Serie Apuntes de Clase ΩBΓ

N° 08
Octubre del 2014

ECONOMETRIA APLICADA MODELOS DE ELECCION BINARIA EN STATA 14

Rafael Bustamante Romaní

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú, DECANA DE AMÉRICA

FACULTAD DE CIENCIAS ECONÓMICAS

La **Serie Apuntes de Clase Omega Beta Gamma** tiene por objetivo difundir los materiales de enseñanza generados por los docentes que tienen a su cargo el desarrollo de las asignaturas que forman parte de los Planes de Estudios de las Escuelas Académico-Profesionales de la Facultad de Ciencias Económicas de la Universidad Nacional Mayor de San Marcos. Estos documentos buscan proporcionar a los estudiantes la explicación de algunos temas específicos que son abordados en su formación universitaria.

Encargados de la serie:

Bustamante Romaní, Rafael.
rbustamanter@unmsm.edu.pe

Cisneros García, Juan Manuel.
jcisnerosg@unmsm.edu.pe

Facultad de Ciencias Económicas.
Universidad Nacional Mayor de San Marcos.
Calle Germán Amézaga N° 375.
Ciudad Universitaria, Lima 1. Perú.

La **Serie Apuntes de Clase ΩBΓ** es promovida y desarrollada por un colectivo de docentes del Departamento de Economía de la Universidad Nacional Mayor de San Marcos.
El contenido de cada publicación es íntegramente responsabilidad de cada autor, no representa necesariamente los puntos de vista de los integrantes del colectivo, ni de la Universidad.

Modelos de elección Discreta en Stata 14.

Rafael Bustamante◊

Resumen

Los modelos de elección discreta se han desarrollado en los últimos tiempos dentro de la rama de la Microeconometria. Estas notas son una introducción a los modelos de elección binario, con aplicaciones a Stata 14.0. Asimismo se presentan todas las herramientas en proceso de estimación y análisis de indicadores de bondad de ajuste, parsimonia y pruebas para detectar los vicios econométricos.

Palabras claves: Modelos Logit, Probit, efecto marginal, Stata 14

Clasificación JEL: E12, E62.

◊ Estudios concluidos de Doctorado en Economía con mención en los Recursos Naturales (c), Universidad Nacional Autónoma de México. MBA Gerencial, CENTRUM Pontificia Universidad Católica del Perú. Maestría en Economía con mención en Finanzas, Universidad Nacional Mayor de San Marcos. B. Sc. Economía, UNMSM. Profesor Auxiliar del Departamento de Economía de la UNMSM. Investigador asociado al Instituto de Investigaciones FCE - UNMSM. Contacto: rbustamanter@unmsm.edu.pe.

1. ESTIMACIÓN Y ANÁLISIS

Las estimaciones lineales clásicas permiten la modelización de variables dependientes cuantitativas para identificar relaciones estadísticas en las que se asume una serie de supuestos sobre la forma del error de la ecuación lineal (homocedasticidad, normalidad, etc.). Sin embargo, en muchos contextos, el fenómeno que se quiere modelizar no es continuo sino discreto, por ejemplo cuando se quiere modelar la elección de compra de un bien o servicio; o la decisión de participar o no en el mercado laboral. Estos son los modelos conocidos como modelos de respuesta cualitativa. Llamamos variables cualitativas a aquellas que no aparecen en forma numérica, sino como categorías o atributos como por ejemplo, el sexo o la profesión de una persona. En general, se dice que una variable es discreta cuando está formada por un número finito de alternativas que miden cualidades (Del Carpio Gonzales, 2008).

1.1. INTERPRETACIÓN ESTRUCTURAL

Existen tres enfoques para la interpretación estructural de los modelos de elección discreta. El primero hace referencia a la modelización de una variable latente a través de una función índice, que trata de modelizar una variable inobservable o latente. El segundo de los enfoques permite interpretar los modelos de elección discreta bajo la teoría de la utilidad aleatoria, de tal manera que la alternativa seleccionada en cada caso será aquella que maximice la utilidad esperada. El tercero pasa por plantear un modelo de probabilidad no lineal.

