Econometría I

Profesor: Rómulo Chumacero
Ayudantes: Adolfo Fuentes[1], Rodrigo Miranda

Apuntes de Stata: Introducción y OLS
(Versión Preliminar)
Otoño 2014

1. Iniciando Stata

Para el inicio de Stata tenemos que usar el comando:

set mem 400m

Que sirve para destinar una parte de la memoria RAM del computador para el uso exclusivo de Stata. El número 400 puede variar, dependiendo de la cantidad de memoria que se quiera dedicar al programa, sin embargo, 400 MB suele ser una buena cantidad.

current memory allocation

settable	current value	description	memory usage (1M = 1024k)
set maxvar	5000	max. variables allowed	1.909M
set memory	400M	max. data space	400.000M
set matsize	400	max. RHS vars in models	1.254M
			403.163M

Figura 1: Resultado de fijar la memoria en 400 MB

Un comando adicional que se sugiere incluir al momento de iniciar Stata es el comando:

set more off

Que hace que el programa nos entregue las tablas y distribuciones de manera completa inmediatamente. Sino incluimos este comando tendremos situaciones en donde tendremos que apretar "more" para poder hacer que se nos muestre todo el contenido de una tabla. Un ejemplo de esto es el siguiente:

Variable	Obs	Mean	Std. Dev.	Min	Max
segmento	246924	8637131	3432485	1101101	1.52e+07
idviv	246924	109.0835	90.38125	1	468
hogar	246924	1.01884	.1615758	1	7
o	246924	2.639128	1.592028	1	16
folio	0				
region	246924	8.39596	3.449086	1	15
provincia	246924	86.30477	34.30732	11	152
comuna	246924	8636.99	3432.489	1101	15202
zona	246924	1.357948	.4793978	1	2
estrato	246924	86371.26	34324.85	11011	152022
contesta	246924	2.031901	.8815967	1	3
expr	246924	68.75555	105.6603	1	4630
expp	246924	68.73118	105.2808	1	4467
expc	246924	68.64969	105.932	1	4166
pco1	246924	3.28821	2.784826	1	14
sexo	246924	1.509558	.4999097	1	2
edad	246924	35.28934	22.40283	0	107
ecivil	246924	4.412228	2.78017	1	7
nucleo	246924	1.148357	.4101702	0	10
pco2	246924	2.760999	2.50388	1	13
numper	246924	4.275838	1.810455	1	16
r7	84956	66.19727	37.04895	1	88
r8	13358	1.593502	.491198	1	2
r9	13358	4.811648	1.955714	1	6
r10a	84956	.2252813	.466667	0	9
r10b	84956	.094955	.4954623	0	99
r11a	84956	1.550332	.9005899	1	9
r11b	84956	1.30158	.720503	1	9
r11c	84956	1.650372	.9178229	1	9
r11d	84956	1.889814	1.014297	1	9

--more--

Figura 2: Efectos de no utilizar el comando "set more off"

[1] adfuente@fen.uchile.cl

2. Variables Númericas y de Texto

2.1. Creación

Al momento de crear una variable debemos fijar la cantidad de observaciones con las cuales vamos a trabajar. Para ello se utiliza el comando:

set obs 2000

Que hará que cada vez que se cree una variable está tendrá 2000 observaciones (dado el valor del ejemplo).

Con esto fijado, vamos a crear una variable númerica llamada "var1" que tomará el valor 0 para todas las observaciones. Esto se hace mediante el comando:

gen var1=0

Sin embargo, no siempre se busca tener variables númericas, sino que queremos una variable de texto. Para ello generamos el siguiente comando:

gen var2="texto"

Luego, si queremos visualizar las variables que hemos creado, utilizamos el siguiente comando[2]

browse var1 var2

Que nos arrojará una ventana de esta forma:

	var1	var2
1	0	texto
2	0	texto
3	0	texto
4	0	texto
5	0	texto
6	0	texto
7	0	texto
8	0	texto
9	0	texto
10	0	texto
11	0	texto
12	0	texto
13	0	texto
14	0	texto
15	0	texto
16	0	texto
17	0	texto
18	0	texto
19	0	texto
20	0	texto
21	0	texto

Figura 3: Visualización de las variables var1 y var2

[2]También se puede utilizar el comando ' edit". Sin embargo, este comando permite cambiar manualmente los valores de las variables.

