnmed lntotal) (lfit predmco lntotal, lcolor(blue)) (lfit
predp50  lntotal,  lcolor(red)) (lfit predp25 lntotal, lcolor(green)) (lfit
predp90 lntotal, lcolor(purple)), legend(on order(1 "" 2  "mco" 3 "mediana" 4
"q25" 5 "q90")) title(Logaritmo gasto médico y logaritmo gasto total del hogar)
```



Luego podríamos estimar un beta para cada cuantil y obtener una relación entre el cuantil y el coeficiente estimado para la elasticidad del gasto en médico con respecto al gasto total del hogar:

```
matrix Q=J(99,2,0)

local i=0.01
while `i'<1{

qreg lnmed lntotal, quantile(`i')
matrix Q[`i'*100,1]=e(q)
matrix Q[`i'*100,2]=_b[lntotal]

local i=`i'+0.01
}

svmat Q, name(quantile)

rename quantile1 quantile
rename quantile2 beta
```

```
twoway (line  beta quantile, msize(vtiny) mstyle(p1) clstyle(p1)),
yline(.5736545, lcolor(red)) title(Beta estimado para cada quantil)
```



Podemos apreciar que mientras menor es el nivel de gasto en médico del hogar (cuantiles más bajos), menor es la elasticidad del gasto en médico con respecto al gasto total del hogar. La línea roja del gráfico representa la estimación MCO del coeficiente de interés.

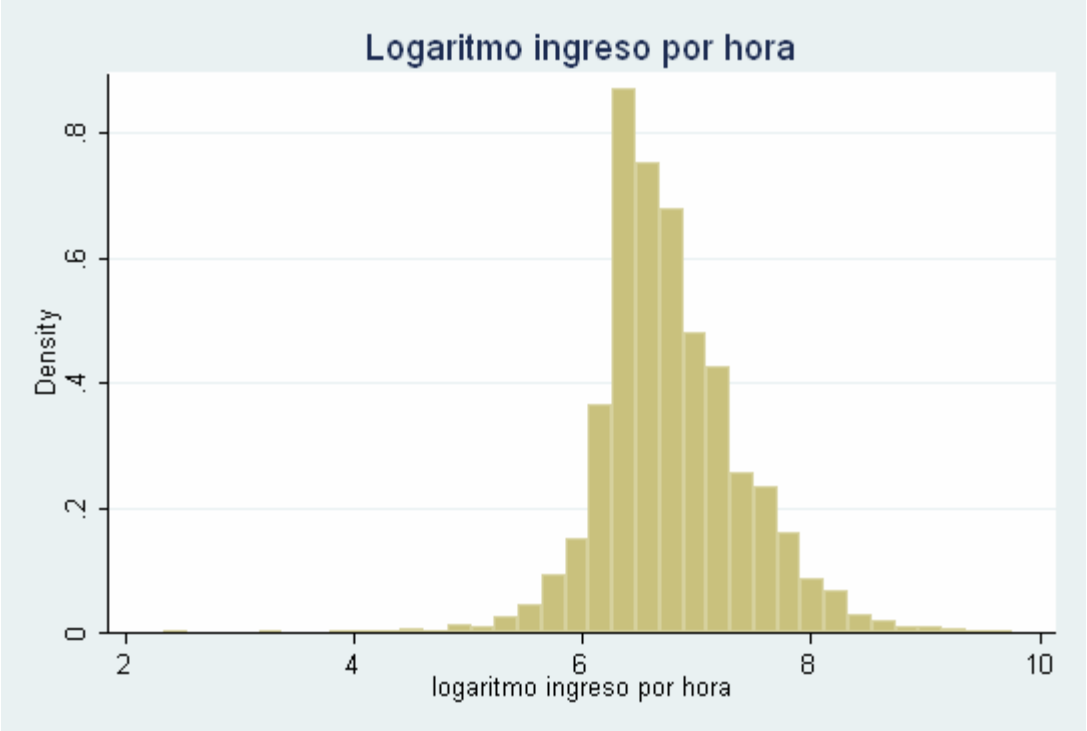
Capítulo X. Métodos no paramétricos y semiparamétricos

En esta sección presentaremos métodos para el análisis de datos que buscan realizar la menor cantidad de supuestos sobre el proceso que genera los datos. Los primeros son los métodos no paramétricos, los que nos permitirán estimar la densidad de una variable. También se verá la regresión no paramétrica, la que sólo se puede realizar en función de una variable explicativa, aunque teóricamente la regresión no paramétrica se puede realizar en función de más de una variable explicativa, en la práctica esto no es factible. Es por esta razón que surgen los métodos semi-paramétricos, en los que por ejemplo no se supone una forma funcional especifica para la relación entre la variable dependiente y explicativa (media, mediana, etc…) sino que se deja que los datos revelen esta función, estimando los parámetros beta que forman parte del argumento de esta relación.

X.1. Estimación no paramétrica de funciones de densidad

La primera aproximación para estimar la densidad de una variable es mediante el **histograma** de la misma, el histograma divide el espacio posible de los valores de la variable en intervalos de igual distancia y calculando la fracción de las observaciones en cada uno de estos intervalos se aproxima la distribución empírica de la variable. Sin embargo, el histograma es una estimación tosca o no suave de la densidad. El siguiente gráfico muestra el histograma del ingreso del salario por hora obtenido de la EPS 2004, para las personas entre 18 y 41 años (Base de Datos Clase 2).

