Modelos con Variable Dependiente Discreta

Econometría II

Alarcón Castillo Henry
Champa Del Valle Katherine
Mayhuasca Gutierrez Victor
Bautista Ramos Luis

Índice general

Parte Uno

1. Introducción

El presengte trabajo tiene como objetivo dar a conocer las bondades de los modelos Logit y probit dentro del campo de la estimación de modelos con variable endógena discreta dicotómica.

Estos modelos surgen porque en situaciones en que la variable endógena es discreta y asume un pequeño número de valores, no tiene sentido tratarla como una variable aproximadamente continua. Por sí misma, la discrecionalidad de la variable endógema no significa que los modelos lineales sean inapropiados. No obstante, el modelo de probabilidad lineal tiene ciertas desventajas. Los modelos logit y probit, superan las desventajas del Modelo de Probabilidad Lineal (MPL); la desventaja es que son más difíciles de interpretar.

Existen numerosos tipos diferentes que se aplican en diferentes situaciones. Lo que tienen en común es que son modelos en los que la variable dependiente es un indicador de una elección discreta, como un "sí o no"decisión. En general, los métodos de regresión convencionales no son adecuadas en estos casos.

En la mayoría de los casos, el método de estimación es de máxima verosimilitud. Existen diversas propiedades de los estimadores de máxima verosimilitud. Para el desarrollo de este libro, se asumirá que se cumplan las condiciones necesarias detrás de las propiedades de optimalidad de los estimadores de máxima verosimilitud.

Además, se desarrollará el modelo con datos de Conteo. Para datos de conteo se suele utilizar la distribución Poisson como componente aleatorio en el proceso de ajuste de un modelo lineal generalizado. Esta distribución se caracteriza por la igualdad entre su media y su variancia, supuesto difícil de verificar ya que en la práctica las observaciones de conteos frecuentemente exhiben una variabilidad que excede la supuesta para una variable del tipo Poisson. El fenómeno por el cual un modelo lineal generalizado tiene mayor variabilidad que la presupuesta por el componente aleatorio del mismo se denomina sobredispersión.

2. Modelos de Elección Binaria

2.1 Modelo Logit

2.1.1 Introducción

En el siguiente capítulo se dará a conocer las bondades del modelo Logit dentro del campo de la estimación de modelos con variable endógena discreta dicotómica. Si bien hemos trabajado hasta ahora con variables discretas en nuestras estimaciones , éstas solo se han comportado como variables exógenas, es decir, han sido tratadas como variables independietes que tratan de explicar a otra variable, dejando de lado la posibilidad de ser modeladas como variables endógenas.

Es preciso entonces, abordar un nuevo tema: modelos con variable endógena discreta. En este caso, los modelos lineales convencionales trabajados hasta ahora ya no son válidos y tampoco la estimación por Mínimos Cuadrados Ordinarios (MCO), por lo que introduciremos un modelo nuevo para tales estimaciones. Es conveniente recalcar que esta variable endógena puede ser discreta dicotómica, discreta sin orden o discretas ordenadas.

De acuerdo a la forma de la variable endógena, (entre los tres mencionados anteriormente) el modelo tiene un tratamiento especial. Centrándonos en el presente trabajo, se pasará a decribir el caso especial de los modelos con variable endógena discreta dicotómica. En un modelo de respuesta binaria, el interés yace principalmente en conocer la probabilidad de respuesta.

Por sí misma, la discrecionalidad de la variable endógena no significa que los modelos de probabilidad lineal (MPL) sean inapropiados. Estimar y utilizar el modelo de probabilidad lineal es simple, pero tiene algunas desventajas. Las dos desventajas más importantes son que las probabilidades ajustadas pueden ser menores que cero o mayores que uno y el efecto parcial de cualquier variable explicativa (si aparece en la ecuación en su nivel) es constante. Estas limitaciones del MPL pueden superarse si se usan modelos de respuesta binaria más sofisticados. Entre ellos el modelo Logit.

2.1.2 Motivación

Los modelos Logit se comportan como una herramienta científica avanzada, genera instrumentos y procedimientos que permitirán validar, mejorar y actualizar los procesos estadísticos.

Los modelos de elección cualitativa son muy útiles y muy utilizados en la economía, porque muchas decisiones pueden ser tomadas a partir de simples respuestas como un sí o un no, podemos mencionar por ejemplo la decisión de una empresa si va decidir retribuir servicio de sus utilidades a sus accionistas o no, votar por un político o no, si un individuo viene a trabajar o no. Estos son distintos casos de los modelos tradicionales. El objetivo de los modelos de elección cualitativa es encontrar la probabilidad de que algo ocurra; por ello los modelos de elección cualitativa son también conocidos como modelos de probabilidad.

2.1.3 Descripción Teórica del Modelo

Los modelos Logit son de respuesta binaria (0 y 1) se usan como un instrumento recomendable para calcular la probabilidad de respuesta, indicando la construcción y forma del modelo y el análisis de algunos estadísticos requeridos.

La modelización Logit es similar a la regresión tradicional salvo que utiliza como función de estimación a la función logística en lugar de utilizar a la lineal. Con la modelización Logit, el resultado del modelo es la estimación de la probabilidad de que un nuevo individuo pertenezca a un grupo o a otro (probabilidad de éxito o fracaso, si o no, etc.). Además, al tratarse de un análisis de regresión, también es posible identificar las variables más importantes que explican las diferencias entre grupos.

$$P(y = 1/x) = P(y = 1/x_1, x_2, ..., x_k) \tag{2.1.1}$$

donde x denota el conjunto total de variables explicativas. En el MPL, se supone que la probabilidad de respuesta es lineal en un conjunto de parámetros β_k. Para evitar las limitaciones del MPL, considere una clase de modelos de respuesta binaria de la forma:

$$P(y = 1/x) = F(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k) = F(x\beta) \tag{2.1.2}$$

donde F es una función que asume valores estrictamente entre cero y uno, para todos los números reales z. Esto asegura que las probabilidades de respuesta estimada están estrictamente entre cero y uno. La función F, entre las muchas sugeridas, es la función logística, cuya representación es:

$$F(x\beta) = \Lambda(z) = \frac{e^{x\beta}}{1 + e^{x\beta}} \tag{2.1.3}$$

que está entre cero y uno para todos los números reales z. Esta es la función de distribución acumulada (fda) para una variable aleatoria logística estándar. La función logística es creciente, y aumenta con más rapidez en z = 0. El comportamiento de la función es el siguiente: F(z) $\rightarrow$ 0 a medida que z $\rightarrow -\infty$, y F(z)$\rightarrow$1 a medida que z$\rightarrow \infty$. (Ver gráfica en **Anexo1**).