Bajo el primero de los enfoques se trata de modelizar una variable índice, inobservable o latente no limitada en su rango de variación y*. Cuando la variable latente supera un determinado nivel, la variable discreta toma el valor 1, y si no lo supera toma el valor 0. La variable latente depende de un conjunto de variables

explicativas que generan las alternativas que se dan en la realidad y que permiten expresar el modelo dicotómico como (Abanto Orihuela, 2010):

$$Y = \begin{cases} 1, & \text{si } Y^* > 0 \\ 0, & \text{si } Y^* < 0 \end{cases} \qquad \textbf{1.}$$

Donde el supuesto sobre la distribución de error determina el tipo de modelo a estimar. Si se supone una función de distribución uniforme, se utiliza el Modelo Lineal de Probabilidad truncado; si se distribuye como una normal con media cero y varianza uno, el modelo generado será un Probit; mientras que si se supone que se distribuye como una curva logística, se trataría de un modelo Logit. La hipótesis de que el umbral a superar por la variable latente sea cero se puede modificar por cualquier otro valor sugiriéndose, en determinados estudios, que el valor crítico sea el definido por el término constante. Bajo este enfoque, el modelo probabilístico quedaría (Abanto Orihuela, 2010):

$$\begin{aligned} &Y^* = X\beta + \varepsilon \\ &\Pr(Y = 1/X) = \Pr(Y^* > 0/X) = \Pr(\varepsilon > -(X\beta)/X) \\ &\Pr(Y = 1/X) = F(X\beta) \end{aligned} \qquad \textbf{2.}$$

Con el modelo así definido, la variable endógena del modelo dicotómico representa la probabilidad de ocurrencia del fenómeno analizado, siendo la probabilidad de que ocurra la opción 1 más elevada cuando mayor sea el valor de Y*.

El segundo de los enfoques para la interpretación de los modelos de respuesta dicotómica es el que hace referencia a la modelización a través de la formulación de una utilidad aleatoria. Bajo este enfoque un individuo debe adoptar una decisión que le permita elegir entre dos alternativas excluyentes, la 1 o la de 0, lo que hará maximizando la utilidad esperada que le proporciona cada una de las alternativas posibles sobre las que tiene que decidir. Es decir, el individuo i-ésimo elegirá una de las dos alternativas dependiendo de que la utilidad que le proporciona dicha decisión sea superior a la que le proporciona su complementaria (Abanto Orihuela, 2010).

La formulación del modelo bajo esta teoría parte del supuesto de que la utilidad derivada de una elección, Ui0o Ui1, es función de las variables explicativas de dicha decisión, que son las características propias de cada una de las alternativas de elección y las características personales propias del individuo, de manera que suponiendo linealidad en las funciones, se tiene (Abanto Orihuela, 2010):

$$U_{i0} = \alpha + X_{i0}\beta + \varepsilon_{i0}$$
$$U_{i1} = \alpha + X_{i1}\beta + \varepsilon_{i1} \qquad \textbf{3.}$$

Donde los eij recogen las desviaciones que los agentes tienen respecto a lo que sería el comportamiento del agente medio y que se debe a factores aleatorios. El agente i elegirá la opción 1 si la utilidad de esa decisión supera la de la opción 0 y viceversa, de manera:

$$Y = \begin{cases} 1, & \text{si } U_{i1} > U_{i0} \\ 0, & \text{si } U_{i1} < U_{i0} \end{cases} \qquad \textbf{4.}$$

Y el modelo dicotómico quedaría definido por:

$$\Pr(\mathrm{Y} = 1/\mathrm{X}) = \Pr(U_{i1} > U_{i0}/\mathrm{X}) = \Pr(\varepsilon_{i1} - \varepsilon_{i0}/(\beta\,\mathrm{X})/\mathrm{X}))$$
$$\Pr(\mathrm{Y} = 1/\mathrm{X}) = F(\beta\,\mathrm{X}) \qquad \textbf{5.}$$