```
histogram lyph, title(Logaritmo ingreso por hora)
```



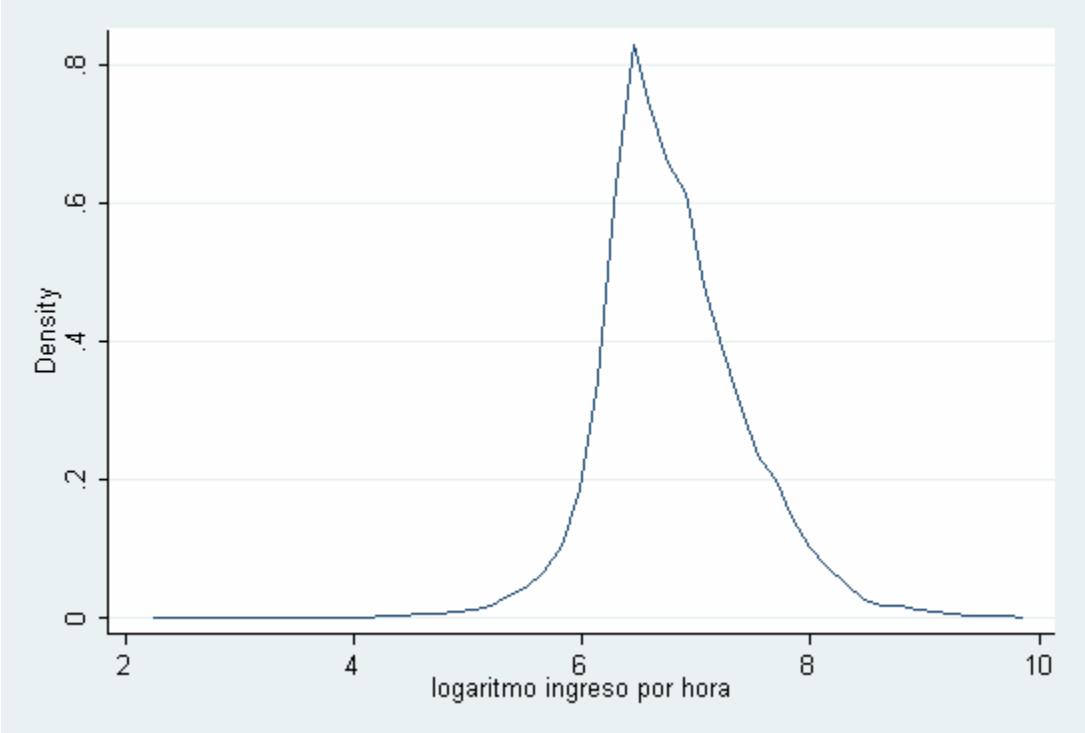
Para obtener una estimación más suave de la función de densidad en vez de tomar intervalos de valores de la variable, se podría tomar cada observación puntual de la variable y darle un peso de 1/N a cada una de estas observaciones, el problema de esta metodología es que no se le asigna probabilidad a los valores de x que no son observados en la muestra. Entonces la alternativa que surge a esto es no darle el peso o probabilidad 1/N al punto x_i sino a la densidad de la variable entorno a x_i. Esto es justamente lo que hace la estimación KERNEL, obtiene la densidad empírica de la variable tomando una combinación de densidades entorno a los puntos observados de la variable:

$$\hat{f}(x_0) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)$$

Donde K() es la llamado función Kernel, y h es el llamado bandwidth. Dentro de las funciones kernel se encuentra el Gaussiano, Epanechnikov, uniforme, entre otros. Se ha demostrado que la función Kernel óptima es la Epanechnikov. Con respecto al parámetro de suavización h, existe una elección óptima que corresponden a aquel que minimiza el error cuadrático medio integrado de la función de densidad.

El comando que en STATA permite estimar la densidad utilizando la función Kernel es **kdensity.** Por ejemplo, para graficar la densidad kernel del logaritmo del ingreso por hora utilizando un kernel gaussiano se debe hacer lo siguiente:

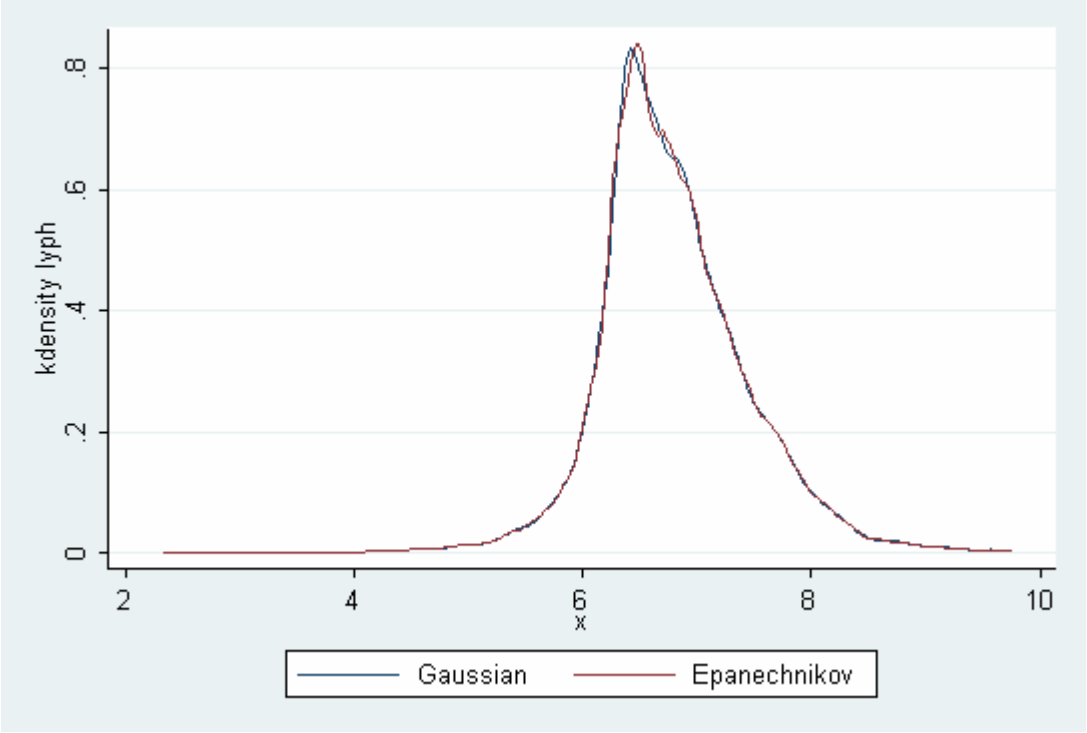
```
kdensity lyph, gaussian generate(estim den)
```



La opción **generate**, genera dos variables `estim` que contiene los puntos de estimación de la densidad kernel y `den` que contiene la densidad estimada para cada uno de estos puntos. En esta estimación se ha utilizado el bandwidth óptimo, que corresponde al default de STATA.

El siguiente gráfico muestra la estimación kernel utilizando la función gaussiana y epanechnikov, en ambas utilizando el bandwidth óptimo que mínimo el error cuadrático medio integrado.

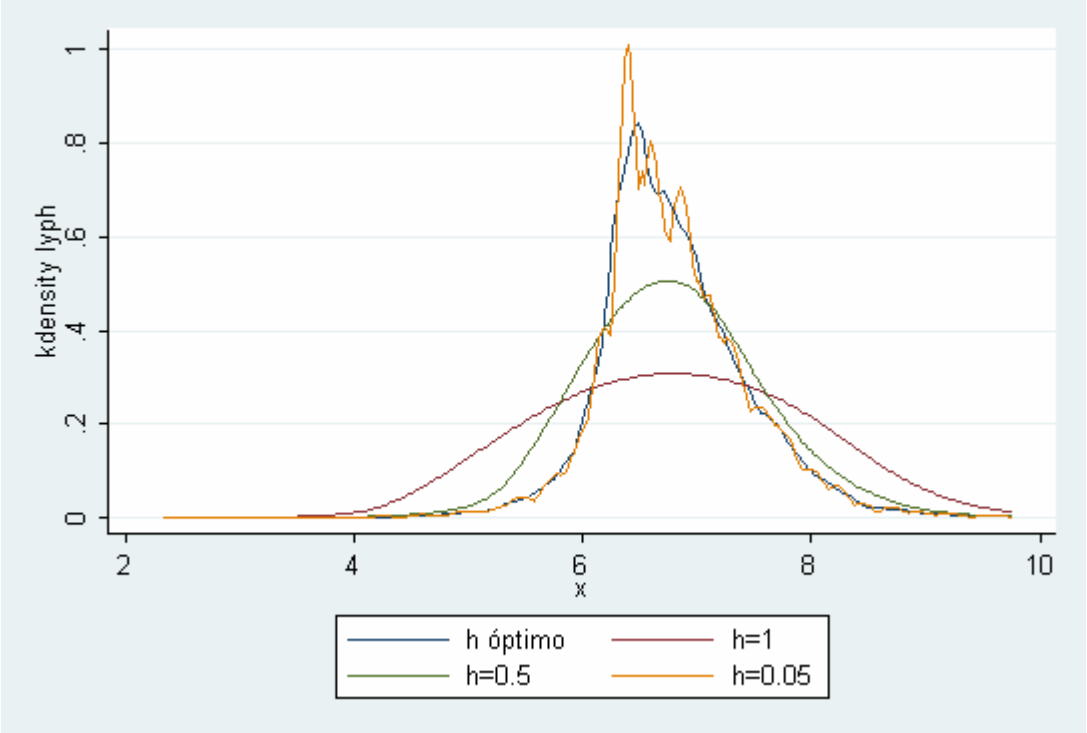
```
twoway (kdensity lyph, gaussian) (kdensity lyph), legend(on order(1
"Gaussian" 2  "Epanechnikov"))
```



Las diferencias entre ambas estimaciones son mínimas. Se menciono que se ha demostrado que el Kernel Epanechnikov ha demostrado ser óptimo pero en las

práctica las ventajas son mínimas. Lo que si puede generar grandes diferencias es la elección del parámetro de suavización, bandwidth. El siguiente gráfico muestra cuatro funciones kernel utilizando el kernel epanechnikov con 4 bandwithds distintos, incluyendo el valor óptimo:

```
twoway (kdensity lyph) (kdensity lyph, width(1)) (kdensity lyph,
width(0.5)) (kdensity lyph, width(0.05)), legend(on order(1 "h óptimo"
2  "h=1" 3 "h=0.5" 4 "h=0.05"))
graph export g6.tif, replace
```



Mientras mayor el bandwidth asumido más se suaviza la función de densidad. Este parámetro representa una especie de desviación estándar de cada una de las densidades que estoy combinando, mientras mayor es el parámetro más desviación estándar tienen las densidades ponderadas lo que suaviza la función de densidad final obtenida.

X.2. Estimación no paramétrica de la relación entre dos variables: Nonparametric local regresión

Consideremos la regresión entre la variable dependiente y sobre la variable explicativa x. El modelo de regresión, sin asumir una forma funcional especifica para la relación entre ambas variables, es el siguiente:

$$y_i = m(x_i) + \varepsilon_i \quad i = 1,...,N$$
$$\varepsilon_i \sim iid(0, \sigma_\varepsilon^2)$$

Donde la forma funcional m() no ha sido especificada.

El método general denominado local weighted average estimator toma la siguiente forma:

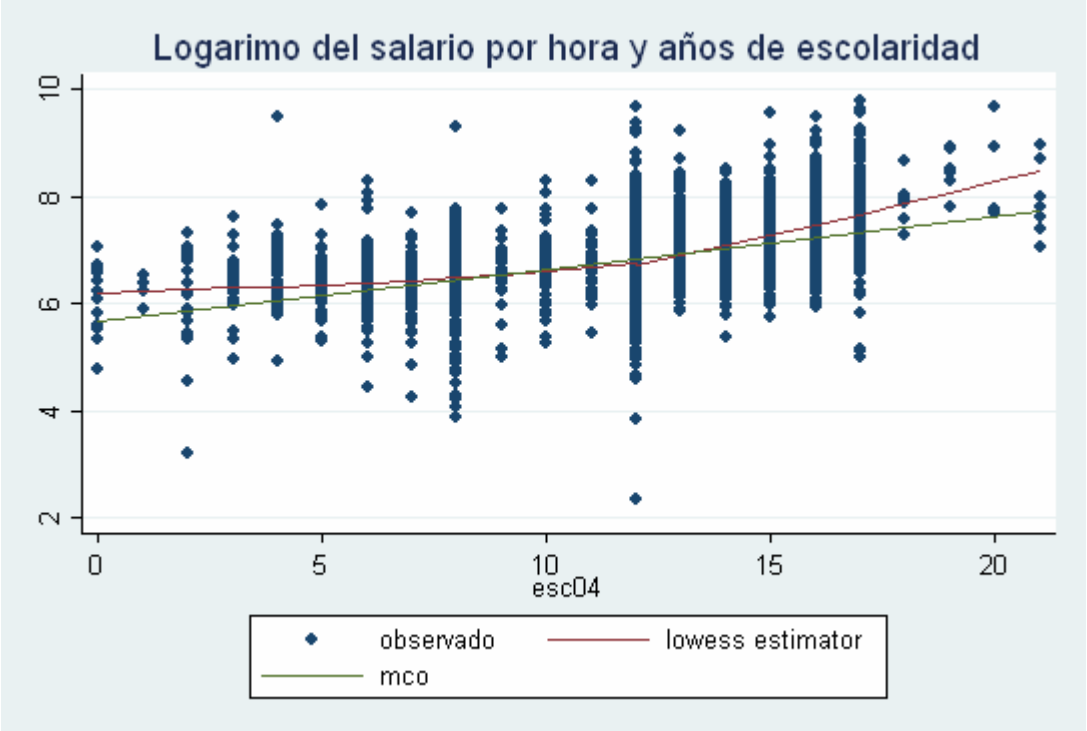
$$\hat{m}(x_0) = \sum_{i=1}^{N} \omega_{i0,h} y_i$$
$$\omega_{i0,h} = \omega(x_i, x_0, h) \;\; y \;\; \sum_{i=1}^{N} \omega_{i0,h} = 1$$

De esta forma, para cada punto x_0 que observamos se obtiene la relación estimada con la variable dependiente como el promedio ponderado de la variable dependiente, donde el ponderador depende de cuan cerca esta la observación de x_i a x_0. El estimador Lowess utiliza la función kernel como ponderador. De esta forma, el Lowess Estimator minimiza la siguiente función objetivo:

$$\min_{m_0} \sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)(y_i - m_0)^2$$

El comando en STATA que permite realizar esta estimación es **lowess**, el siguiente gráfico muestra la estimación no paramétrica de la relación entre logaritmo del salario por hora y los años de escolaridad, y la estimación paramétrica MCO de la misma relación.

```
twoway (scatter lyph esc04) (lowess lyph esc04) (lfit lyph esc04),
legend(on order(1 "observado" 2  "lowess estimator" 3 "mco"))
title(Logarimo del salario por hora y años de escolaridad)
```



Otros estimadores no paramétricos son el estimador de Nadaraya-Watson (kernreg) y el estimador del vecino más cercano (knnreg), lo que cambia son las definiciones del ponderador.

X.3. Modelos semiparamétricos

El problema de los modelos no paramétricos vistos en la sub-sección anterior es no son manejables con más de una variable explicativa. Por eso surgen los modelos semi-paramétricos, donde una parte se deja libre de supuestos, y otra parte se parametriza. Dos ejemplos de modelos semi-paramétricos son:

Modelo parcialmente lineal:

$$E[Y_i \mid X, Z] = \underbrace{X\beta}_{parametrico} + \underbrace{\lambda(Z)}_{no\ parametrico}$$

Modelos Single Index:

$$E[Y_i \mid X] = g(X\beta)$$
$$\beta : parametrico$$
$$g() : no\ parametrico$$

Lamentablemente estas técnica de estimación aún no están programadas en STATA, quienes estén interesados en más detalles acerca de estas técnicas ver sección 9.7 del libro "Microeconometrics: Methods and Applications" de A. Colin Cameron y Pravin K. Trivedi.

X.4. Estimación de la función Hazard en Modelo de Cox

Recordemos que en el modelo proporcional de duración de Cox, la función Hazard tiene un componente común que indica como cambia el riesgo de falla a medida como función del tiempo, lo que se denomina baseline Hazard y otro componente que depende de las características individuales:

$$h(t \mid x_j) = h_0(t)\exp(\beta_0 + x_j\beta)$$
$$h(t \mid x_j) = \underbrace{h_0(t)\exp(\beta_0)}_{\tilde{h}_0(t)}\exp(x_j\beta)$$

Para la estimación del modelo (recordemos) no se necesita hacer ningún supuesto sobre $\tilde{h}_0(t)$, ya que se estiman los Hazard ratios. Pero si queremos analizar como el riesgo de fallar cambia por efecto del tiempo debemos estimar esta función. Los métodos paramétricos asumen cierta distrubción de probabilidad: exponencial, weibull, entre otras. Sin embargo, se podría estimar en forma no paramétrica mediante kernels:

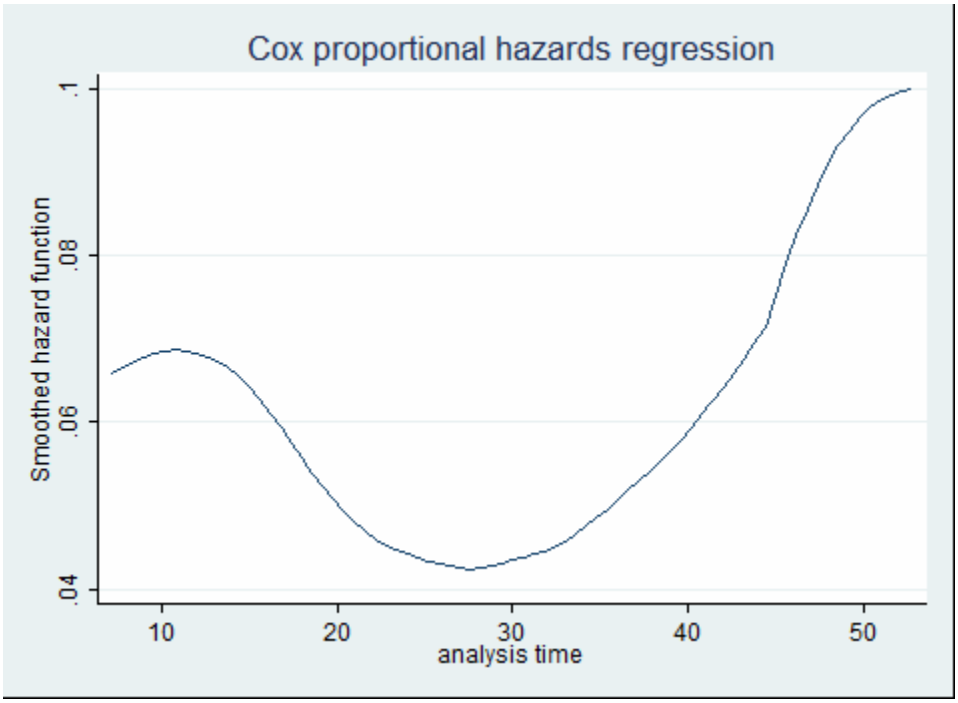
$$\hat{h}_0(t) = \frac{1}{h}K\left(\frac{t - t_j}{h}\right)\hat{h}_{t_j}$$

La forma de estimar la baseline Hazard en STATA en forma no paramétrica es la siguiente:

```
. stcox edad_ini sexo esc04, basehc(h0)

         failure _d:  censura
   analysis time _t:  spell

Iteration 0:   log likelihood = -76670.416
Iteration 1:   log likelihood = -76458.443
Iteration 2:   log likelihood = -76458.346
Refining estimates:
Iteration 0:   log likelihood = -76458.346

Cox regression -- Breslow method for ties

No. of subjects =        10422                     Number of obs   =     10422
No. of failures =         9106
Time at risk    =       115183
                                                   LR chi2(3)      =    424.14
Log likelihood  =   -76458.346                     Prob > chi2     =    0.0000
```

```
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    edad_ini |   .9846358   .0010489   -14.53   0.000     .9825821    .9866938
        sexo |   1.350909   .0291383    13.94   0.000     1.294989    1.409243
       esc04 |   .9938347   .0029107    -2.11   0.035     .9881461     .999556
------------------------------------------------------------------------------
```

```
. stcurve, hazard
```



Capítulo XI. Modelo de datos de conteo

XI.1. Introducción

En muchos contextos económicos la variable dependiente toma sólo valores enteros positivos, es decir, corresponde a una cuanta o conteo de algo y esto es lo que queremos explicar en función de algunas variables explicativas. Cuando la variable dependiente tiene estas características no es apropiado utilizar el modelo de regresión lineal (MCO), este tipo de modelo, al igual que los modelos probit y logit, son no lineales, por lo cual la forma correcta de estimar este tipo de modelos es por Máxima Verosimilitud.

Algunos ejemplos de modelo de conteo son:

- Estudios de fertilidad: se estudia el número de nacimientos y como estos varían en función de la escolaridad de la madre, la edad, y el ingreso del hogar.
- Estudio del número de accidentes de una aerolinea como medida de seguridad de la aerolinea, que puede ser explicado por los beneficios de la empresa y la salud financiera de la misma.
- Estudios de demanda recreacional, que modelan el número de viajes a lugares recreacionales.
- Estudios de demanda por salud, donde se trata de modelar el número de veces que los individuos demandan serivicios de salud como número de visitas al doctor o número de días en el hospital.

La distribución Poisson es para variables discretas no negativas, de esta forma, podemos asumir que la variable dependiente tiene este tipo de distribución para plantear la función de verosimilitud. Entonces la variable dependiente tiene distribución Poisson, podemos escribir su función de masa de proabilidad como:

$$\Pr[Y = y] = \frac{e^{-\mu}\mu^{y}}{y!} \quad y = 0,1,2,...$$

Donde μ es lo que se denomina intensidad. Notemos que E[Y]=μ y V[Y]=μ, es decir, se tiene **equidispersion** o igual media y varianza.

Luego el **Modelo de Regresión Poisson** es derivado de la distribución Poisson parametrizando la relación entre el parámetro de media m y las variables explicativas. El supuesto estándar es usar la siguiente parametrización:

$$E[y_i \mid x_i] = \mu_i = \exp(x_i'\beta)$$

Debido a que,

$$V[y_i \mid x_i] = \mu_i = \exp(x_i'\beta)$$

El modelo de regresión poisson es intrínsicamente heterocedastico.

La función de verosimilitud a optimizar en este caso es:

$$\max_{\beta} \ \ln L(\beta) = \sum_{i=1}^{N} [y_i x_i'\beta - \exp(x_i'\beta) - \ln(y_i!)]$$

Lo que nos interesa en este caso, al igual que en todos los modelos no lineales, no son los coeficientes β estimados, sino los efectos marginales:

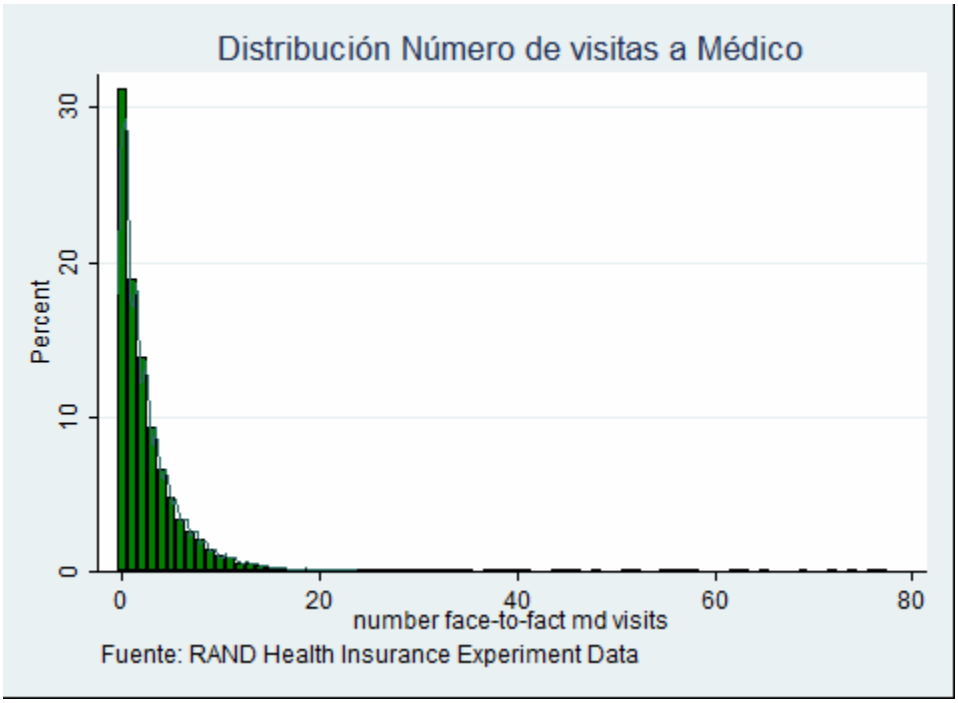
$$\frac{\partial E[y \mid x]}{\partial x_k} = \exp(x'\beta)\beta_k$$

XI.3. Aplicación: Número de visitas al Médico

Para esta aplicación se utiliza la base de datos del RAND Experimento de Seguros de Salud (RAND Health Insurance Experiment) utilizada por Deb y Trivendi (2002). El experimento conducido por la Coporación RAND entre los años 1974 y 1982, ha sido el experimento social controlado más grande en el área de la investigación en seguros de salud. El objetivo principal del experimento era evaluar como el uso de los servicios de salud por parte de los pacientes se ve afectado por los tipos de seguros medicos, los cuales fueron asignados aleatoriamente. En el experimento los datos fueron recolectados para cerca de 8.000 personas en 2.823 familias. Cada familia fue suscrita a uno de los 14 diferentes planes de salud por 3 o 5 años. Los planes van desde servicio libre hasta 95% de cobertura bajo cierto nivel de gasto (con un tope).

El siguiente gráfico muestra un histograma con el número de visitas al médico, podemos ver que poco más del 30% realiza cero visitas al año al médico, y cerca de un 18% realiza una visita al año.