2.2. Modificación

Supongamos que ahora queremos modificar la variable var2 y hacer que para las primeras 1000 observaciones tome el texto "primeros", y las últimas 2000 observaciones tome el texto de "ultimos". El comando con el que hacemos esto corresponde a:

```
replace var2 = "Primero" in 1/1000
replace var2 = "Segundo" in 1001/2000
```

Luego, también se pueden crear variables mas complejas mediante modificaciones. Un ejemplo corresponde a crear una variable vacia, y a continuación, hacer que tome el valor 1 para las primeras 200 observaciones, el valor 2 para las observaciones 201 a 600, el valor 3 para las observaciones 601 a 1500, y el valor 4 para las 500 últimas.

```
gen var3 = .
replace var3 = 1 in 1/200
replace var3 = 2 in 201/600
replace var3 = 3 in 601/1500
replace var3 = 4 in 1501/2000
```

2.3. Eliminación

Para eliminar una variable (sea númerica o de texto), se utiliza el comando:

```
drop var1
```

Ahora, si queremos eliminar condicional a otra variable, podemos hacer algo del estilo:

```
keep if var2==2 | var2==3
```

En este caso, conservaremos la variable var2 cuando tome los valores 2 **o** 3, y eliminaremos todos las otras observaciones.

Otro ejemplo sería este comando:

```
keep if var2>=2 & var2<=3
```

Asi, conservaremos la variable var2 siempre que su valor sea mayor igual que 2 **y** menor igual que 3[3].
Un último ejemplo corresponde a:

```
keep if var2!=1
```

Que corresponde a conservar la variable var2 cuando no su valor sea **distinto** de 1.

2.4. Cambio de Nombre

Para cambiar de nombre a una variable utilizamos el comando:

```
remane var3 sexo
```

Y ahora, haremos que esta tome el valor 1 cuando su valor es 2, y el valor 0 cuando su valor es 3. Esto lo hacemos con el comando:

```
replace sexo=1 if sexo==2
replace sexo=0 if sexo==3
```

[3]En los comandos condicionales el signo = siempre va al final. Por otro lado, también se pueden

2.5. Etiquetas

En la creación de variables también es útil poner descripciones para que otras personas puedan ver lo que hace la variable. Esto se hace con el siguiente comando:

label var sexo “0 Mujer ; 1 Hombre”

Donde lo que estamos explicando es que la variable cuando toma el valor 0 nos referimos a una mujer, mientras que si si toma el valor 1 nos referimos a un hombre.

3. Comandos de descripción estadística

Cuando contamos con bases de datos amplias, suele ser útil tener una visión estadística de los datos con los que contamos. Así hay varios comandos que sirven para mostrar información sobre las bases de datos.

3.1. Summarize

Uno de los comandos más utilizados corresponde a:

summarize

El cual, nos entregará una tabla que incluye la siguiente información:

- Nombre de la variable
- Media
- Desviación Estandar
- Valor mínimo
- Valor máximo

Si el comando se utiliza solo nos entregará esta información para todas las variables de la base de datos. Sin embargo, tambien podemos pedirle que nos entregue esta información solo para las variables que nosotros queramos. Para esto utilizamos:

summarize yopraj

Que nos entregará la siguiente tabla:

Variable	Obs	Mean	Std. Dev.	Min	Max
yopraj	86009	304561.6	416102.5	1105	1.44e+07

Figura 4: Detalle comando summarize

3.2. Mean

Otro comando de este tipo corresponde a:

mean

Que entrega el nombre de la variable, su media, desviación estandar e intervalo de confianza. Sin embargo, a diferencia del comando summarize, este comando si requiere la especificación de la variable de la cual queremos estos datos.

Mean estimation Number of obs = 86009

	Mean	Std. Err.	[95% Conf.	Interval]
yopraj	304561.6	1418.824	301780.7	307342.4

Figura 5: Detalle comando mean

3.3. Correlate

A veces queremos determinar el nivel de correlación entre dos variables, por diversos motivos. El comando que hace esto corresponde a:

correlate

Donde nos calculará la correlación entre las diversas variables que sean incluidas en el comando. A modo de ejemplo, si queremos ver la relación entre el nivel de ingreso y el sexo de las personas, tenemos el siguiente comando:

correlate yopraj sexo

Obteniendo la siguiente tabla:

(obs=86009)

	yopraj	sexo
yopraj	1.0000	
sexo	-0.0860	1.0000

Figura 6: Detalle comando correlate

3.4. Xtile

Otras veces queremos trabajar con los datos por cuantiles, de forma de poder trabajar la variable desagregada. La forma de trabajar esto es crear una variable que tome el valor correspondiente al cuantil con el que estamos trabajando. Así, para crear esta variable, utilizamos el comando:

xtile decil = yopraj, nquatiles(10)

Donde dentro del comando nquantile colocamos la cantidad de partes en las cuales queremos dividir la variable. Así, al ingresar 10, estamos dividiendo la variable en deciles.