Según que la función asociada a la perturbación aleatoria ε_{i1} (que será la función de distribución, $F(\beta\,\mathrm{X})$, que se suponga siga dicha probabilidad), sea una función de distribución uniforme, la función de distribución de la normal tipificada o la de la curva logística, se obtienen el Modelo Lineal de Probabilidad Truncado, el Probit o el Logit, respectivamente. El tercer enfoque pasa por estructurar un modelo de probabilidad no lineal, como lo sugiere Theil -1970, de tal manera que:

$$\Pr(\mathrm{Y}_i = 1/\mathrm{X}) = \mathrm{M}_i = \frac{\exp(\mathrm{X}_i\,\beta)}{1+\exp(\mathrm{X}_i\,\beta)}$$
$$\Lambda(\mathrm{X}_i) = \frac{\Pr(\mathrm{Y}_i = 1/\mathrm{X}_i)}{\Pr(\mathrm{Y}_i = 0/\mathrm{X}_i)} = \exp(\mathrm{X}_i\,\beta) \qquad \textbf{6.}$$
$$\ln\Lambda(\mathrm{X}_i) + \varepsilon_i = \mathrm{X}_i\,\beta + \varepsilon_i$$

Es decir medir que tan a menudo ocurre algo (Y=1), respecto a que tan a menudo no ocurre (Y=0).

1.1.2. MODELO DE PROBABILIDAD LINEAL

La primera alternativa teórica desarrollada para estudiar modelos con variables dicótomas se planteó como una extensión del modelo lineal general:

$$Y_i = X_i \beta + \varepsilon_i$$

$$Y = \begin{cases} 1, & \text{si ocurre el evento} \\ 0, & \text{si no ocurre el evento} \end{cases} \qquad \textbf{7.}$$

$$\varepsilon_i \approx N(0, \sigma_\varepsilon^2)$$

En general, la distribución de los modelos de elección binaria se caracteriza por establecer una nube de puntos de tal manera que las observaciones se dividen en dos subgrupos. Uno de ellos está conformado por las observaciones en las que ocurrió el acontecimiento objeto de estudio (Yi =1), y el otro, por los puntos muéstrales en los que no ocurrió (Yi =0). Para el desarrollo de los modelos de elección discreta se utilizará la base de datos "labora.dta"[1].

```
use labora.dta, clear
```

Antes de desarrollar el modelo de probabilidad lineal, es posible obtener una descripción rápida de la base de datos a utilizar, el comando describe mostrará el tipo de información con la que se cuenta. Esta base de datos hipotética contiene 400 observaciones en las que se detalla si el postulante es admitido a un programa de Post Grado (admit), el puntaje obtenido en la prueba Graduate Record Exam (gre), el puntaje obtenido en el pregrado (Grade Point Average, gpa) y finalmente se considera si el postulante proviene de una universidad de prestigio o no (topnotch). Seguidamente se procederá a estimar la regresión lineal en donde la variable dependiente admit esta explicada por el puntaje obtenido en el gpa (Abanto Orihuela, 2010).

[1] Puede acceder a la data a través del siguiente enlace
https://drive.google.com/file/d/0B4B7bhYQMcKmTy1hSVBPaUxRMU0/view?usp=sharing

```
. d

Contains data from C:\Users\Rafael\Desktop\material Stata\stata Avanzado\Reto1\labora.dta
  obs:           400
 vars:             4                          29 Jan 2007 11:20
 size:         3,200
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
admit           byte    %8.0g                 admitido_al_post
gre             int     %8.0g                 examen de graduados
topnotch        byte    %8.0g                 prestigo univ
gpa             float   %9.0g                 promedio académico
-------------------------------------------------------------------------------
Sorted by:
```

Luego procedemos a realizar la estimación MCO.

regress admit gpa

```
. regress admit gpa

      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F(  1,   398) =   81.14
       Model |  15.0345865     1  15.0345865            Prob > F      =  0.0000
    Residual |  73.7429135   398  .185283702           R-squared     =  0.1694
-------------+------------------------------           Adj R-squared =  0.1673
       Total |     88.7775   399       .2225           Root MSE      =  .43045

------------------------------------------------------------------------------
       admit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |   .3074274   .0341284     9.01   0.000      .240333    .3745218
       _cons |  -.6440456   .1105248    -5.83   0.000    -.8613309   -.4267603
------------------------------------------------------------------------------
```