```
histogram MDU, discrete percent fcolor(green) lcolor(black) kdensity
title(Distribución Número de visitas a Médico) caption(Fuente: RAND
Health Insurance Experiment Data)
```



La siguiente tabla muestra las principales estadísticas de cada una de laa variables que serán utilizadas como factores determinantes en la cantidad de visitas al médico realizadas al año. La variable BLACK toma valor 1 si el jefe de hogar es de raza negra, la variable AGE corresponde a la edad en años, FEMALE toma valor 1 si la persona es mujer, EDUCDEC representa los años de educación del jefe de hogar, MDU es la variable que queremos explicar (variable dependiente) que mide el número de visitas ambulatorias a un médico, NDISEASE es el número de enfermedades crónicas, PHYSLIM toma valor 1 si la persona tiene limitaciones físicas, CHILD toma valor 1 si la persona tiene menos de 18 años, FEMCHILD corresponde a la interacción de la Dummy FEMALE y la Dummy CHILD, LFAM es el logaritmo del tamaño familiar, LPI es el logaritmo del pago anual de incentivo por participación,

IDP si el plan tiene deducible, LC es el logaritmo del copago, FMDE es el logaritmo del tope de cobertura sobre 0.01 el copago, HLTHG es 1 si declara que su estado de salud es bueno, HLTHF es 1 si declara su estado de salud regular, HLTHP si declara estado de salud malo, y LINC es el logaritmo del ingreso familiar.

Variable	Obs	Mean	Std. Dev.	Min	Max
BLACK	20186	.1815343	.3827365	0	1
AGE	20186	25.71844	16.76759	0	64.27515
FEMALE	20186	.5169424	.4997252	0	1
EDUCDEC	20186	11.96681	2.806255	0	25
MDU	20186	2.860696	4.504765	0	77
NDISEASE	20186	11.2445	6.741647	0	58.6
PHYSLIM	20186	.1235247	.3220437	0	1
CHILD	20186	.4014168	.4901972	0	1
FEMCHILD	20186	.1937481	.3952436	0	1
LFAM	20186	1.248404	.5390681	0	2.639057
LPI	20186	4.708827	2.697293	0	7.163699
IDP	20186	.2599822	.4386354	0	1
LC	20186	2.383588	2.041713	0	4.564348
FMDE	20186	4.030322	3.471234	0	8.294049
HLTHG	20186	.3620826	.4806144	0	1
HLTHF	20186	.0772813	.2670439	0	1
HLTHP	20186	.0149609	.1213992	0	1
LINC	20186	8.708167	1.22841	0	10.28324

El commando en STATA para estimar un modelo de regression poisson es **poisson**, la siguiente tabla muestra el resultado de este commando para el modelo que buscar explicar el número de veces que la persona va al medico al año en función de las caracteristicas de los planes de salud y características familiares.

```
. poisson MDU LC IDP LPI FMDE LINC LFAM AGE FEMALE CHILD FEMCHILD BLACK EDUCDEC
PHYSLIM NDISEASE HLTHG HLTHF HLTHP

Iteration 0:   log likelihood = -60097.599
Iteration 1:   log likelihood = -60087.636
Iteration 2:   log likelihood = -60087.622
Iteration 3:   log likelihood = -60087.622

Poisson regression                                      Number of obs   =      20186
                                                        LR chi2(17)     =   13106.07
                                                        Prob > chi2     =     0.0000
Log likelihood = -60087.622                             Pseudo R2       =     0.0983

------------------------------------------------------------------------------
         MDU |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          LC |  -.0427332   .0060785    -7.03   0.000    -.0546469   -.0308195
         IDP |  -.1613169   .0116218   -13.88   0.000    -.1840952   -.1385385
         LPI |   .0128511   .0018362     7.00   0.000     .0092523    .0164499
        FMDE |   -.020613   .0035521    -5.80   0.000     -.027575   -.0136511
        LINC |   .0834099   .0051656    16.15   0.000     .0732854    .0935343
        LFAM |  -.1296626   .0089603   -14.47   0.000    -.1472245   -.1121008
         AGE |   .0023756   .0004311     5.51   0.000     .0015306    .0032206
      FEMALE |   .3487667   .0113504    30.73   0.000     .3265203     .371013
       CHILD |   .3361904   .0178194    18.87   0.000     .3012649    .3711158
    FEMCHILD |  -.3625218   .0179396   -20.21   0.000    -.3976827   -.3273608
       BLACK |  -.6800518   .0155484   -43.74   0.000    -.7105262   -.6495775
     EDUCDEC |   .0176149   .0016387    10.75   0.000     .0144031    .0208268
     PHYSLIM |   .2684048   .0123624    21.71   0.000     .2441749    .2926347
    NDISEASE |    .023183   .0006081    38.12   0.000     .0219912    .0243749
       HLTHG |   .0394004   .0095884     4.11   0.000     .0206074    .0581934
       HLTHF |   .2531119    .016212    15.61   0.000     .2213369    .2848869
       HLTHP |   .5216034   .0272382    19.15   0.000     .4682176    .5749892
       _cons |  -.1898766   .0491731    -3.86   0.000    -.2862541    -.093499
------------------------------------------------------------------------------
```

Este output nos muestra como resultado el coeficiente beta estimado, pero al igual que en la mayoría de los modelos no lineales, no estamos interesados en el coeficiente sino en los efectos marginales, y nuevamente el comando que nos permite obtener los efectos marginales es **mfx**:

```
. mfx

Marginal effects after poisson
      y  = predicted number of events (predict)
         =  2.5526948
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.      z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
      LC |  -.1090849      .01551    -7.03   0.000  -.139482 -.078688   2.38359
     IDP*|   -.396583      .02749   -14.43   0.000   -.45046 -.342706   .259982
     LPI |   .0328049      .00469     7.00   0.000    .02362   .04199   4.70883
    FMDE |  -.0526187      .00907    -5.80   0.000  -.070389 -.034848   4.03032
    LINC |   .2129199      .01313    16.22   0.000   .187188  .238652   8.70817
    LFAM |  -.3309892      .02283   -14.50   0.000  -.375737 -.286241    1.2484
     AGE |   .0060642       .0011     5.51   0.000   .003907  .008221   25.7184
  FEMALE*|   .8895421      .02887    30.81   0.000   .832963  .946121   .516942
   CHILD*|   .8912944      .04911    18.15   0.000   .795039   .98755   .401417
FEMCHILD*|  -.8327063      .03698   -22.52   0.000   -.90519 -.760223   .193748
   BLACK |  -1.735965      .03863   -44.94   0.000  -1.81168 -1.66025   .181534
 EDUCDEC |   .0449655      .00418    10.76   0.000   .036773  .053158   11.9668
 PHYSLIM |   .6851556      .03153    21.73   0.000   .623364  .746947   .123525
NDISEASE |   .0591792      .00155    38.26   0.000   .056148  .062211   11.2445
   HLTHG*|   .1011317      .02474     4.09   0.000   .052633   .14963   .362083
   HLTHF*|   .7210038      .05129    14.06   0.000   .620471  .821537   .077281
   HLTHP*|   1.734312      .11537    15.03   0.000   1.50819  1.96044   .014961
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Recoerdemos que en este modelo los errores son heterocedásticos por construcción, y además la varianza depende de los mismos coeficientes estimados, en este modelo la forma apropiada de ver la significancia de los coeficientes es obteniendo los intervalos de confianza mediante bootstrap:

```
. bs "poisson MDU LC IDP LPI FMDE LINC LFAM AGE FEMALE CHILD FEMCHILD BLACK
EDUCDEC PHYSLIM NDISEASE HLTHG HLTHF HLTHP" "_b", reps(100)

Bootstrap statistics                              Number of obs    =     20186
                                                  Replications     =       100

------------------------------------------------------------------------------
Variable     |  Reps  Observed      Bias  Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
        b_LC |   100 -.0427332  .0007178  .0159466  -.0743747  -.0110918   (N)
             |                                     -.0739939  -.0141433   (P)
             |                                     -.0756542  -.0161271  (BC)
       b_IDP |   100 -.1613169  .0014252  .0345362  -.2298443  -.0927895   (N)
             |                                     -.2258377  -.0872168   (P)
```