3.5. Table

Si queremos tener una tabla personalizada, que nos entregue los datos que nosotros deseamos, entonces debemos utilizar el comando Table. Supongamos que teniendo la variable de ingreso y la variable decil (creada antes), queremos ver cual es la media de ingreso y la desviación estandar por decil. Para esto, utilizamos el comando:

```
table decil, c(n yopraj mean yopraj sd max yopraj min yopraj)
```

De forma que nos hará una tabla de la variable decil, que nos entregará la cantidad de datos que contiene cada grupo (n yopraj), la media (mean yopraj), la desviación estandar (sd yopraj), el valor máximo (max yopraj) y el valor mínimo (min yopraj).

Al hacer esto obtenemos la siguiente tabla:

10 quantiles of yopraj	N(yopraj)	mean(yopraj)	max(yopraj)	min(yopraj)
1	8,608	62825.31842	99450	1105
2	8,594	124566.6327	145663	99548
3	12,089	167611.4996	176800	145759
4	7,276	182220.9977	182325	176819
5	6,611	193919.1472	198900	182419
6	8,464	215718.8834	226525	198907
7	8,566	261290.6604	289776	226850
8	8,652	328164.1356	375700	289813
9	9,170	454274.3411	552500	375955
10	7,979	1112994.42	14378000	552656

Figura 7: Detalle comando table

3.6. Tabulate

Supongamos que queremos ver la frecuencia y la frecuencia acumulada que lleva asociada cada valor dentro de una variable. El comando utilizado para ver esto es:

```
tabulate decil
```

El cual requiere indicar la variable sobre la cual queremos ver la distribución. Así, veamos la distribución de la variable decil, la cual teóricamente, debiese tener que cada decil tiene una frecuencia de 10 %.

10 quantiles of yopraj	Freq.	Percent	Cum.
1	8,608	10.01	10.01
2	8,594	9.99	20.00
3	12,089	14.06	34.06
4	7,276	8.46	42.52
5	6,611	7.69	50.20
6	8,464	9.84	60.04
7	8,566	9.96	70.00
8	8,652	10.06	80.06
9	9,170	10.66	90.72
10	7,979	9.28	100.00
Total	86,009	100.00	

Figura 8: Detalle comando tabulate

Ahora, si incluimos dos variables, el comando nos entregará la distribución de la segunda variable dado un valor de la primera variable, perdiendo en el camino los datos de frecuencia y frecuencia acumulada. Así, si utilizamos el comando:

tabulate decil sexo

Obtenemos el siguiente resultado:

10 quantiles of yopraj	r2: sexo hombre	mujer	Total
1	3,930	4,678	8,608
2	4,594	4,000	8,594
3	7,731	4,358	12,089
4	4,931	2,345	7,276
5	4,498	2,113	6,611
6	5,702	2,762	8,464
7	6,064	2,502	8,566
8	6,325	2,327	8,652
9	6,610	2,560	9,170
10	5,865	2,114	7,979
Total	56,250	29,759	86,009

Figura 9: Detalle comando tabulate

4. Gráficos

Muchas veces es útil graficar las series, de forma de ver la distribución que tiene y observar los valores extremos. Para graficar, tenemos un par de comandos útiles.