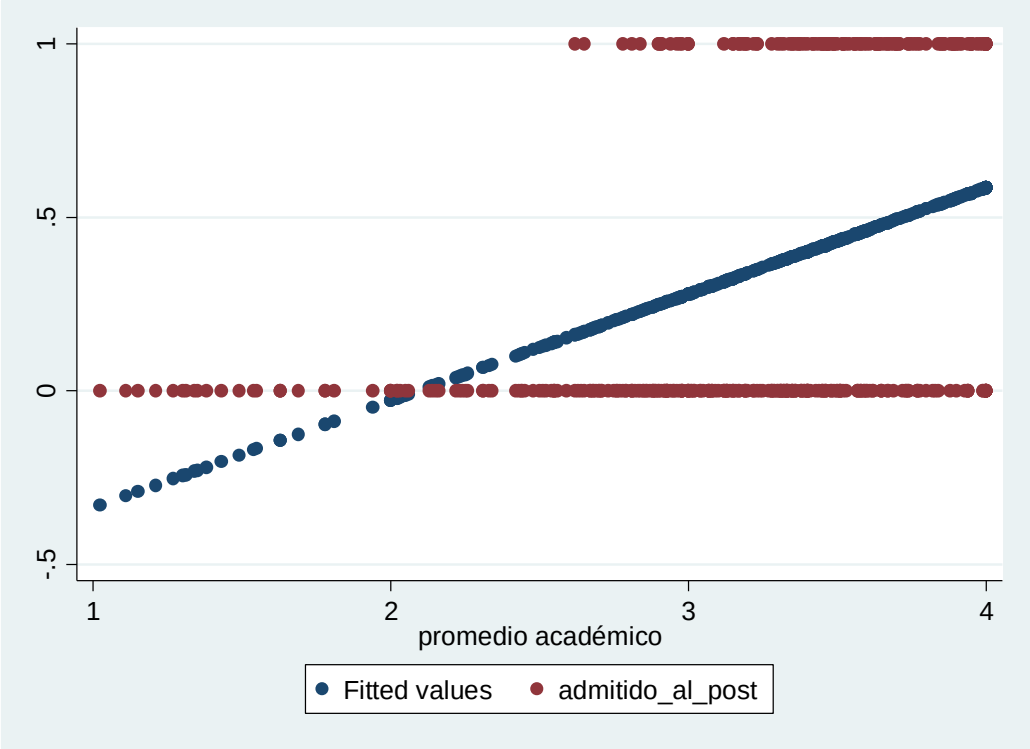
1.1.3. PROBLEMAS CON ESTA ESTIMACIÓN

La interpretación de los coeficientes en los modelos de probabilidad es similar a la de los modelos de regresión lineal, en donde el valor de los parámetros recoge el efecto de una variación unitaria en cada una de las variables explicativas sobre la probabilidad de ocurrencia del acontecimiento objeto de estudio, sin embargo, el MPL presenta algunas inconsistencias.

Se puede apreciar en el modelo inicial que algunos de los valores estimados se encuentran fuera de rango, lo cual carece de lógica considerando que deben