2.1.4 Definición Matemática

El modelo Logit puede derivarse a partir de un modelo de variable latente subyacente. Sea y* una variable inobservable, o latente, determinada por:

$$y^* = \beta_0 + x\beta + e, y = 1[y^* > 0] \tag{2.1.4}$$

donde se introduce la notación 1[.] para definir un resultado binario. La función 1[.] recibe el nombre de función de indicador, que asume el valor de uno si el evento dentro de los corchetes es verdadero y de cero si no lo es. Por tanto, y es uno si y* > 0 y y es cero si y* ≤ 0.

Bajo el supuesto que "x" es independiente de "e" y que este último tiene la distribución logística estándar, "e" se distribuye simétricamente en torno a cero, lo cual significa que 1 - F(-z) = F(z) para todos los números reales z. A partir de (3.4) y de los supuestos establecidos al inicio del párrafo, es posible calcular la probabilidad de respuesta para y:

$$\begin{aligned} P(y=1/x) &= P(y* > 0/x) = P[x\beta + e > 0/x] = P[e > -(\beta_0 + x\beta)/x] \\ &= 1 - F[-(\beta_0 + x\beta)] = F(\beta_0 + x\beta) \end{aligned} \tag{2.1.5}$$

2.1.5 Impacto marginal

Como en todo modelo de estimación, el objetivo principal del modelo Logit es explicar los efectos de las x_j sobre la probabilidad de respuesta P(y =1/x). La formulación de la variable latente tiende a dar la impresión de que lo que principalmente interesa son los efectos de cada x_j sobre y*. Pero la variable latente y* rara vez tiene una unidad de medición bien definida. (Por ejemplo, y* puede ser la diferencia en niveles de utilidad de dos acciones diferentes.) Por tanto, las magnitudes de cada β_k no son, por sí mismas, especialmente útiles en contraste con el modelo de probabilidad lineal.

Para la mayoría de los propósitos, se quiere estimar el efecto de x_j sobre la probabilidad de éxito P(y =1/x), pero esto se complica por la naturaleza no lineal de la función logística. Para hallar el efecto parcial de las variables aproximadamente continuas sobre la probabilidad de respuesta, es necesario recurrir al cálculo. Si x_j es una variable aproximadamente continua, su efecto parcial sobre p(x) = P(y = 1/x) se obtiene de la derivada parcial:

$$\frac{\partial p(x)}{\partial x_j} = \frac{\partial F(x\beta)}{\partial x} = \frac{\partial F(x\beta)}{\partial x\beta}\frac{\partial x\beta}{\partial \beta} = f(\bar{x}\beta)\beta_j \tag{2.1.6}$$

Ahora, si por ejemplo, x_j es una variable explicativa binaria discreta, entonces el efecto parcial de cambiar x_j de cero a uno, manteniendo todas las demás variables fijas, simplemente es:

$$\begin{aligned} \frac{\Delta P(y=1/x)}{\Delta x_j} &= P(y=1/x_j=1) - P(y=1/x_j=0) \\ &= F(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k / x_j = 1) - F(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k / x_j = 0) \end{aligned} \tag{2.1.7}$$

2.2 Modelo Probit

Los Modelo Probit son aquellos que pertenecen a la clase de modelos de respuesta binaria, es decir, la variable dependiente es una variable dicotómica, donde toma 1 para indicar el éxito en la variable de análisis y 0 en el caso de no ser así.
Por ejemplo se asume una variable observada (latente) que debe traspasar un umbral para que la variable dependiente tome el valor de 1,la estimación d estos modelos no puede ser realizada por MCO (Mínimos cuadrados ordinarios)ya que la variable dependiente es inobservable por lo que se recurre al uso de Máxima Verosimilitud haciendo supuestos sobre la distribución de los errores.Cuando los errores se consideran distribuidos de manera normal, entones se obtiene un Modelo Probit .

Con esta especificación,la variable dependiente dicotómica tiene la probabilidad de 2 opciones Pr(y=1/x) o la Pr (y=0/x) que dependen de los valores que toman las variables de control especificadas como las variables sociodemográficas, socioeconómicas representadas mediante una combinación lineal ($x_i\beta$).El modelo se especifica de la siguiente forma :

$$P(y=1/x) = Pr(y^* > 0) = F(x\beta) \tag{2.2.1}$$

Si definimos el modelo de la siguiente manera:

$$P(y=1/x) = G(\beta_0 + x_1\beta_1 + ... + x_K\beta_K) = G(\beta_0 + x\beta) \tag{2.2.2}$$

donde G es una funcion que adopta valores entre cero y uno para todos los numeros reales Z,donde G representa la funcion de distribucion acumulativa.
Debido a que el modelo Probit es un modelo de vaiable dependiente limitada,la estimacion de parametros se hace por el metodo de Maxima Verosimilitud.Este modelo sugiere que se elijan como estimados los valores de los parametros que maximizen el logaritmo de la funcion de verosimilitud. La funcion logaritmica de verosimilitud para la observacion i se define como:

$$\lambda(\beta) = yilog(G(Xi\beta)) + (1-yi)log(1-G(Xi\beta)) \tag{2.2.3}$$

El logaritmo de la funcion de verosimilitud para una muestra de tamano n se define como:

$$L = \sum_{i=1}^{n} \lambda(\beta) \tag{2.2.4}$$

El estimador de maxima verosimilitud de β,denotado por β que maximize el logaritmo de verosimilitud.Las propiedades de los estimadores de maxima verosimiltud del modelo son conistentes,asintoticamente normales y asintoticamente eficientes.
Ahora conociendo los efectos de los cambios en las variables explicativas sobre las probabilidades de que cualquier observaion perteneza a uno de los 2 grupos (y=0,y=1) se emplea una derivada parial definida como:

$$\frac{\partial x}{\partial xj} = g(\beta 0 + X\beta)\beta \tag{2.2.5}$$

El termino g(z) corresponde a una funcion de densidad de probabilidad.Dado que en el modelo Probit G(.) es una funcion de distribucion acumulativa estrictamente positiva,g(z)>0 para toda Z,el signo del efecto parcial es el mismo que el de β.
Ahora para probar la significania de cada uno de los coeficientes estimados se lleva a cabo la prueba hipotesis Ho :β=0,con un t estadistico.Para probar la significancia de variables conjuntamente existen diferentes estadisticos como el estadistico Wald y el estadistico de la razon de verosimilitud entre otros. En estos 2 casos se emplea una distribucion chi cuadrado.
Mediante un caso practico analizaremos ambos modelos e interpretaremos los resultados Estimamos en Stata el siguiente modelo para la probabilidad de estar desempleado en Colombia en función de la edad, el genero, la situacion marital, la educacion, el ingreso no laboral y la localizacion geografica.

. probit desocupado edad mujer soltero educ jefe inla caba Ver resultados en **Anexo2**.

A diferencia de los modelos de Mínimos Cuadrados Ordinarios (MCO), estos modelos tienen que ser interpretados cuidadosamente.Empezando que los valores de estos coeficientes no tienen una interpretación cuantitativa (solo es interpretable el signo de los mismos).A la vez analizaremos los efectos marginales de cada variable para realizar una interpretación cuantitativa del efecto de cada variable sobre la probabilidad de estar desocupado.

Interpretando cuantitativamente cada uno de los efectos marginales.Las variables explicativas que son continuas:

.La interpretación del valor -0.0020344, que corresponde al efecto marginal de la variable años de educación (educ) donde para una persona con las características consideradas un aumento en un año de educación provoca un cambio en la probabilidad predicha de -0.0020344, es decir, las 2 probabilidades de estar desocupado se reduciría en 0.203 puntos porcentuales (-0.0020344*100),dado todo lo demás constante. .La interpretación para el efecto marginal de la variable edad es equivalente. Para una persona con las características consideradas, un aumento en un año de edad reduce la probabilidad predicha de estar desempleado en 0.022 puntos porcentuales (-0.0002215*100), ceteris paribus.

Para el caso del efecto marginal de las variables dummies (como mujer, soltero, jefe y caba) recuerden que se computan de diferente manera pero se interpreta de manera equivalente.
• El hecho de ser jefe de hogar, para un hombre casado que es jefe de familia, con 17 años de educación, edad e ingreso no laboral promedio y que resida en la CABA, reduce su probabilidad predicha de estar desempleada en 1.87 puntos porcentuales (-0.0187869*100). • De la misma forma, el hecho de residir en CABA, dado todo lo demás, reduce su probabilidad predicha de estar desempleada en 0.19 puntos porcentuales (-.0019124*100).

Como notarán, se ha hecho énfasis en aclarar que en el caso de los modelos de elección binaria si se multiplica por 100 al efecto marginal, se está midiendo el efecto del cambio en una unidad de X sobre la probabilidad predicha. Ese cambio es en puntos porcentuales y no en tanto por ciento.En el primer caso se usa para indicar un cambio marginal, mientras que el segundo se aplica cuando se trata de cambios proporcionales. Por ejemplo, según se muestra en la segunda salida de Stata, la probabilidad de desempleo para un hombre casado que es jefe de familia, con 17 años de educación, edad e ingreso no laboral promedio y que resida en la CABA es de 0.02056653 (es decir, 2 por ciento de probabilidad). Dijimos que el efecto marginal de la educación (educ) para este caso es de 0.20 puntos porcentuales, es decir si en lugar de tener 17 años de educación tuviera 18 (1 año más) entonces la probabilidad pasaría a ser 1.8 % (es decir, el 2 por ciento original menos 0.20 puntos porcentuales). La forma incorrecta de interpretar los modelos probit y logit es si habláramos del cambio de probabilidad como una reducción del 0.02 % (cambio proporcional), porque en ese caso se entiende que la probabilidad predicha para ese caso seria 1.9996 por ciento,es decir hacer 2*(1-0.0002),lo cual es incorrecto.

2.3 Problema Aplicativo

La entidad financiera ABC, destina $800,000,000 de su capital a otorgar créditos personales de acuerdo a las siguientes convenciones:
-El Supervisor bancario, establece una tasa de severidad (LGD) de 45 % para el banco, ya que este no cuenta con un modelo interno para la estimación de dicho parámetro.
-El Supervisor, establece las categorías crediticias basándose en la probabilidad de incumplimiento (PD), de la siguiente manera: Cliente normal(0 – 20 %), cliente con problemas potenciales(20 %-40 %), cliente deficiente(40 %-60 %), cliente dudoso(60 %-80 %) y pérdida: (80 %-100 %)

-Basándose en los lineamientos de riesgo que sigue el banco, se establece que los préstamos personales en mención se harán de la siguiente manera: Clientes normales: 35 %, cliente con problemas potenciales: 30 %, cliente deficiente: 20 %, cliente dudoso: 10 % y pérdida: 5 % del capital invertido en préstamos.

-Se pide al banco declarar el gasto en provisiones que hará, teniendo en cuenta que para su cálculo sigue una metodología de Pérdidas Esperadas.