4.1. Graph Bar

Este comando realiza gráficos de barras indicando un estadístico que se le solicite dentro de una gama de opciones. Así, podemos pedirle a este gráfico que nos entregue la media del ingreso y diferencie hombres de mujeres. Esto se hace con el comando:

graph bar (mean) yopraj, over(sexo)

Entregandonos la siguiente figura:

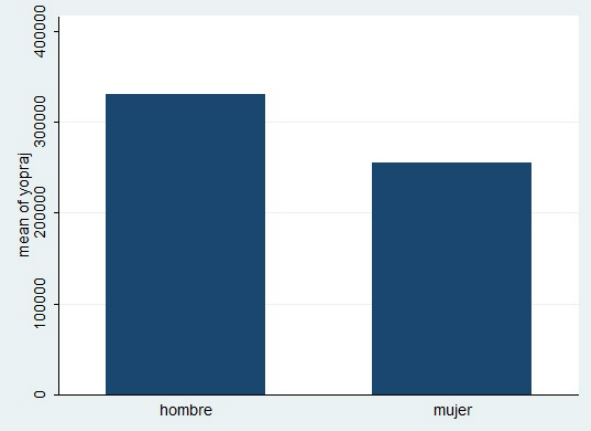

Figura 10: Ejemplo Comando graph bar

4.2. Graph Pie

Este comando realiza gráficos de torta, utilizados para mostrar como se distribuye porcentualmente un dato en varias subcategorias. Por ejemplo, si queremos ver como se distribuye el ingreso de las personas entre ingreso autónomo (yautaj) y subsidios monetarios (ysubaj), tenemos el siguiente comando:

```
graph pie yautaj ysubaj
```

Que nos entrega la siguiente figura:

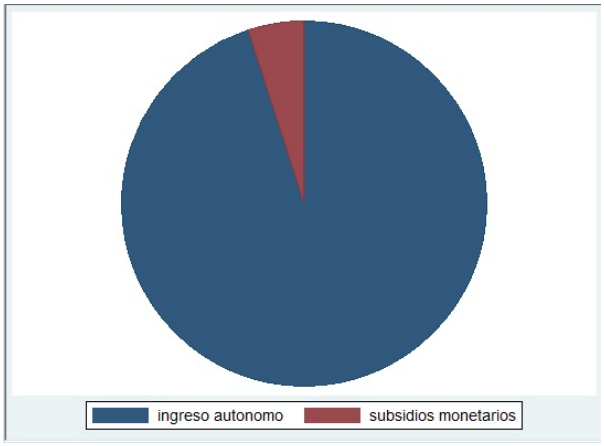

Figura 11: Ejemplo Comando graph pie

4.3. Graph Box

A veces, puede ser util comparar las frecuencias de distribución de una misma variable separando por una dummy. Así, supongamos que queremos un gráfico de frecuencias que nos muestre la distribución del ingreso de los hombres y de las mujeres. Esto, lo hacemos con el siguiente comando:

```
graph box yopraj, over(sexo)
```

Que nos arroja la siguiente figura:

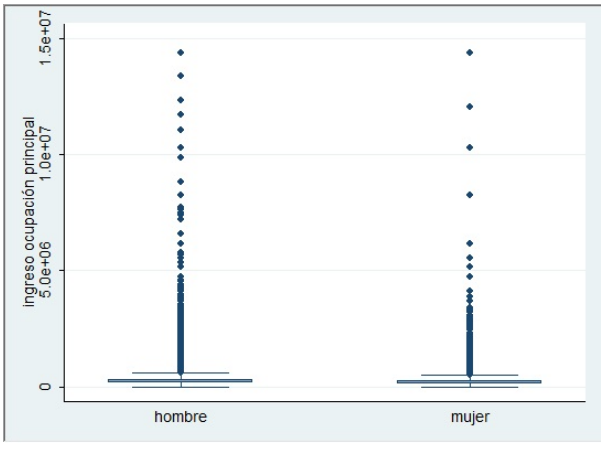

Figura 12: Ejemplo Comando graph box

4.4. Histogram

Un comando muy útil para ver la distribución de una variable corresponde al comando histogram, el cual, entrega el histograma de la variable, vale decir un gráfico que asocia a cada valor de la varaible la frecuencia que esta tiene. Así, si le pedimos el histograma a la variable ingreso de la ocupación principal (yopraj), esta nos dirá la cantidad de personas que logra un determinado nivel de ingreso, para todos los niveles de ingreso.