					-.2100605	-.0798584	(BC)
b_LPI	100	.0128511	.0003724	.0044547	.0040121	.0216901	(N)
					.0059388	.0223223	(P)
					.0030672	.0213983	(BC)
b_FMDE	100	-.020613	-.0008264	.0091856	-.0388392	-.0023868	(N)
					-.0387147	-.0032838	(P)
					-.035657	.0046425	(BC)
b_LINC	100	.0834099	.0018282	.0135125	.0565982	.1102215	(N)
					.0563673	.1088733	(P)
					.052815	.1033517	(BC)
b_LFAM	100	-.1296626	.0009137	.0233243	-.1759431	-.0833822	(N)
					-.1799356	-.0818748	(P)
					-.1793034	-.0752146	(BC)
b_AGE	100	.0023756	-.0001072	.0009951	.0004012	.00435	(N)
					.000192	.0039354	(P)
					.000192	.0039354	(BC)
b_FEMALE	100	.3487667	-.0010908	.0301244	.2889933	.4085401	(N)
					.2870255	.4171301	(P)
					.2999384	.4198574	(BC)
b_CHILD	100	.3361904	.000713	.0354193	.2659109	.4064699	(N)
					.2615154	.3928844	(P)
					.2515882	.3920318	(BC)
b_FEMCHILD	100	-.3625218	-.001857	.0463087	-.4544083	-.2706353	(N)
					-.4669854	-.2924513	(P)
					-.5062459	-.2936182	(BC)
b_BLACK	100	-.6800519	-.0013074	.0341245	-.7477622	-.6123415	(N)
					-.7536808	-.6212551	(P)
					-.7536808	-.6212551	(BC)
b_EDUCDEC	100	.0176149	-.0000167	.0045235	.0086393	.0265905	(N)
					.0069081	.02805	(P)
					.0067609	.0259771	(BC)
b_PHYSLIM	100	.2684048	.00063	.0328184	.2032859	.3335237	(N)
					.2070561	.340555	(P)
					.2070561	.340555	(BC)
b_NDISEASE	100	.023183	.0001385	.001655	.0198992	.0264668	(N)
					.019857	.026076	(P)
					.0193031	.0257115	(BC)
b_HLTHG	100	.0394004	-.0004796	.0223977	-.0050416	.0838424	(N)
					-.0029975	.0867478	(P)
					-.0029975	.0867478	(BC)
b_HLTHF	100	.2531119	.0012493	.038593	.1765349	.3296889	(N)
					.1826011	.3211362	(P)
					.1773578	.3182556	(BC)
b_HLTHP	100	.5216034	.0103248	.0654646	.3917073	.6514995	(N)
					.3883481	.6682352	(P)
					.3856528	.6546603	(BC)
b_cons	100	-.1898766	-.016589	.1332961	-.454365	.0746119	(N)
					-.4715162	.0103337	(P)
					-.4362686	.1106389	(BC)

Note: N = normal
P = percentile
BC = bias-corrected

Capítulo XII. Matching y propensity score

XII.1. Introducción

El objetivo de esta clase es estudiar la técnica de matching y propesity score, técnicas que se utilizan en la evaluación de programas cuando el diseño del mismo no contempla la asignación aleatoria o experimental del grupo de tratamiento y el grupo de control. Estas técnicas nos permiten evaluar el resultado de un programa sin que este haya sido asignado aleatoriamente en la población.

XII.2. Estimación Matching y Propensity Score

XII.2.1. Medidas de impacto del tratamiento

En evaluación de programas el objetivo es medir el impacto de cierto tratamiento T sobre la variable de interés o de resultado, existen varias medidas para medir el impacto de un tratamiento:

i) Comparación antes después (D): se compara la variable de resultado antes y después del tratamiento. El problema de esta medida que captura todos los otros factores en el tiempo que pueden estar afectando la variable de resultado pero que no tienen relación con el tratamiento.

$$D = \underset{T=1}{\overline{Y}_{1,i}} - \underset{T=1}{\overline{Y}_{0,i}}$$

ii) Diferencias en Diferencias (D-D): se tiene que tener un grupo de control asignado aleatoriamente, para garantizar que tengas las mismas características del grupo de tratamiento. Al grupo de control también se le

mide la variable de interés o resultado en dos momentos del tiempo (antes y después). Luego se toma la diferencia entre la diferencia antes-después del grupo de tratamiento y la diferencia antes-después en el grupo de control.

$$D - D = \left(\underset{T=1}{\overline{Y}_{1,i}} - \underset{T=1}{\overline{Y}_{0,i}} \right) - \left(\underset{T=0}{\overline{Y}_{1,i}} - \underset{T=0}{\overline{Y}_{0,i}} \right)$$

Esta segunda medida supone un diseño experimental donde el grupo de tratamiento y control han sido asignados en forma aleatoria dentro de la población objetivo a ser tratada.

Las medidas de efectos causales antes mostradas suponen que este efecto es el mismo para todos los individuos de la población, así se esta suponiendo que la población en homogénea. Sin embargo, en la práctica la población puede ser heterogénea, específicamente, el efecto causal puede variar de un individuo a otro dependiendo de las circunstancias del individuo, su entorno, y otras características. Por ejemplo, en un programa de capacitación sobre habilidades para escribir curriculums probablemente el efecto causal sea mayor en aquellos que no tienen esta habilidad que los que ya poseía al menos algo de esta habilidad. Análogamente, el efecto causal de cierta medicina dependerá de la alimentación, y hábitos de alcohol y fumar del paciente. Entonces cuando existe heterogeneidad, lo que uno trata de estimar es un efecto causal promedio, correspondiente al promedio de los efectos causales individuales.

La primera medida es el Average Treatment Effect (ATE), y la segunda es Average Treatment on the treated (ATET), la primera medida promedio para todos los individuos, y la segunda medida solo en los individuos tratados. El ATE es relevante cuando el tratamiento tiene una aplicabilidad universal.

i) **Average Treatment Effect (ATE):** que consiste en la diferencia en la variable de resultado, entre los tratados y no tratados pero condicional a un set de variables observables x:

$$ATE = E[Y_{1,i} \mid X, T = 1] - E[Y_{0,i} \mid X, T = 0]$$

ii) **Average Treatment on the treated (ATET):** solo se promedia para el grupo de tratados. La fórmula general para el estimador matching ATET es:

$$ATET = E[Y_{1,i} \mid X, T = 1] - E[Y_{0,i} \mid X, T = 1]$$

Cuando no se posee un diseño experimental en la asignación de tratamiento, y no se cuenta con un grupo de control que permita obtener el efecto causal, es necesario buscar técnicas que nos permitan de alguna forma encontrar un grupo de control, en este sentido las técnicas de propensity score y matching permiten encontrar un contrafactual considerando un vector de variables explicativas (X) para determinar la similitud con el grupo de tratamiento.

XII.2.2. Técnica de Matching y propensity score para datos no experimentales

Un **matching exacto** se puede realizar cuando el conjunto de variables x es discreto y la muestra contiene muchas observaciones para cada valor distinto de x. Si el vector de variables x tiene una dimensión elevada, es decir, si pretendo buscar el contrafactual tomando 10 variables, o si las variables son continuas, hacer un matching exacto no es practicable. Este problema ha motivado la realización de matching utilizando **propensity score**.

El **propensity score** resume la información de todas las variables x en un solo valor, **p(x)**. El que no es más que la probabilidad de recibir tratamiento condicional a las variables explicativas x. Así, el propensity score (p(x)) se obtiene de estimar un logit o probit de la variable binaria que toma valor 1 si la persona recibe tratamiento y cero si no recibe tratamiento, luego el valor predicho de la probabilidad de recibir tratamiento condicional a las variables explicativas es el propensity score.

XII.3. Aplicación: El efecto de la capacitación sobre ingresos

Para la siguiente aplicación se utilizará una muestra de 185 hombres que recibieron una capacitación durante 1976 y 1977, y el grupo de control se obtiene de una muestra de 2.490 hombres jefes de hogar menores de 55 años y que no se encuentran pensionados, muestra que fue obtenida del PSID. La variable TREAT indica si la persona ha sido tratada o no. La siguiente tabla presenta el promedio de algunas de las variables claves, mostrando la diferencia entre tratados y grupo de control:

		TRATADOS	CONTROL
AGE	Edad	25.8	34.9
EDUC	Años de escolaridad	10.3	12.1
NODEGREE	1 si educ<12	0.71	0.31
BLACK	1 si es negro	0.84	0.25
HISP	1 si es hispano	0.06	0.03
MARR	1 si es casado	0.19	0.87
U74	1 si estaba desempleado 1n 1974	0.71	0.09
U75	1 si estaba desempleado 1n 1975	0.60	0.10
RE74	Ingreso real en 1974	2096	19429
RE75	Ingreso real en 1975	1532	19063
RE78	Ingreso real en 1978	6349	21554

Existe un comando en STATA **pscore** que estima el propensity score, y permite guardarlo para posteriormente utilizarlo en la estimación del ATET.