Desarrollo

2.3.1 Estimación con el Modelo Logit

Lo primero que se realizó fue realizar una estimación mediante el modelo Logit. Se regresionó la variable dependiente "default" (variable dicotómica discreta que toma el valor de 1 si el individuo cayó en default, y 0 en caso contrario) con respecto a las variables explicativas edad, rcuota_ingreso, ingreso, nro_ctas, nro_default_anterior, nro_prest_hipotec y nro_depend. Como resultado de la estimación, obtuvimos que todos los parámetros eran significativos excepto el coeficiente de la variable nro_prest_hipotec (Ver en **Anexo3**).

Para comprobar que dicha variable no era significativa, aplicamos el test de Wald, el test nos permite asegurar que dicha variable no era significativa. Por tanto, regresioanamos nuevamente el modelo logit, pero esta vez sin la variable en cuetión. El resultado obtenido es que ahora todas las variables consideradas son significativas. (Ver **Anexo4** y **Anexo5**)

2.3.2 Estimación con el Modelo Probit

Análogamente al caso anterior, realizamos una regresión mediante el modelo Probit de la variable cualitativa discreta dcicotómica "default" con respecto a todas las variables exógenas encontradas en la base de datos "data_pd". De la misma manera que con el modelo Logit, los resultados arrojan que la variable independiente nro_prest_hipotec es la única que no es significativa, al estimar nuevamente el modelo sin considerar esta vez dicha variable, se obtiene un modelo con todas las variables significativas. (Ver **Anexo6** y **Anexo7**)

2.3.3 Comparando entre Modelos

Una vez que hemos realizado las estimaciones con los modelos Logit y Probit, el siguiente paso es elegir entre estos dos modelos, el criterio de elección es: elegir el modelo que tenga mayor capacidad de predicción acetdad, esto será posible analizando la Potencia recurriendo al comando "lstat". Los resultados del test indican que con el modelo Logit se acierta en el 67.45 % de los casos, mientras que el modelo Probit acierta en el 67.44 %. (Ver **Anexo8** y **Anexo9**)
Al contrastar ambos resultados, se aprecia que el modelo logit es ligeramente mejor que el modelo Probit, debido a que la diferencia obtenida del test entre ambos modelos es mínima; se podría decir, en este caso particular que es indistinto optar por cualquiera de ellos. Sin embargo, el modelo elegido para desarrollar los pasos siguientes es el Modelo Logit.

Finalmente para validar nuestro modelo obtenido, analizamos la Curva ROC mediante el comando "lroc", el resultado muestra que el área es 0.7436, valor superior a 0.5. Por lo tanto, es correcto decir que nuestro modelo de elección discreta dicotómica: Logit, está bien especificado. (Ver **Anexo10**).

2.3.4 Probabilidad de Default

Ya que contamos con el modelo adecuado, además que está validado, lo que realizaremos ahora es estimar las probabilidades de default. Lo primero a hacer es obtener la probabuilidad de default para cada individuo. Es decir, obtendremos la probabilidad que cada individuo con sus características específicas cumpla sus pagos.
Después de esto, se ordena dichas probabilidades de menor a mayor, para poder facilitar la agrupación, ya que se categorizará a las personas en 5 niveles de riesgo, de acuerdo al nivel de probabilidas obtenida, dicha categorización será de la siguiente manera:

Cuadro 2.3.1: *Ranking Crediticio*

Categorías	
Cliente	PD(%)
Normal	$[0-20]$
CPP	$[20-40]$
Deficiente	$[40-60]$
Dudoso	$[60-80]$
Pérdida	$[80-100]$

Una vez categorizado a cada individuo, se debe calcular la probabilidad default promedio de cada categoría. Dichos valores representan el valor esperado de la PD por cada categoría. Los resultados de esta operación se meustran en el **Anexo11**.

Estos resultados nos permite corroborar con la teoría, ya que se aprecia que la esperanza que los individuos normales caigan caigan en default es baja (17.08 %), mientras la esperanza que los individuos categorizados en pérdida caigan en defaul es muy alta (92.07 %)

2.3.5 Pérdida Esperada

Contamos ya con el promedio de la probabilidad de incumplimiento de cada categoría crediticia que se ha calculado anteriormente, con la tasa de severidad (LGD) de 45 % establecido por el Supervisor bancario (SBS para el caso peruano) y el saldo expuesto determinado por la entidad financiera ABC de la siguiente manera:

Cuadro 2.3.2: *Saldo Expuesto*

Categorías	
Cliente	Porcentaje del capital invetido
Normal	35 %
CPP	30 %
Deficiente	20 %
Dudoso	10 %
Pérdida	5 %

Ahora, a partir de estos 3 datos es posible hallar la pérdida esperada para dicha entidad.(Ver **Anexo12**)
Los resultados nos dicen que el banco deberá tener una mayor cantidad de provisiones para las

categorías de clientes que se encuentren con problemas potenciales y/o sean deficientes; aunque sus probabilidad de incumplimiento no sean las más altas, la causa se debe a que tienen un mayor porcentaje del capital invertido.

Los clientes normales y dudosos presentan una menor perdida esperada, pero no son la categoría que necesitan menos provisiones. En el caso de clientes normales aunque tengan una baja probabilidad de incumplimiento, pero presentan un alto porcentaje del capital invertido (el más alto entre las cinco categorías). Para los clientes dudosos, es la situación contraria; presentan una alta probabilidad de incumplimiento y por lo tal el capital invertido no es tan alto.

Y con menor cantidad de provisiones se encuentra los clientes que son categorizados como pérdida ya que cuentan con una alta probabilidad de incumplimiento; justamente se espera que la perdida esperada no sea tan alta, y para esto el banco asigna un menor porcentaje de su capital.
En suma la perdida esperada total es $132,404,686.20; por lo tal el banco tendrá que declarar el gasto en provisiones igual a ese mismo monto.