La forma de utilizarlo es la siguiente:

histogram yopraj

Que nos arroja la siguiente figura:

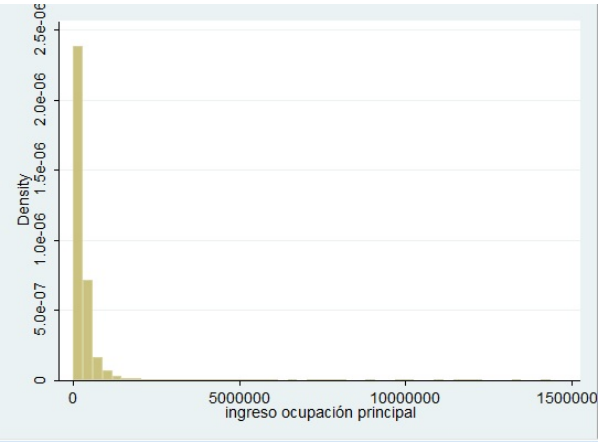

Figura 13: Ejemplo Comando histogram

Un dato útil sobre las distribuciones del ingreso, es que este generalmente sigue una distribución log normal, vale decir, es una distribución tal que al aplicarle logaritmo adquiere distribución normal. Así creemos el logaritmo de la variable yopraj para verificar esto:

gen lnyopraj = ln(yopraj)
histogram lnyopraj

Donde obtenemos la siguiente figura:

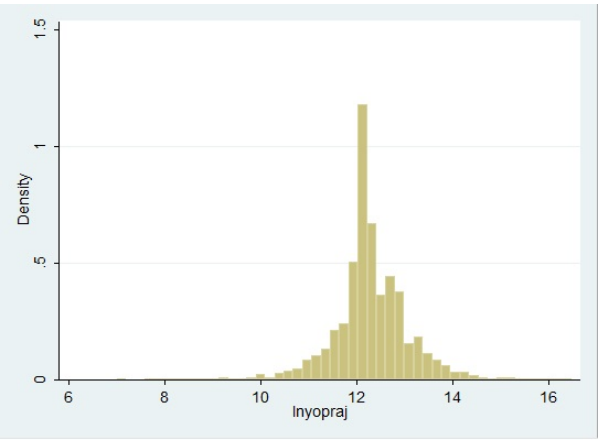

Figura 14: Ejemplo Distribución log normal

5. Modelo de Mínimos Cuadrados Ordinarios (OLS)

De clases ya sabemos que el estimador de OLS está determinado por un modelo de regresión lineal de la forma:

$$Y = X\beta + u$$

Donde el estimador de β, $\hat{\beta}$, es aquel parámetro que minimiza la suma de los errores al cuadrado, o bien, en terminos matriciales:

$$\hat{\beta} \in \text{argmin} \;\; u'u$$

Luego, el estimador está dado por:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Donde el estimador existe si y solo si la matrix X'X tiene inversa, o bien, las columas de X son linealmente independientes.

Supongamos que queremos estimar una ecuación de Mincer, es decir, una ecuación que determina el ingreso laboral de las personas en base a caracteristicas personales, como el sexo y los años de educación. En concreto, queremos realizar la siguiente estimación:

$$\text{yopraj} = \beta_0 + \beta_1 sexo + \beta_2 esc + \beta_3 edad + \beta_4 exp + \beta_5 exp^2$$

Donde se define la variable exp (experiencia laboral), como:

$$exp = edad - esc - 6$$

Sin embargo, antes de poder estimar este modelo debemos generar la variable experiencia que no está en la base de datos. Esto lo hacemos con el siguiente comando:

```
gen exp = edad - esc - 6
gen exp = exp*exp
```

Por otro lado, la variable sexo está definida para tomar el valor 1 cuando es hombre y el valor 2 cuando es mujer. En este sentido, sabemos que una variable dummy debe estar definida como 0 y 1, dado que demuestra algo cualitativo y no cuantitativo. La intuición de esto, es que si dejamos la variable como está, le estamos diciendo que ser mujer equivale a ser dos hombres. Para corregir esto, hay varios caminos, pero uno de los mas rápidos es generar una nueva variable que sea equivalente a sexo menos 1. Esto lo hacemos con el siguiente comando:

```
gen genero = sexo - 1
```

A continuación, notaremos que tenemos un problema con el modelo a estimar. Si nos fijamos un poco, veremos que la matriz X tiene las siguientes columnas asociadas:

	edad	esc	genero	exp	constant	exp2
1	74	2	0	66	1	4356
2	70	2	1	62	1	3844
3	15	6	0	3	1	9
4	35	0	0	29	1	841
5	29	7	1	16	1	256
6	5	.	1	.	1	.
7	2	.	1	.	1	.
8	49	5	1	38	1	1444
9	23	11	1	6	1	36
10	39	7	0	26	1	676
11	37	10	1	21	1	441
12	11	.	1	.	1	.
13	9	.	1	.	1	.
14	1	.	1	.	1	.
15	43	13	0	24	1	576
16	30	12	1	12	1	144
17	55	9	1	40	1	1600
18	32	14	1	12	1	144
19	19	12	1	1	1	1
20	9	.	0	.	1	.
21	13	.	1	.	1	.