El siguiente comando obtiene la estimación del propensity score:

```
pscore  TREAT  AGE AGESQ EDUC EDUCSQ NODEGREE BLACK HISP MARR RE74 RE75 RE74SQ
RE75SQ U74BLACK, pscore(propensity) blockid(estratos) logit comsup numblo(8)
```

Luego de ejecutar el comando se genera una variable llamada propensity que contiene el propensity score estimado, otra variable estratos que contiene el número de bloques en que se ha divido la muestra según el propensity score, el número de bloques por default es 5 en este caso yo impuse que fueran 8. La opción comsup es para que se genere una variable que indique si se cumple la condición de soporte común que requiere el matching. El output de este comando es el siguiente:

```
**************************************************
Algorithm to estimate the propensity score
**************************************************


The treatment is TREAT

      TREAT |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,490       93.08       93.08
          1 |        185        6.92      100.00
------------+-----------------------------------
      Total |      2,675      100.00
```

Estimation of the propensity score

Logistic regression — Number of obs = 2675; LR chi2(13) = 935.44; Prob > chi2 = 0.0000; Log likelihood = -204.9295; Pseudo R2 = 0.6953

TREAT	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
AGE	.3305734	.1203353	2.75	0.006	.0947206	.5664262
AGESQ	-.0063429	.0018561	-3.42	0.001	-.0099808	-.0027049
EDUC	.8247711	.3534216	2.33	0.020	.1320775	1.517465
EDUCSQ	-.0483153	.0186057	-2.60	0.009	-.0847819	-.0118488
NODEGREE	.1299868	.4284278	0.30	0.762	-.7097163	.96969
BLACK	1.132961	.352088	3.22	0.001	.4428814	1.823041
HISP	1.962762	.5673735	3.46	0.001	.8507302	3.074793
MARR	-1.884062	.2994614	-6.29	0.000	-2.470996	-1.297129
RE74	-.0001047	.0000355	-2.95	0.003	-.0001743	-.0000351
RE75	-.0002172	.0000415	-5.23	0.000	-.0002986	-.0001357
RE74SQ	2.36e-09	6.57e-10	3.59	0.000	1.07e-09	3.65e-09
RE75SQ	1.58e-10	6.68e-10	0.24	0.813	-1.15e-09	1.47e-09
U74BLACK	2.137042	.4273667	5.00	0.000	1.299419	2.974665
_cons	-7.552458	2.451721	-3.08	0.002	-12.35774	-2.747173

Note: 19 failures and 0 successes completely determined.

Note: the common support option has been selected
The region of common support is [.00065257, .97487544]

Description of the estimated propensity score
in region of common support

Estimated propensity score

Percentiles	Smallest		
1% .0006813	.0006526		
5% .0008363	.0006581		
10% .0011416	.0006593	Obs	1331
25% .0024351	.0006598	Sum of Wgt.	1331
50% .0111854		Mean	.1388772
	Largest	Std. Dev.	.275571
75% .0779976	.9744237		
90% .6200607	.9747552	Variance	.0759394
95% .9494181	.9747918	Skewness	2.17177
99% .970738	.9748754	Kurtosis	6.296349


```
**************************************************
Step 1: Identification of the optimal number of blocks
Use option detail if you want more detailed output
**************************************************


The final number of blocks is 8

This number of blocks ensures that the mean propensity score
is not different for treated and controls in each blocks



**********************************************************
Step 2: Test of balancing property of the propensity score
Use option detail if you want more detailed output
**********************************************************


The balancing property is satisfied


This table shows the inferior bound, the number of treated
and the number of controls for each block

  Inferior |
  of block |         TREAT
of pscore  |         0          1 |     Total
-----------+----------------------+----------
  .0006526 |     1,043         11 |     1,054
      .125 |        43         13 |        56
       .25 |        22         13 |        35
      .375 |        12         13 |        25
        .5 |         7         21 |        28
      .625 |        11         10 |        21
       .75 |         1         18 |        19
      .875 |         7         86 |        93
-----------+----------------------+----------
     Total |     1,146        185 |     1,331

Note: the common support option has been selected


****************************************
End of the algorithm to estimate the pscore
****************************************
```

Cuando estamos interesados en el efecto causal de un programa de aplicabilidad universal, la medida apropiada es el ATE, esta medida consiste en un promedio de los efectos causales individuales, donde se requiere comparar cada inidviduo del grupo de tratamiento con su contrafactual en el grupo de control. En la práctica, lo

que se ha realizado es buscar el número óptimo de estratos o grupos donde el propensity score es similar entre grupo de tratamiento y grupo de control, lo que garantiza la similitud entre ambos grupos, luego tomando el promedio de la variable de resultado al interior de cada estrato y para cada grupo, y luego sacando un promedio ponderado de estas diferencias, se obtiene finalmente el estimador ATE.

En este ejemplo en particular la medida ATE será obtenida tomando el promedio de la variable de resultado, en este caso el ingreso real post capacitación (1978), en cada estrato tanto para el grupo de tratamiento como para el grupo de control, calcular la diferencia de los promedios de ambos grupos y ponderar esta diferencia por la fracción de individuos tratados en el estrato:

$$ATE = \sum_{s=1}^{8} \omega_s (\overline{RE78}_{s,T=1} - \overline{RE78}_{s,T=0})$$

La información para construir este estimador se obtiene de las siguientes tablas:

```
. tab estratos TREAT

 Number of |         TREAT
     block |         0          1 |     Total
-----------+----------------------+----------
         1 |     1,043         11 |     1,054
         2 |        43         13 |        56
         3 |        22         13 |        35
         4 |        12         13 |        25
         5 |         7         21 |        28
         6 |        11         10 |        21
         7 |         1         18 |        19
         8 |         7         86 |        93
-----------+----------------------+----------
     Total |     1,146        185 |     1,331
```

```
. tab estratos TREAT, summarize(RE78) means

                           Means of RE78

 Number of |        TREAT
     block |         0          1 |     Total
-----------+----------------------+----------
         1 | 13768.653  9541.4008 | 13724.536
         2 | 9278.8254  6028.3812 |  8524.258
         3 | 9008.1485  5720.2962 | 7786.9462
         4 | 3859.3258  4948.4329 | 4425.6615
         5 | 6824.7656  5072.4689 | 5510.5431
         6 | 4173.6352  8235.7418 | 6107.9717
         7 |   16809.1  6709.9257 | 7241.4612
         8 | 1563.8577  6312.9782 | 5955.5175
-----------+----------------------+----------
     Total | 13198.626  6349.1454 | 12246.594
```

El valor estimado del impacto del tratamiento es:

$$ATE = \sum_{s=1}^{8} \omega_s (\overline{RE78}_{s,T=1} - \overline{RE78}_{s,T=0}) = 6.009$$

Recuerde que el estimador ATE tiene validez cuando se supone una universalidad en el tratamiento, es decir, es razonable considerar una ganancia hipotética de asignar el tratamiento aleatoriamente a miembros de la población.

Por otra parte, cuando el interés se centra en las ganancias en el grupo de tratamiento, el estimador pertinente es ATET.

$$ATET = \frac{1}{N_T} \sum_{i \in T=1} \left[Y_{1,i} - \sum_{j} \omega(i,j) Y_{0,j} \right]$$

Donde j son las personas en el grupo de control, y ω(i,j) pondera el resultado de los individuos en el grupo de control de acuerdo a su cercanía con el individuo en el grupo de tratamiento.