3. Modelos de Conteo

3.1 Introducción

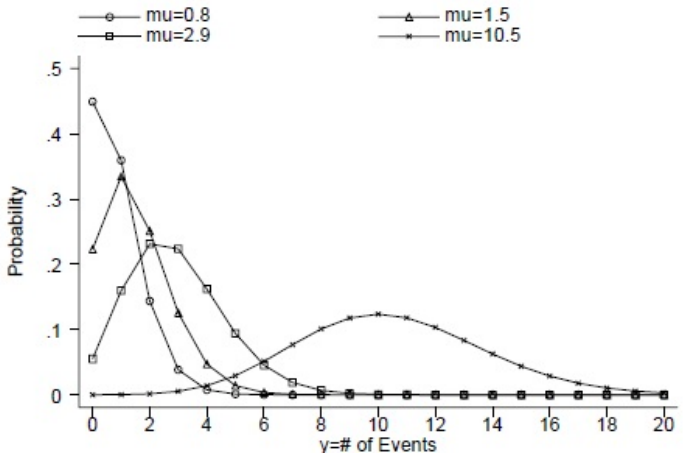
Contar las variables indica cuántas veces ha ocurrido un evento. Mientras que el uso de la regresión modelos de conteo es relativamente reciente, incluso una breve encuesta de aplicaciones recientes ilustra cómo estos resultados son comunes y la importancia de este tipo de modelos. Los ejemplos incluyen el número de pacientes, hospitalizaciones, homicidios diarios, conflictos internacionales, bebidas consumidas, accidentes de trabajo, nuevas empresas, y las detenciones por la policía, por nombrar sólo algunos.
Mientras que el modelo de regresión lineal a menudo se ha aplicado para contar los resultados, esto puede resultar en que las estimaciones sean ineficientes, inconsistentes y sesgadas. A pesar de que hay situaciones en las que el la regresión lineal proporciona resultados razonables, es mucho más seguro de usar modelos diseñados específicamente para el conteo de resultados. En este capítulo se estudiara el modelo de regresión de Poisson (PRM).

3.2 Distribución de Poisson

La distribución de Poisson univariado es fundamental para la comprensión de los modelos de conteo. En consecuencia, comenzamos explorando esta distribución. Sea y una variable aleatoria que indica la número de veces que se ha producido un evento. Si Y tiene una distribución de Poisson, a continuación:

$$Pr(y|\mu) = \frac{e^{\mu}\mu^{y}}{y!} \tag{3.2.1}$$

donde $\mu > 0$• es el único parámetro que define la distribución. La manera más fácil de conseguir un sentido de esta distribución es comparar la trama de la probabilidad pronosticada para diferentes valores de la tasa parámetro μ (etiquetado como mu en el gráfico):



La trama muestra cuatro características de la distribución de Poisson que son importantes para la comprensión modelos de regresión para el recuento:

- μ es la media de la distribución. Como μ aumenta, la masa de la distribución se desplaza hacia la derecha.
- μ es también la varianza. Por lo tanto, $Var(y) = \mu$, que se conoce como equidispersión. En los datos reales, muchas variables de recuento tienen una varianza mayor que la media, que se llama sobredispersión.
- Como μ aumenta, la probabilidad de que un cero disminución de los recuentos. Para muchas variables de recuento, hay ceros que las predichas por la distribución de Poisson más observado.
- Como μ aumenta, la distribución de Poisson se aproxima a una distribución normal. Esto se muestra por la distribución de $\mu = 10,5$.

3.3 Modelo de Regresión de Poisson

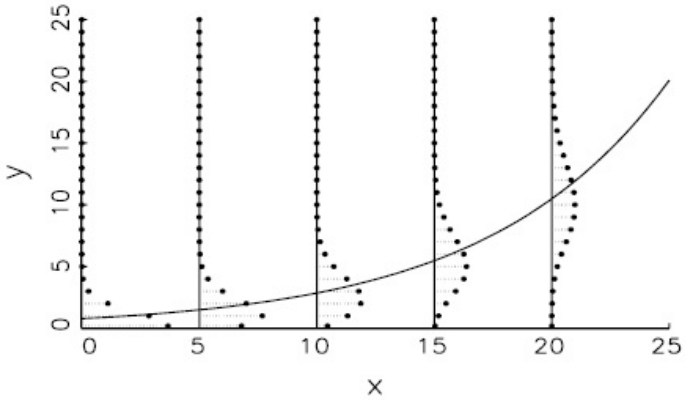
El modelo de regresión de Poisson (PRM) se extiende de la distribución de Poisson al permitir que cada observación tener un valor diferente de μ. Más formalmente, el PRM asume que el recuento observado para la observación i se extrae de una distribución de Poisson con μ_i de media, donde μ_i se estima a partir de las características observadas. Esto se refiere a veces como la incorporación de heterogeneidad observada, y conduce a la ecuación estructural:

$$\mu_i = E(y_i|x_i) = exp(x_i\beta) \tag{3.3.1}$$

Por lo tanto la distribución de Possion con la variables explicativas x, seria:

$$Pr(y|x) = \frac{e^{\mu_i}\mu_i^y}{y!} \tag{3.3.2}$$

Tomando el exponencial de $x\beta$ para μ debe ser positivo, lo cual necesario ya que el conteo sólo puede ser 0 o positivo. Para ver cómo funciona esto, considere el modelo de regresión de Poisson con una sola variable independiente μ =exp $(\alpha + \beta x)$, que puede ser trazada como:



En este gráfico, la media μ, que se muestra por la línea curva, aumenta a medida que aumenta x. Para cada valor de μ, la distribución alrededor de la media se muestra por los puntos y que representan la probabilidad de cada conteo. Interpretación del modelo implica evaluar cómo los cambios en las variables independientes afectan a la media condicional y las probabilidades de varios conteos.