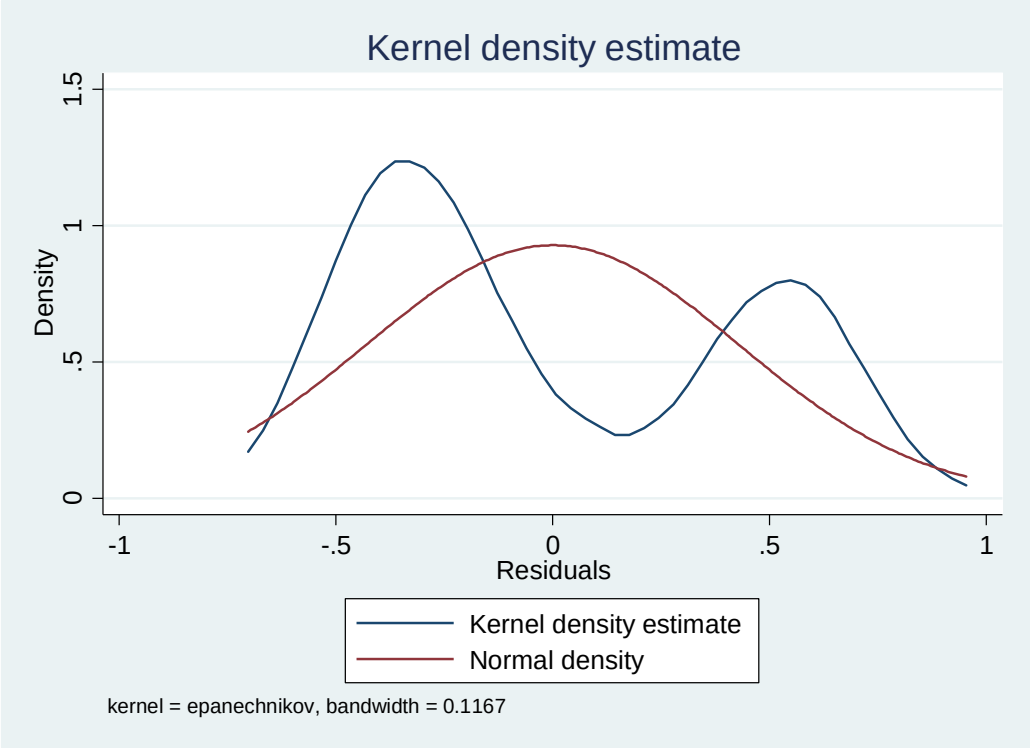
interpretarse como probabilidades.

tw sc y admit gpa



1.1.4 MODELO DE PROBABILIDAD TRUNCADA

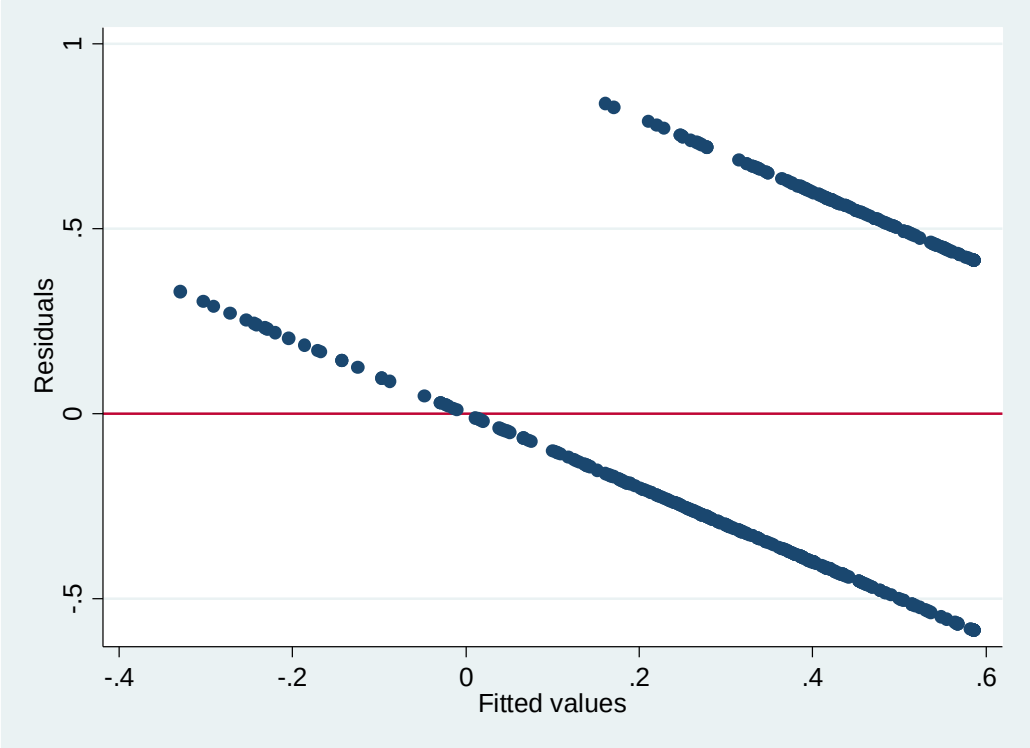
A través del gráfico de la densidad de Kernel para el modelo que incluye todas las variables, se observa que los residuos no se distribuyen de manera normal, por lo tanto no es eficiente, es decir, pueden presentarse problemas de minimización de la varianza a medida que la muestra aumenta (Abanto Orihuela, 2010).
kdensity r, normal