Figura 15: Columnas de la matriz X

Donde la columna edad puede ser formada con la siguiente operación lineal:

$$edad = esc + exp + 6 \cdot constant$$

De forma que las columnas de esta matriz no son linealmente independientes, por lo tanto, la regresión no se puede realizar.

A continuación estimemos el siguiente modelo:

$$\text{yopraj} = \beta_0 + \beta_1 genero + \beta_2 esc + \beta_3 exp + \beta_4 exp^2$$

Este modelo se estima con el siguiente comando:

```
regress yopraj genero esc exp exp2
```

El cual arroja los siguientes resultados:

Source	SS	df	MS
Model	1.9933e+15	4	4.9833e+14
Residual	1.2896e+16	85970	1.5001e+11
Total	1.4890e+16	85974	1.7319e+11

Number of obs = 85975
F(4, 85970) = 3321.93
Prob > F = 0.0000
R-squared = 0.1339
Adj R-squared = 0.1338
Root MSE = 3.9e+05

yopraj	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
genero	-110078.7	2806.713	-39.22	0.000	-115579.9	-104577.6
esc	44956.05	409.5638	109.77	0.000	44153.31	45758.8
exp	8684.642	278.0245	31.24	0.000	8139.716	9229.567
exp2	-38.70738	4.743153	-8.16	0.000	-48.00391	-29.41084
_cons	-294655.7	6422.099	-45.88	0.000	-307243	-282068.4

Figura 16: Tabla de resultados de una regresión lineal

La tabla entrega bastante información, asi que la veremos por partes. En primer lugar, tenemos los valores estimados de los coeficientes de la regresión. Estos valores salen de la fórmula:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Notar ademas que dada la forma del modelo, corresponden a los efectos marginales sobre la variable. Así, tenemos por ejemplo que:

$$\frac{\partial \text{yopraj}}{\partial esc} = \beta_1 = 44956,05$$

Vale decir, cada año adicional de educación agrega al ingreso mensual $44.956,05.

Source	SS	df	MS		Number of obs =	85975
					F(4, 85970) =	3321.93
Model	1.9933e+15	4	4.9833e+14		Prob > F =	0.0000
Residual	1.2896e+16	85970	1.5001e+11		R-squared =	0.1339
					Adj R-squared =	0.1338
Total	1.4890e+16	85974	1.7319e+11		Root MSE =	3.9e+05

yopraj	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
genero	-110078.7	2806.713	-39.22	0.000	-115579.9	-104577.6
esc	44956.05	409.5638	109.77	0.000	44153.31	45758.8
exp	8684.642	278.0245	31.24	0.000	8139.716	9229.567
exp2	-38.70738	4.743153	-8.16	0.000	-48.00391	-29.41084
_cons	-294655.7	6422.099	-45.88	0.000	-307243	-282068.4

Figura 17: Coeficientes de los estimadores

En segundo lugar, veamos la desviación estandar de los parámetros estimados. Estos valores salen de la fórmula:

$$\sqrt{\hat{V}(\hat{\beta})} = \sqrt{\tilde{\sigma}(X'X)^{-1}}$$

Donde $\tilde{\sigma}$ corresponde al estimador insesgado de la varianza.

Source	SS	df	MS		Number of obs =	85975
					F(4, 85970) =	3321.93
Model	1.9933e+15	4	4.9833e+14		Prob > F =	0.0000
Residual	1.2896e+16	85970	1.5001e+11		R-squared =	0.1339
					Adj R-squared =	0.1338
Total	1.4890e+16	85974	1.7319e+11		Root MSE =	3.9e+05

yopraj	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
genero	-110078.7	2806.713	-39.22	0.000	-115579.9	-104577.6
esc	44956.05	409.5638	109.77	0.000	44153.31	45758.8
exp	8684.642	278.0245	31.24	0.000	8139.716	9229.567
exp2	-38.70738	4.743153	-8.16	0.000	-48.00391	-29.41084
_cons	-294655.7	6422.099	-45.88	0.000	-307243	-282068.4