3.3.1 Estimación por máxima verosimilitud

$$lnt = \sum_{i=1}^{n}(-\mu + ylnu - ln(y!)) \tag{3.3.3}$$

$$lnt = \sum_{i=1}^{n}(-e^{x\beta} + yx\beta - ln(y!)) \tag{3.3.4}$$

Derivamos la ecuación respecto de β

$$\frac{\partial lnt}{\partial \beta} = \sum_{i=1}^{n}(-xe^{x\beta} + yx) = 0 \tag{3.3.5}$$

$$\sum_{i=1}^{n}(xe^{x\beta}) = \sum_{i=1}^{n}(y_i x_i) \tag{3.3.6}$$

$$\frac{\partial^2 lnt}{\partial \beta^2} = -\sum_{i=1}^{n}(xxe^{x\beta}) \tag{3.3.7}$$

3.4 Ejemplo de una estimación del modelo de regresión de Poisson en Stata

Para este ejemplo, utilizamos datos de Long (1990) sobre el número de publicaciones producido por Ph.D. bioquímicos. Las variables consideradas son
. use couart2, clear
. describe

variable name	storage type	display format	value label	variable label
art	byte	%9.0g		Articles in last 3 yrs of PhD
fem	byte	%9.0g	sexlbl	Gender: 1=female 0=male
mar	byte	%9.0g	marlbl	Married: 1=yes 0=no
kid5	byte	%9.0g		Number of children < 6
phd	float	%9.0g		PhD prestige
ment	byte	%9.0g		Article by mentor in last 3 yrs

. summarize

Variable	Obs	Mean	Std. Dev.	Min	Max
art	915	1.692896	1.926069	0	19
fem	915	.4601093	.4986788	0	1
mar	915	.6622951	.473186	0	1
kid5	915	.495082	.76488	0	3
phd	915	3.103109	.9842491	.755	4.62
ment	915	8.767213	9.483916	0	77

Las diferencias entre los científicos en sus índices de productividad podría deberse a factores como el género, el estado civil, el número de jóvenes niños, el prestigio del programa de postgrado, y el número de artículos escritos por el mentor de un científico. Para dar cuenta de estas diferencias, añadimos estas variables como variables independientes, donde la variable dependiente sera el numero de artículos en los últimos 3 años de doctorado.

Ahora utilizaremos el siguiente comando para estimar el modelo.
. poisson art fem mar kid5 phd ment, nolog

Poisson regression — Number of obs = 915; LR chi2(5) = 183.03; Prob > chi2 = 0.0000; Log likelihood = -1651.0563; Pseudo R2 = 0.0525

art	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
fem	-.2245942	.0546138	-4.11	0.000	-.3316352	-.1175532
mar	.1552434	.0613747	2.53	0.011	.0349512	.2755356
kid5	-.1848827	.0401272	-4.61	0.000	-.2635305	-.1062349
phd	.0128226	.0263972	0.49	0.627	-.038915	.0645601
ment	.0255427	.0020061	12.73	0.000	.0216109	.0294746
_cons	.3046168	.1029822	2.96	0.003	.1027755	.5064581

La manera en la cual se interpreta un modelo de conteo depende si se está interesado en el valor esperado de la variable de recuento o en la distribución de los recuentos. Si el interés está en el

recuento esperado, varios métodos se pueden utilizar para calcular el cambio en la expectativa de un cambio en una independiente variable.
Si el interés está en la distribución de los recuentos o tal vez sólo la probabilidad de que un recuento específico, la probabilidad de que un recuento para un nivel dado de las variables independientes se puede calcular.

- Factor de Cambio en la E (y / x)
  Quizás el método más común de interpretación es el factor de cambio en la tarifa. Si definimos E (y / x,x_k) como el recuento esperado para un determinado x donde notamos explícitamente el valor de x_k, y definir E (y / x, $x_k + \delta$) como el recuento de espera después de aumentarx_k por unidades δ, entonces

$$\frac{E(y/x, x_k + \delta)}{E(y/x, x_k)} = e^{\beta_k \delta} \tag{3.4.1}$$

  Por lo tanto, los parámetros pueden ser interpretados como Para un cambio de δ en x_k, el recuento esperados aumenta en un factor de $exp(\beta_k \delta)$, teniendo a todas las otras variables constantes.
- Cambio porcentual en el E (y / x)
  Por otra parte, el porcentaje de cambio en el recuento esperado para un cambio unitario δ en x_k, la celebración de otra las variables constantes, se puede calcular como:

$$100 * \frac{E(y/x, x_k + \delta) - E(y/x, x_k)}{E(y/x, x_k)} = 100 * [exp(\beta_k * \delta) - 1] \tag{3.4.2}$$

Calculamos el factor y el cambio en el E (y / x)

Coeficientes de cambio Factor se pueden calcular utilizando listcoef:
. poisson art fem mar kid5 phd ment, nolog
. listcoef fem ment, help

```
poisson (N=915): Factor Change in Expected Count

 Observed SD: 1.926069
```

art	b	z	P>|z|	e^b	e^bStdX	SDofX
fem	-0.22459	-4.112	0.000	0.7988	0.8940	0.4987
ment	0.02554	12.733	0.000	1.0259	1.2741	9.4839

```
      b = raw coefficient
      z = z-score for test of b=0
  P>|z| = p-value for z-test
    e^b = exp(b) = factor change in expected count for unit increase in X
e^bStdX = exp(b*SD of X) = change in expected count for SD increase in X
  SDofX = standard deviation of X
```

Por ejemplo, los coeficientes de fem y ment pueden ser interpretados como: Ser una científica disminuye el número esperado de artículos por un factor de 0.80, manteniendo las demás variables constantes.
Para un aumento de una desviación estándar de la productividad del mentor, aproximadamente 9,5 artículos, un medias aumento de la productividad del científico por un factor de 1,27, manteniendo constante otras variables.Para calcular el porcentaje de cambio utilizamos el comando:
listcoef fem ment, percent help

```
poisson (N=915): Percentage Change in Expected Count

 Observed SD: 1.926069
```

art	b	z	P>|z|	%	%StdX	SDofX
fem	-0.22459	-4.112	0.000	-20.1	-10.6	0.4987
ment	0.02554	12.733	0.000	2.6	27.4	9.4839

```
      b = raw coefficient
      z = z-score for test of b=0
  P>|z| = p-value for z-test
      % = percent change in expected count for unit increase in X
  %StdX = percent change in expected count for SD increase in X
  SDofX = standard deviation of X
```

Por ejemplo, los coeficientes de variación porcentual de fem y ment pueden ser interpretados como: Ser una científica disminuye el número esperado de artículos en un 20 por ciento, manteniendo todas las otras variables constantes. Por cada artículo adicional por parte del mentor, predijo de un científico de la productividad media aumenta en un 2,6 por ciento, manteniendo constantes otras variables.

