Programa de Especialización en Econometría Aplicada
Centro de Formación Continua -UNI
Modelos Variable conteo
Clase 3

Edinson Tolentino
MSc Economics
email: edinson.tolentino@gmail.com

Twitter: @edutoleraymondi

Universidad Nacional de Ingeneria

30 de noviembre de 2024

Contenido

Introducción

- Los datos de conteo son variables cuantitativas discretas restringidas a valores no negativos ($y_i = 0, 1, 2, 3, ...$). Es decir, "y" asume, relativamente, pocos valores incluido el cero. Algunos ejemplos:
  - N° de hijos de una madre (fecundadidad).
  - N° de veces que los individuos demandan servicios de salud (visitar al doctor o número de días en el hospital).
  - N° de patentes registradas por una empresa en un año.
- Cuando "y" tiene estas características no es apropiado utilizar MCO, para tal caso la forma correcta de estimar este tiepo de modelos es el de Máxima Verosimilitud.
- Un enfoque estándar para analizar variables de conteo son los modelos de regresión Poisson y Binomial no Negativa.

Función de Distribución Poisson

- Una variable aleatoria "y", cuyo soporte es $\mathbb{N} \cup \{0\} = \{0, 1, 2, ...\}$, sigue una distribución Poisson si y solo si la función de probabilidad es:

$$Pr(y = j) = \frac{e^{-\lambda}\lambda^{j}}{j!} \quad ; \lambda > 0 \;\; \text{y} \;\; j = 0, 1, 2, ...$$

otra manera de exéresarlo es $y \sim \text{Poisson}(\lambda)$

- La distribución Poisson es una distribución de un solo parámetro (λ). Donde λ determina la media y varianza:

$$E(y) = Var(y) = \lambda$$

- La igualdad entre la media y la varianza es una propiedad de la distribución Poisson y es llamda **equidispersión**.
- Desafortunadamente, esta propiedad es frecuentemente violada en aplicaciones empíricas: **overdispersión** $(Var(y) > E(y))$ o **subdispersión** $(E(y) > Var(y))$.

Modelo

- El modelo de Poisson puede derivarse de la función de distribución Poisson, considerando 2 supuestos:
  1. La distribución condicional de y dado x esta distribuida como Poisson con parámetro $\lambda_i = \lambda(x_i, \beta)$.
  $$y_i|x_i \sim \text{Poisson}(\lambda_i)$$
  2. Parametrizamos λ_i en términos de x_i y β:
  $$\lambda_i = exp(x_i'\beta)$$
- Combinando ambos supuestos, obtenemos la función de probabilidad condicional del modelo Poisson:

$$f(y_i|x_i'\beta) = \frac{exp(-exp(x_i'\beta))exp(x_i'\beta)}{y_i!} \quad , \quad y_i = 0, 1, 2, 3, ...$$

- LA propiedad de Equidispersión se mantiene en el modelo de regresión.

$$E(y_i|x_i'\beta) = Var(y_i|x_i'\beta) = exp(x_i'\beta)$$

- Esto implica que el modelo sea intrinsicamente heterocedástico.

Efectos marginales

- El efecto margina de una variable continua x_j sobre $E(y_i|x_i)$ es:

$$\frac{\partial E(y_i|x_i)}{\partial x_j} = \frac{\partial exp(x_i'\beta)}{\partial x_j} = exp(x_i'\beta)\beta_j$$

- Alternativamente, es posible reportar el cambio relativo en $E(y_i|x_i)$ asociado a un cambio en un regresor continuo:

$$\frac{\partial E(y_i|x_i)}{\partial x_j} = exp(x_i'\beta)\beta_j$$

$$\Rightarrow \beta_j = \frac{\partial E(y_i|x_i)}{\partial x_j} \cdot \frac{1}{exp(x_i'\beta)} = \frac{\partial E(y_i|x_i)}{\partial x_j} \cdot \frac{1}{E(x_i'\beta)} = \frac{\partial E(y_i|x_i)/E(y_i|x_i)}{\partial x_j} = \frac{\partial \ln E(y_i|x}{\partial x_j}$$

- Esta expresión es constante para todo i y se interpreta como una semielasticidad: Si x_j se incrementa en una unidad, la β_j resulta el cambio porcentual en $E(y_i|x_i)$.

Estimación

- Si la especificación para la distribución condicional de la variable de respuesta, así como para la media condicional es correcta, y bajo el supuesto de que se tienen ibservaciones indeoendientes, entonces se obtiene la función de log-verosimilitud:

$$\log L(\beta, y, x) = \sum_{i=1}^{n} \{y_i x_i' \beta - exp(x_i' \beta) - \ln(y_i!)\}$$

- Los paquetes estadísticos ofrecen etimadores de β que maximizan esta función.

Introducción

- El modelo Poisson presenta algunos problemas:
  1. Una de las principales razones por las que el modelo Poisson falla es la heterogeneidad no observada. Esto significa que hay factores no observados, en especial características individuales, que ejercen alguna influencia sobre la variabilidad de y.
  2. El problema de heterogeneidad no observada puede tener consecuencias en la inferencia estadística:
     - Pueden introducir sobredispersión.
     - Número excesivo de ceros.
  3. Esta heterogeneidad, ignorada por el modelo Poisson, puede modelarse de manera explicita mediante el uso de la regresión binomial negativa.