Problemas de Heterocedasticidad. Aún en el caso de que se cumpliesen las hipótesis de media y correlación nula en la perturbación aleatoria E(ei) = 0 E(ei, ej) = 0 para todo i = j, no se cumple la hipótesis de varianza constante, es decir, la perturbación aleatoria no es homocedástica. En STATA es posible realizar un análisis tanto gráfico como a través de números índice para verificar la presencia de heterocedasticidad.

rvfplot, yline (0)

hettest

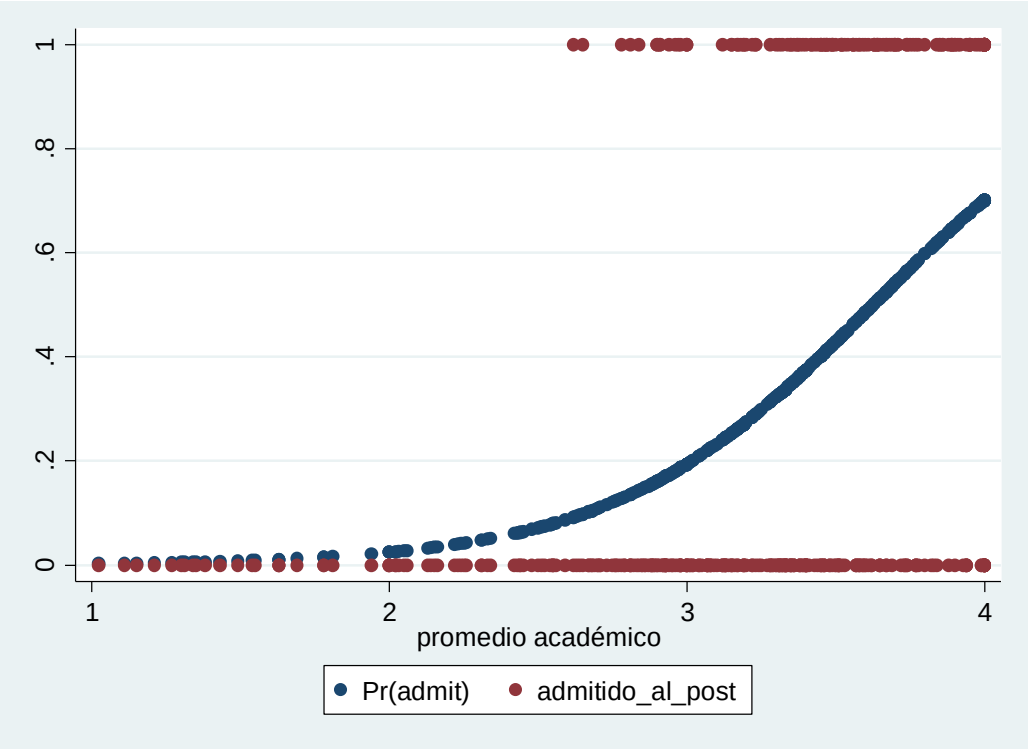


Para el presente ejemplo la hipótesis nula de varianza constante (homo-cedasticidad) será rechazada debido a que el p value de la distribución del estadístico chicuadrado es muy pequeño, aceptándose la hipótesis alterna de varianza no homogénea.