Cambio marginal en E (y / x)

Otro método de interpretación es el cambio marginal en E (y / x)

$$\frac{\partial E(y/x_k)}{\partial x} = E(y/x)\beta_k \tag{3.4.3}$$

Para $\beta_k > 0$ es mayor el valor actual de E (y | x), mayor es la tasa de cambio; para $\beta_k < 0$,es menor es la tasa de cambio. El marginal respecto de x_k depende tanto β_k y E (y/ x). Por lo tanto, el valor de la marginal depende de los niveles de todas las variables en el modelo. En la práctica, este medida a menudo se calcula con todas las variables se encuentren en su medio.

Ejemplo de cambio marginal utilizando mfx compute

Por default, mfx compute calcula el cambio marginal con variables se encuentren en su medio:
. mfx compute

3.4.1 Interpretación utilizando probabilidades predichas

Los parámetros estimados se pueden utilizar también para calcular probabilidades predichas utilizando la siguiente fórmula:

$$\widehat{Pr}(y = m|x) = \frac{e^{-x\widehat{\beta}}(x\widehat{\beta})^m}{m!} \tag{3.4.4}$$

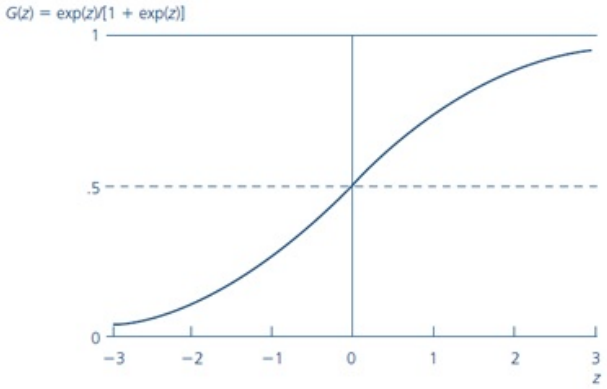
Probabilidades pronosticadas en los valores especificados se pueden calcular utilizando prvalue. Las predicciones de los valores observados para todas las observaciones se pueden calcular usando prcounts.
. poisson art fem mar kid5 phd ment, nolog
. prcounts prm, plot max(9)
. d prm*

II

Parte Dos

Anexos

Anexo1: Representación gráfica de la función logística



Anexo2: Resultados de la aplicación del Modelo Probit

```
Logistic regression                             Number of obs   =      6669
                                                LR chi2(7)      =    189.95
                                                Prob > chi2     =    0.0000
Log likelihood = -1615.0284                     Pseudo R2       =    0.0555

------------------------------------------------------------------------------
  desocupado |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        edad |  -.0109965   .0048867    -2.25   0.024    -.0205744   -.0014187
       mujer |   .2553919   .1050513     2.43   0.015     .0494951    .4612887
     soltero |   .2554484   .1181695     2.16   0.031     .0238404    .4870564
        educ |   -.100993    .013732    -7.35   0.000    -.1279073   -.0740787
        jefe |  -.6682848   .1161409    -5.75   0.000    -.8959168   -.4406529
        inla |    .000687   .0000901     7.63   0.000     .0005105    .0008635
        caba |   -.090867   .1227719    -0.74   0.459    -.3314955    .1497614
       _cons |  -.9996536   .2698181    -3.70   0.000    -1.528487   -.4708199
------------------------------------------------------------------------------
```

Anexo3: Resultados de la estimación con el Modelo Logit

```
. logit   default edad rcuota_ingreso ingreso nro_ctas nro_default_anterior nro_prest_
> end

Iteration 0:   log likelihood = -11124.315
Iteration 1:   log likelihood = -9635.8759
Iteration 2:   log likelihood = -9512.9024
Iteration 3:   log likelihood = -9509.4344
Iteration 4:   log likelihood =  -9509.433
Iteration 5:   log likelihood =  -9509.433

Logistic regression                             Number of obs   =      16049
                                                LR chi2(7)      =    3229.76
                                                Prob > chi2     =     0.0000
Log likelihood =  -9509.433                     Pseudo R2       =     0.1452

------------------------------------------------------------------------------
     default |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        edad |  -.0275301   .0013794   -19.96   0.000    -.0302337   -.0248264
rcuota_ingreso |   .7539031   .0632351    11.92   0.000     .6299646    .8778415
     ingreso |  -.0000301   5.33e-06    -5.65   0.000    -.0000406   -.0000197
     nro_ctas |   .0189821   .0040483     4.69   0.000     .0110476    .0269167
nro_default_anterior |   1.508588   .0480815    31.38   0.000     1.41435    1.602826
nro_prest_hipotec |  -.0136214   .0209849    -0.65   0.516    -.054751    .0275082
  nro_depend |   .0956232   .0151696     6.30   0.000     .0658912    .1253551
       _cons |   .6555473   .0761852     8.60   0.000     .5062269    .8048676
------------------------------------------------------------------------------
```

Anexo4: Resultados del Test de Wald

```
( 1)  [default]edad = 0
( 2)  [default]rcuota_ingreso = 0
( 3)  [default]ingreso = 0
( 4)  [default]nro_ctas = 0
( 5)  [default]nro_default_anterior = 0
( 6)  [default]nro_depend = 0

           chi2(  6) = 1785.00
         Prob > chi2 =    0.0000
```