Modelo

- Para derivar la distribución Binomial Negativa, se asume que se está ante la presencia de una mezcla de distribuciones, en la cual los datos observados se distribuyen como una Poisson, pero se presupone un elemento de heterogeneidad individual no observado, que sigue una distribución gamma, y que refleja el hecho de que la verdadera media no se ha medido perfectamente.
- Siendo más formales, remplazamos μ por $\mu\nu$, donde ν es una variable aleatorio, así $y \sim \text{Poisson}(y|\lambda\nu)$.
- Supungamos que $E(\nu) = 1$ y $Var(\nu)\sigma^2$, asi ante la presencia de ν se mantiene la media pero se incrementa la dispersión.

$$E(y) = \mu \;\; ; \;\; Var(y) = \mu(1 + \lambda\sigma^2) > E(y) = \mu$$

- La sobredispersión describe la característica $Var(y) > E(y)$ o, siendo más precisos, $Var(y|x) > E(y|x)$ en el modelo de regresión.
- Se asume $\nu \sim \text{Gamma}(1, \alpha)$, donde α es el Parámetro de varianza de la distribución Gamma.

Modelo

- Así, la función de densidad para la distribución Binomial Negativa, construida a partir de un mix entre la distribución Poisson y Gamma es:

$$f(y|x) = \frac{\Gamma(\alpha^{-1}+y)}{\Gamma(\alpha^{-1})\Gamma(y+1)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\mu+\alpha^{-1}}\right)^{y}$$

  donde α es el parámetro de dispersión $(\alpha \geq 0)$ y $\Gamma(\cdot)$ es la función Gamma.
- El parámetro de dispersión α es el que ayuda a definir la relación entre la media y la varianza, conocida así como función de varianza.
- Si por ejemplo $\alpha \to 0$, entonces la media y la varianza son iguales y se tiene el modelo Poisson. Por otro lado, las funciones más comunes para la relación media-varianza de la distribución Binomial NEgativa son la lineal y la cuadratica.
- Los momentos de la Binomial Negativa son:

$$E(y|\mu,\alpha) = \mu \quad ; \quad Var(y|\mu,\alpha) = \mu(1+\alpha\mu)$$

Estimación

- La función de log-verosimilitud para el modelo Binomial Negativa, con función de varianza cuadrática, es la siguiente:

$$\log L(\beta, y, x) = \sum_{i=1}^{n} \{ \sum_{j=0}^{y_i - 1} \log(j + \alpha^{-1}) - \log(y_i!) - (y_i + \alpha^{-1}) \times \log(1 + \alpha . exp(x_i' \beta) + y_i \log(\alpha + y x_i' \beta)) \}$$

- Esta función es utilizada por la paquetería estadística para encontrar los estimadores de β.

ZIP Models

- Es posible que el mecanismo aleatorio que dio origen a los datos de conteo muestre una mayor concentración para algún valor especifico, como el cero.
- Esto implica que dicho valor tienen una mayor probabilidad de ocurrencia en ceros que la especificada por los modelos Poisson y Binomial Negativa. En genera, ambos modelos suelen subpredecir la ocurrencia de ceros.
- Es posible que los ceros tengan un doble origen: Por ejemplo, si analizamos el número de veces que una persona práctica tennis en un mes. Los ceros pueden ser originados porque la persona nunca práctico tennis o proque en esos 30 días no práctico el deporte.
- Esto significa que se tiene una mezcla de distribuciones, por lo que no sería adecuado, asumir que los ceros y no ceros se han generado por un mismo proceso.
- La idea básica en un modelo cero inflado es introducir una variable binaria C_i que toma el valor de 1 si la observación es cero, y toma el valor de 0 para las observaciones positivas (enteros no negativos).

ZIP Models

- Siendo w la probabilidad de $C_i = 1$ y suponemos una variable de conteo latente y_i^* que sigue una distribución Poisson con $\lambda = exp(x_i'\beta)$. El valor observado de y_i es dado por:

$$y_i = \begin{cases} 0 & \text{si } C_i = 1 \\ y_i^* & \text{si } C_i = 0 \end{cases}$$

La función de probabilidades es:

$$f(y_i|x_i) = w_i d_i + (1 - w_i)\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \quad ; y_i = 0, 1, 2, ...$$

Donde $d_i = I(y_i = 0)$. Siendo más específico:

$$f(y_i|x_i) = \begin{cases} w_i + (1 - w_i)e^{\lambda} & \text{si } y_i = 0 \\ (1 - w_i)\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} & \text{si } y_i \leq 1 \end{cases}$$

- Usualmente w es parametrizado a través de los modelos logit o probit. Asimismo, es posible cambiar la distribución de y_i^* a una Binomial Negativa.

Implementación en Stata

Stata implementa los modelos de conteo Poisson y Binomial Negativa mediante los comandos `poisson` y `poinbregsson` cuya sintaxis es:

Syntax

poisson `depvar [indepvars] [if] [in] [weight] [,options]`
nbreg `depvar [indepvars] [if] [in] [weight] [,options]`

Las opciones más importantes son:

- `exposure()`: especifica una variable indicando la cantidad de tiempo durante el cual la observación esta expuesta a la ocurrencia del evento.
- `offset()`: a diferencia de la opción `exposure()`, con `offset()` se especifica una variable que es igual al logaritmo de la "exposición al tiempo".

Implementación en Stata

Stata implementa el modelo cero-inflado mediante los comandos `zip` y `zinb` cuya sintaxis es:

Syntax

zip `depvar [indepvars] [if] [in] [weight], inflation(varlist) [options]`
zinb `depvar [indepvars] [if] [in][weight], inflation(varlist) [options]`

Las opciones más importantes son:

- `inflation(varlist)`: especifica los regresores que determinan si la observación de conteo es cero.
- `probit`: especifica que se utilice un modelo Probit para caracterizar el exceso de ceros en la data. Por default es Logit.

Las demás opciones son comunes a los comandos `poisson` y `poinbregsson`.