2. MCG VERSUS MCP

Los problemas en la interpretación y estimación de los parámetros del modelo de probabilidad lineal han llevado a la búsqueda de modelos alternativos que permitan estimaciones más confiables de las variables dicótomas. Es el caso de los modelos de probabilidad no lineal, donde la función de especificación utilizada garantiza un resultado en la estimación comprendido en el rango 0-1. Estos son los modelos logit y probit. Analizaremos a continuación los datos a través de una regresión logística, la cual se formula a continuación.

```
logit admit gpa
predict l
tw sc l admit gpa
```



3. POST -ESTIMACIÓN

3.1. TEST DE EFECTOS INDIVIDUALES

Si los supuestos bases del modelo se sostienen, los estimadores son distribuidos de manera asintótica y normal:

$$\overset{\Lambda}{\beta}_k \rightarrow N(\beta_k, \sigma^2_{\beta_k})$$

Donde la hipótesis nula de significancia estadística del parámetro puede ser testeada a partir del siguiente estadístico asintótico.

$$Z = \frac{\overset{\Lambda}{\beta}_k - \beta^*}{\sigma^2_{\beta_k}}$$

Si la hipótesis nula es verdadera entonces z se distribuirá aproximadamente como una normal con media cero y varianza unitaria para muestras grandes.

3.2 TEST DE WALD

Podemos analizar el modelo una vez estimado, mediante un testeo de hipótesis que validen una correcta especificación. Para esto el test de Wald calculado para hipótesis lineales sobre los parámetros de los modelos estimados nos será de mucha utilidad. También puede usarse el test bajo una estructura no lineal, la cual no abordaremos en esta sección.

```
logit admit gre gpa topnotch
test gpa=0
test gre=gpa, accumulate
```

3.3 TEST LR

El estadístico de verosimilitud también nos será de gran utilidad para evaluar mediante hipótesis la significancia de modelos. Este estadístico compara modelos anidados.

```
logit admit gre gpa topnotch
lrtest, saving(0)
logit admit gre gpa
lrtest
```

```
. lrtest, saving(0)

. logit admit gre gpa

Iteration 0:   log likelihood = -254.37399
Iteration 1:   log likelihood = -196.01023
Iteration 2:   log likelihood = -192.17149
Iteration 3:   log likelihood = -192.13262
Iteration 4:   log likelihood = -192.13259
Iteration 5:   log likelihood = -192.13259

Logistic regression                             Number of obs   =        400
                                                LR chi2(2)      =     124.48
                                                Prob > chi2     =     0.0000
Log likelihood = -192.13259                     Pseudo R2       =     0.2447

------------------------------------------------------------------------------
       admit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gre |   .0072571    .001305     5.56   0.000     .0046994    .0098149
         gpa |   1.727508   .2942894     5.87   0.000     1.150712    2.304305
       _cons |  -10.80411   1.179211    -9.16   0.000    -13.11532   -8.492897
------------------------------------------------------------------------------

. lrtest
You ran lrtest using the old syntax.  Click here to learn about the new syntax.

Likelihood-ratio test                                 LR chi2(1)  =      5.74
(Assumption: . nested in LRTEST_0)                    Prob > chi2 =    0.0166

.
end of do-file
```