Anexo5: Resultados de la segunda estimación con el Modelo Logit

```
. logit default edad rcuota_ingreso ingreso nro_ctas nro_default_anterior nro_depend

Iteration 0:   log likelihood = -11124.315
Iteration 1:   log likelihood = -9636.1744
Iteration 2:   log likelihood = -9513.1138
Iteration 3:   log likelihood = -9509.6451
Iteration 4:   log likelihood = -9509.6438
Iteration 5:   log likelihood = -9509.6438

Logistic regression                             Number of obs   =      16049
                                                LR chi2(6)      =    3229.34
                                                Prob > chi2     =     0.0000
Log likelihood = -9509.6438                     Pseudo R2       =     0.1451

------------------------------------------------------------------------------
     default |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        edad |  -.0275599   .0013787   -19.99   0.000    -.0302621   -.0248577
rcuota_ingreso |   .7325004    .053816    13.61   0.000     .6270229    .8379779
     ingreso |  -.0000317   4.73e-06    -6.71   0.000     -.000041   -.0000225
     nro_ctas |   .0184824   .0039743     4.65   0.000     .0106929    .0262718
nro_default_anterior |   1.509618   .0480635    31.41   0.000     1.415416    1.603821
  nro_depend |   .0954537   .0151671     6.29   0.000     .0657266    .1251807
       _cons |   .6652562   .0747031     8.91   0.000     .5188408    .8116715
------------------------------------------------------------------------------
```

Anexo6: Resultados de la estimación con el Modelo Probit

Probit regression — Number of obs = 16049; LR chi2(7) = 3118.61; Prob > chi2 = 0.0000; Log likelihood = -9565.0115; Pseudo R2 = 0.1402

default	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
edad	-.0168142	.000828	-20.31	0.000	-.0184371	-.0151913
rcuota_ingreso	.4502273	.0366519	12.28	0.000	.3783908	.5220638
ingreso	-.0000187	3.18e-06	-5.87	0.000	-.0000249	-.0000124
nro_ctas	.010948	.0024582	4.45	0.000	.0061299	.0157661
nro_default_anterior	.7725406	.0219723	35.16	0.000	.7294756	.8156055
nro_prest_hipotec	-.0080191	.0126349	-0.63	0.526	-.0327829	.0167448
nro_depend	.0585236	.0092452	6.33	0.000	.0404032	.0766439
_cons	.4204389	.0462804	9.08	0.000	.329731	.5111469

Anexo7: Resultados de la segunda estimación con el Modelo Probit

Probit regression — Number of obs = 16049; LR chi2(6) = 3118.20; Prob > chi2 = 0.0000; Log likelihood = -9565.2129; Pseudo R2 = 0.1402

default	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
edad	-.0168306	.0008276	-20.34	0.000	-.0184528	-.0152085
rcuota_ingreso	.438375	.031522	13.91	0.000	.3765931	.5001569
ingreso	-.0000196	2.83e-06	-6.92	0.000	-.0000251	-.000014
nro_ctas	.0106277	.0024059	4.42	0.000	.0059123	.0153431
nro_default_anterior	.7729776	.0219627	35.20	0.000	.7299315	.8160237
nro_depend	.058424	.0092438	6.32	0.000	.0403065	.0765416
_cons	.4259722	.0454531	9.37	0.000	.3368858	.5150586

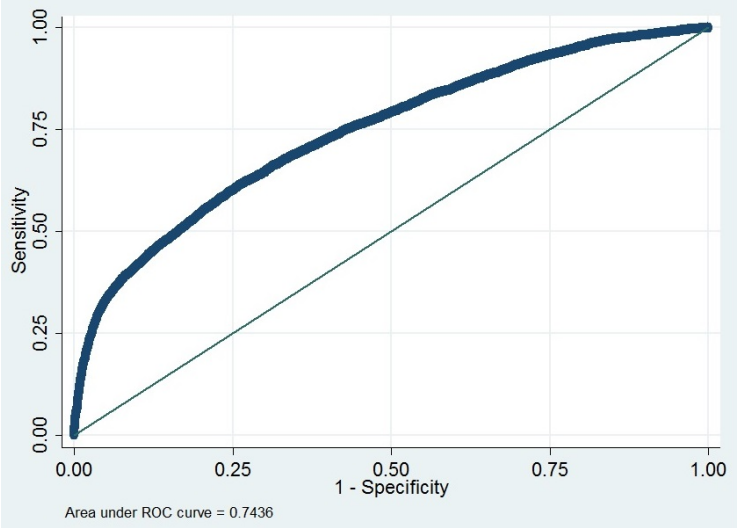
Anexo8: Potencia de la predicción con el Modelo Logit

	True		
Classified	D	~D	Total
+	4610	1815	6425
-	3409	6215	9624
Total	8019	8030	16049
Correctly classified			67.45%

Anexo9: Potencia de la predicción con el Modelo Probit

	True		
Classified	D	~D	Total
+	4683	1889	6572
-	3336	6141	9477
Total	8019	8030	16049
Correctly classified			67.44%

Anexo10: Representación gráfica de la Curva ROC



Anexo11: Valor esperado de la PD por categoría

Categoria	pr_logit
normal	.1708423
cpp	.3160654
deficiente	.48738
dudoso	.696633
perdida	.9207414

Anexo12: Pérdida esperada de la entidad financiera ABC por categoría

	Categoria	PE
1.	normal	2.15e+07
2.	cpp	3.41e+07
3.	deficiente	3.51e+07
4.	dudoso	2.51e+07
5.	perdida	1.66e+07

Bibliografía

Books

- GREENE, W.H. (2003) “Econometric Analysis”5ª edición. Prentice Hall N.J. Capítulo 21

- WOOLDRIDGE, J.M. (2010) “Introducción a la Econometría: Un Enfoque Moderno". 4ª edición. Cengage Learning. Capítulo 17

Índice alfabético