Figura 18: Desviación estandar de los estimadores

En tercer lugar, está el valor de tabla del test t y el p-value asociado. El valor de tabla del test t sale de hacer un test de hipótesis donde la nula es evaluar que el valor del parámetro es 0. Entonces la fórmula es:

$$t = \frac{\hat{\beta} - 0}{\sqrt{\hat{V}(\hat{\beta})}}$$

Mientras tanto, el valor de la izquierda representa el p-value asociado a ese valor de tabla. Este valor intuitivamente nos dice el "riesgo" que estamos corriendo al rechazar la hipótesis nula. Así, si el valor fuese 0,000 significa que la hipótesis nula se rechaza a cualquier nivel de riesgo, mientras que si toma el valor 0,060 significa que a un nivel de riesgo del 5 % no podemos rechazar la nula, pero si a un 10 %.

Source	SS	df	MS		Number of obs =	85975
Model	1.9933e+15	4	4.9833e+14		F(4, 85970) =	3321.93
Residual	1.2896e+16	85970	1.5001e+11		Prob > F =	0.0000
					R-squared =	0.1339
Total	1.4890e+16	85974	1.7319e+11		Adj R-squared =	0.1338
					Root MSE =	3.9e+05

yopraj	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
genero	-110078.7	2806.713	-39.22	0.000	-115579.9	-104577.6
esc	44956.05	409.5638	109.77	0.000	44153.31	45758.8
exp	8684.642	278.0245	31.24	0.000	8139.716	9229.567
exp2	-38.70738	4.743153	-8.16	0.000	-48.00391	-29.41084
_cons	-294655.7	6422.099	-45.88	0.000	-307243	-282068.4

Figura 19: Valor de tabla de test t y p-value

En cuarto lugar, tenemos los intervalos de confianza, que corresponde al intervalo dentro del cual aceptamos que esté el estimador, estos invervalos salen de la siguiente fórmula:

$$[\hat{\beta} - t_{df,\alpha/2} \cdot \sqrt{\hat{V}(\hat{\beta})} \quad , \quad \hat{\beta} + t_{df,\alpha/2} \cdot \sqrt{\hat{V}(\hat{\beta})}]$$

Donde $t_{df,\alpha/2}$ corresponde al valor crítico de una t-student con df grados de libertad al $\alpha/2$ porciento de riesgo. Este número para variables con muchos grados de libertad y al 2,5 % de riesgo (recordar que es 5 % divido en dos colas) da asintóticamente 1,96 en valor absoluto.

Source	SS	df	MS		Number of obs =	85975
Model	1.9933e+15	4	4.9833e+14		F(4, 85970) =	3321.93
Residual	1.2896e+16	85970	1.5001e+11		Prob > F =	0.0000
					R-squared =	0.1339
Total	1.4890e+16	85974	1.7319e+11		Adj R-squared =	0.1338
					Root MSE =	3.9e+05

yopraj	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
genero	-110078.7	2806.713	-39.22	0.000	-115579.9	-104577.6
esc	44956.05	409.5638	109.77	0.000	44153.31	45758.8
exp	8684.642	278.0245	31.24	0.000	8139.716	9229.567
exp2	-38.70738	4.743153	-8.16	0.000	-48.00391	-29.41084
_cons	-294655.7	6422.099	-45.88	0.000	-307243	-282068.4

Figura 20: Intervalos de confianza

En quinto lugar, tenemos el test F y el p-value asociado. Este test corresponde a un test de significancia global que tiene por hipótesis nula que todos los coeficientes del modelo (menos la constante) son equivalentes a 0, vale decir, que las variables que tiene nuestro modelo (salvo la constante) son todas no significativas.

El valor de este test viene de la fórmula:

$$F = \frac{T-k}{q} \cdot \frac{\hat{\beta}'(X'X)\hat{\beta}}{\hat{u}'\hat{u}}$$

Lo cual, corresponde a reemplazar $Q = I$ y $c = 0$ en la fórmula de clases.