Notar la diferencia importante con el estimador ATE el que al evaluar los efectos sobre el total de la población, utiliza directamente el grupo de control para la obtención del efecto causal, mediante el propensity score lo que hace es considerar la heterogeneidad en el efecto causal y agrupar tratamiento y control en estratos similares según el vector de variables X. Sin embargo, el estimador ATET sólo se concentra en los efectos sobre los tratados, pero como los tratados no pueden a la vez no recibir el tratamiento busca un contrafactual en el grupo de control, es decir, busca su "clon" en el grupo de control. Esto se hace mediante las técnicas de matching. Existen distintos tipos de matching: vecino más cercano, kernel, y usando la metodología de radios. Las estimaciones de cada una de estas medidas se realizan mediante los siguientes comandos en STATA:

XII.3.1. Kernel Matching Method

```
. attk RE78 TREAT , pscore(propensity) logit comsup epan boot reps(100)


 The program is searching for matches of each treated unit.
 This operation may take a while.

ATT estimation with the Kernel Matching method

------------------------------------------------------------
n. treat.   n. contr.         ATT   Std. Err.           t
------------------------------------------------------------

      185        1146    1246.556           .           .

------------------------------------------------------------
Note: Analytical standard errors cannot be computed. Use
the bootstrap option to get bootstrapped standard errors.
```

```
Bootstrapping of standard errors

command:            attk  RE78  TREAT  ,  pscore(propensity)  logit  comsup  epan
bwidth(.06)
statistic:    attk       = r(attk)

Bootstrap statistics                                  Number of obs    =      2675
                                                      Replications     =       100

------------------------------------------------------------------------------
Variable     |  Reps  Observed      Bias  Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
        attk |   100  1246.556  149.9585  906.1793  -551.5007   3044.612   (N)
             |                                      -536.3026   3214.554   (P)
             |                                      -867.8006   3025.121  (BC)
------------------------------------------------------------------------------
Note:  N   = normal
       P   = percentile
       BC  = bias-corrected


ATT estimation with the Kernel Matching method
Bootstrapped standard errors

------------------------------------------------------------
n. treat.   n. contr.         ATT   Std. Err.           t
------------------------------------------------------------

      185        1146    1246.556     906.179       1.376

------------------------------------------------------------
```

El coeficiente ATET estimado utilizando la metodología kernel para el matching entrega un impacto sobre el ingreso de participar en programa de capacitación dentro de los tratados de 1.247, pero el valor no resulta ser estadísticamente significativo, no se puede rechazar la hipótesis nula de que sea igual a cero, el estadístico t calculado es menor al valor de tabla.

XII.3.2. Método de matching Vecino más cercano

```
. attnd RE78 TREAT , pscore(propensity) logit comsup boot reps(100)


 The program is searching the nearest neighbor of each treated unit.
 This operation may take a while.
```

```
ATT estimation with Nearest Neighbor Matching method
(random draw version)
Analytical standard errors

------------------------------------------------------------
n. treat.   n. contr.         ATT    Std. Err.          t
------------------------------------------------------------

      185          57     560.287     2205.663      0.254

------------------------------------------------------------
Note: the numbers of treated and controls refer to actual
nearest neighbour matches

Bootstrapping of standard errors

command:      attnd RE78 TREAT , pscore(propensity) logit comsup
statistic:    attnd      = r(attnd)

Bootstrap statistics                              Number of obs    =      2675
                                                  Replications     =       100

------------------------------------------------------------------------------
Variable     |  Reps  Observed      Bias  Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
       attnd |   100  560.2872  569.4494  1060.183  -1543.346    2663.92   (N)
             |                                      -968.0416   3148.873   (P)
             |                                      -2029.688   2237.005  (BC)
------------------------------------------------------------------------------
Note:  N   = normal
       P   = percentile
       BC  = bias-corrected


ATT estimation with Nearest Neighbor Matching method
(random draw version)
Bootstrapped standard errors

------------------------------------------------------------
n. treat.   n. contr.         ATT   Std. Err.          t
------------------------------------------------------------

      185          57     560.287    1060.183       0.528

------------------------------------------------------------
Note: the numbers of treated and controls refer to actual
nearest neighbour matches
```

El coeficiente ATET estimado utilizando la metodología del vecino más cercano entrega un impacto sobre el ingreso de participar en programa de capacitación de

560, pero el valor no resulta ser estadísticamente significativo, no se puede rechazar la hipótesis nula de que sea igual a cero, el estadístico t calculado es menor al valor de tabla.

XII.3.3. Matching radius (r=0.001)

```
. attr RE78 TREAT , pscore(propensity) logit radius(0.001) comsup boot
reps(100)


 The program is searching for matches of treated units within radius.
 This operation may take a while.

ATT estimation with the Radius Matching method
Analytical standard errors

-------------------------------------------------------------
n. treat.   n. contr.         ATT   Std. Err.           t
-------------------------------------------------------------

       57         583   -9358.228     997.561      -9.381

-------------------------------------------------------------
Note: the numbers of treated and controls refer to actual
matches within radius


Bootstrapping of standard errors

command:      attr RE78 TREAT , pscore(propensity) logit comsup radius(.001)
statistic:    attr       = r(attr)

Bootstrap statistics                                Number of obs    =      2675
                                                    Replications     =       100

------------------------------------------------------------------------------
Variable     |  Reps  Observed      Bias  Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
        attr |   100 -9358.228  1795.617  2658.161   -14632.6  -4083.859   (N)
             |                                      -12074.71  -1561.896   (P)
             |                                      -12984.94  -5661.869  (BC)
------------------------------------------------------------------------------
Note:  N   = normal
       P   = percentile
       BC  = bias-corrected
```

```
ATT estimation with the Radius Matching method
Bootstrapped standard errors

------------------------------------------------------------
n. treat.   n. contr.         ATT   Std. Err.           t
------------------------------------------------------------

       57         583   -9358.228    2658.161      -3.521

------------------------------------------------------------
Note: the numbers of treated and controls refer to actual
matches within radius
```

El coeficiente ATET estimado utilizando la metodología de radio (r=0.001) entrega un impacto sobre el ingreso de participar en programa de capacitación de -9.358, resulta ser estadísticamente significativo, se rechaza la hipótesis nula de que sea igual a cero, el estadístico t calculado es mayor al valor de tabla. Notar que se puede realizar matching para 57 individuos del grupo de control. Si se entrega un radio más pequeño para hacer el matching se obtiene el siguiente resultado:

XII.3.4. Radius=0.00001

```
. attr RE78 TREAT , pscore(propensity) logit radius(0.00001) comsup boot
reps(100)


 The program is searching for matches of treated units within radius.
 This operation may take a while.


ATT estimation with the Radius Matching method
Analytical standard errors

------------------------------------------------------------
n. treat.   n. contr.         ATT   Std. Err.           t
------------------------------------------------------------

       16          13     223.468    4551.850       0.049

------------------------------------------------------------
Note: the numbers of treated and controls refer to actual
matches within radius
```

```
Bootstrapping of standard errors

command:      attr RE78 TREAT , pscore(propensity) logit comsup radius(.00001)
statistic:    attr       = r(attr)

Bootstrap statistics                                Number of obs    =      2675
                                                    Replications     =       100

------------------------------------------------------------------------------
Variable     |  Reps  Observed      Bias  Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
        attr |   100  223.4685 -608.0974  4402.156  -8511.365   8958.302   (N)
             |                                      -8533.206   6074.561   (P)
             |                                      -11687.47   5927.729  (BC)
------------------------------------------------------------------------------
Note:  N   = normal
       P   = percentile
       BC  = bias-corrected


ATT estimation with the Radius Matching method
Bootstrapped standard errors

-------------------------------------------------------------
n. treat.   n. contr.         ATT   Std. Err.           t
-------------------------------------------------------------

       16          13     223.468    4402.156       0.051

-------------------------------------------------------------
Note: the numbers of treated and controls refer to actual
matches within radius
```

El coeficiente ATET estimado utilizando la metodología de radius (r=0.00001) para el matching entrega un impacto sobre el ingreso de participar en programa de capacitación de 224, pero el valor no resulta ser estadísticamente significativo, no se puede rechazar la hipótesis nula de que sea igual a cero, el estadístico t calculado es menor al valor de tabla. Tan sólo 16 de los individuos en el grupo de tratamiento son asignados a un grupo de control.