Donde nuestra hipótesis nula es H0 = βtopnotch = 0

```
logit admit gre gpa topnotch
lrtest, saving(M1)
logit admit gre gpa
lrtest, using(M1)
```

```
. lrtest, saving(M1)

. logit admit gre gpa

Iteration 0:   log likelihood = -254.37399
Iteration 1:   log likelihood = -196.01023
Iteration 2:   log likelihood = -192.17149
Iteration 3:   log likelihood = -192.13262
Iteration 4:   log likelihood = -192.13259
Iteration 5:   log likelihood = -192.13259

Logistic regression                             Number of obs   =        400
                                                LR chi2(2)      =     124.48
                                                Prob > chi2     =     0.0000
Log likelihood = -192.13259                     Pseudo R2       =     0.2447

------------------------------------------------------------------------------
       admit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gre |   .0072571    .001305     5.56   0.000     .0046994    .0098149
         gpa |   1.727508   .2942894     5.87   0.000     1.150712    2.304305
       _cons |  -10.80411   1.179211    -9.16   0.000    -13.11532   -8.492897
------------------------------------------------------------------------------

. lrtest, using(M1)
You ran lrtest using the old syntax.  Click here to learn about the new syntax.


Likelihood-ratio test                                 LR chi2(1)  =      5.74
(Assumption: . nested in LRTEST_M1)                   Prob > chi2 =    0.0166

.
end of do-file

.
```

```
logit admit gre gpa topnotch
lrtest, saving(0)
logit admit gre gpa
lrtest, saving(1)
lrtest, using(1) model(0)
```

```
. logit admit gre gpa topnotch

Iteration 0:   log likelihood = -254.37399
Iteration 1:   log likelihood = -192.99747
Iteration 2:   log likelihood = -189.30142
Iteration 3:   log likelihood = -189.26321
Iteration 4:   log likelihood = -189.26319
Iteration 5:   log likelihood = -189.26319

Logistic regression                               Number of obs   =        400
                                                  LR chi2(3)      =     130.22
                                                  Prob > chi2     =     0.0000
Log likelihood = -189.26319                       Pseudo R2       =     0.2560

------------------------------------------------------------------------------
       admit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gre |   .0067809    .001324     5.12   0.000      .004186    .0093758
         gpa |   1.597449   .2935206     5.44   0.000     1.022159    2.172738
    topnotch |   .7623019   .3197943     2.38   0.017     .1355167    1.389087
       _cons |  -10.23092    1.18238    -8.65   0.000    -12.54835   -7.913501
------------------------------------------------------------------------------

. lrtest, saving(0)

. logit admit gre gpa

Iteration 0:   log likelihood = -254.37399
Iteration 1:   log likelihood = -196.01023
Iteration 2:   log likelihood = -192.17149
Iteration 3:   log likelihood = -192.13262
Iteration 4:   log likelihood = -192.13259
Iteration 5:   log likelihood = -192.13259

Logistic regression                               Number of obs   =        400
                                                  LR chi2(2)      =     124.48
                                                  Prob > chi2     =     0.0000
Log likelihood = -192.13259                       Pseudo R2       =     0.2447

------------------------------------------------------------------------------
       admit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gre |   .0072571    .001305     5.56   0.000     .0046994    .0098149
         gpa |   1.727508   .2942894     5.87   0.000     1.150712    2.304305
       _cons |  -10.80411   1.179211    -9.16   0.000    -13.11532   -8.492897
------------------------------------------------------------------------------

. lrtest, saving(1)

. lrtest, using(1) model(0)
You ran lrtest using the old syntax.  Click here to learn about the new syntax.

Likelihood-ratio test                                 LR chi2(1)  =      5.74
(Assumption: LRTEST_1 nested in LRTEST_0)             Prob > chi2 =    0.0166

.
end of do-file
```

Muchas medidas escalares han sido desarrolladas para resumir las bondades de ajuste de modelos de regresión continuo o de variables categóricas. Sin embargo no hay evidencia convincente de selección de un modelo que maximice los valores de una medida comparada con la medida de otro modelo. Mientras las medidas de ajuste proveen información, esta es solo parcial, que debería ser sostenida con una teoría económica razonable, o investigaciones anteriores como referencia. El comando Fifstat nos permite obtener una tabla con estadísticos que ayudaran a evaluar la bondad de ajuste del modelo. De los cuales analizaremos algunos (Del Carpio Gonzales, 2008).

3.4 FITSTAT

A continuación proveeremos de una breve descripción de cada una de las medidas que computa el "fitstat". Mayores detalles de las medidas las podemos encontrar en Long(1997). Stata comienza su análisis maximizando iteracciones de verosimilitud y calculando sus logaritmos, para determinado modelo, con todos los parámetros excepto el intercepto en un nivel de cero L(Mintercepto), mientras que cuando los parámetros son diferentes de cero, el logaritmo de verosimilitud calculado será Un test LR donde la hipótesis nula de que todos los coeficientes excepto el intercepto son ceros puede ser calculado comparando el logaritmo de verosimilitud $LR = 2\left[Ln\left(Mfull\right) - Ln\left(Mintercepto\right)\right]$ El LR es reportado por Stata como chi2(GL), donde los GL son el número de parámetros restringidos (Abanto Orihuela, 2010).

Bibliografía

Abanto Orihuela, J. (2010). *Stata Avanzado aplicado a la Investigación Económica.* Lima: Grupo IDDEAS.

Del Carpio Gonzales, J. (2008). *Manual de Stata.* Bolivía, La Paz.