Source	SS	df	MS
Model	1.9933e+15	4	4.9833e+14
Residual	1.2896e+16	85970	1.5001e+11
Total	1.4890e+16	85974	1.7319e+11

Number of obs = 85975
F(4, 85970) = 3321.93
Prob > F = 0.0000
R-squared = 0.1339
Adj R-squared = 0.1338
Root MSE = 3.9e+05

yopraj	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
genero	-110078.7	2806.713	-39.22	0.000	-115579.9	-104577.6
esc	44956.05	409.5638	109.77	0.000	44153.31	45758.8
exp	8684.642	278.0245	31.24	0.000	8139.716	9229.567
exp2	-38.70738	4.743153	-8.16	0.000	-48.00391	-29.41084
_cons	-294655.7	6422.099	-45.88	0.000	-307243	-282068.4

Figura 21: Valor tabla de test F y p-value

Finalmente, tenemos el R^2 y el R^2 ajustado. La idea del concepto es tener una medida de la cantidad de la varianza de las Y que es capturada por los parámetros X. La fórmula que se utiliza está dada por:

$$\begin{aligned} R^2 &= \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} \\ \overline{R}^2 &= 1 - \frac{SSR}{TSS}\frac{T}{T-k} \end{aligned}$$

Source	SS	df	MS
Model	1.9933e+15	4	4.9833e+14
Residual	1.2896e+16	85970	1.5001e+11
Total	1.4890e+16	85974	1.7319e+11

Number of obs = 85975
F(4, 85970) = 3321.93
Prob > F = 0.0000
R-squared = 0.1339
Adj R-squared = 0.1338
Root MSE = 3.9e+05

yopraj	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
genero	-110078.7	2806.713	-39.22	0.000	-115579.9	-104577.6
esc	44956.05	409.5638	109.77	0.000	44153.31	45758.8
exp	8684.642	278.0245	31.24	0.000	8139.716	9229.567
exp2	-38.70738	4.743153	-8.16	0.000	-48.00391	-29.41084
_cons	-294655.7	6422.099	-45.88	0.000	-307243	-282068.4

Figura 22: R^2 y $\overline{R}^2$

6. Anexo

6.1. Mejora de Regresión al utilizar Logaritmo

Como vimos anteriormente, al aplicar logartimo a una variable que distribuye log normal obtenemos una distribución normal. Esta transformación ayuda en general a mejorar los resultados de la regresión. Para ver esto, realizaremos la siguiente estimación:

$$\ln \text{yopraj} = \beta_0 + \beta_1 genero + \beta_2 esc + \beta_3 exp + \beta_4 exp^2$$

Cuyos resultados están en la siguiente figura:

Source	SS	df	MS
Model	11541.988	4	2885.49701
Residual	36412.1999	85970	.423545422
Total	47954.188	85974	.557775467

Number of obs = 85975
F(4, 85970) = 6812.72
Prob > F = 0.0000
R-squared = 0.2407
Adj R-squared = 0.2407
Root MSE = .6508

lnyopraj	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
genero	-.3397971	.0047161	-72.05	0.000	-.3490406	-.3305535
esc	.1026524	.0006882	149.16	0.000	.1013035	.1040012
exp	.0220412	.0004672	47.18	0.000	.0211255	.0229568
exp2	-.0001746	7.97e-06	-21.91	0.000	-.0001902	-.000159
_cons	10.99161	.0107911	1018.58	0.000	10.97046	11.01276

Figura 23: Resultados aplicación logaritmo

Donde podemos ver que ahora los test t y el test F son mucho mas fuertes y aumentan ambos R^2, lo que sería un primer indicio para decir que la regresión es mejor a la anterior.

6.2. Factores de Expansión

Muchas encuestas de nivel social (CASEN,EPS,ELPI) tienen observaciones que van ligadas a factores de expansión. Los factores de expansión son números que nos dicen a cuantas observaciones de la población representa cada observación de la muestra. La forma de aplicar los factores de expansión es a través del comando:

[w=expr]

Así por ejemplo, veamos la composición entre hombres y mujeres en la muestra, para esto utilizamos:

table sexo

La cual viene dada por:

r2: sexo	Freq.
hombre	121,102
mujer	125,822

Figura 24: Mujeres y Hombres de la muestra

Asi, vemos que tenemos en total la suma de 246.924 datos.

Mientras que si utilizamos factor de expansión tenemos:

table sexo [w=expr]

La cual viene dada por:

r2: sexo	Freq.
hombre	**8179631**
mujer	**8797764**

Figura 25: Mujeres y Hombres poblacional estimado

Donde en total tenemos 16.977.395 datos.

El comando del factor de expansión no solo se puede utilizar para tablas descriptivas, sino que también esta habilitado para las regresiones, siendo esta de gran utilidad para tener mejores propiedades estadísticas de los estimadores.