Regresión de Poisson

Microeconomía Cuantitativa

R. Mora

Departmento de Economía
Universidad Carlos III de Madrid

Esquema

Introducción

Ejemplo 1

Relación entre los gastos de investigación y desarrollo de las empresas (I + D) y el número de patentes recibidas por ellos.

- La variable dependiente, número de patentes, es un **recuento** del número total de patentes de una empresa en particular, normalmente en un período de tiempo, como un año.
- Sólo son posibles números no negativos y discretos: algunas empresas tienen cero patentes, mientras que otras empresas pueden muchas patentes.

Ejemplo 2

El uso de los servicios de salud entre los ancianos.

- La variable dependiente, el número de visitas al médico, es un **recuento** del número total de visitas al médico a una persona de edad avanzada ha hecho durante el año pasado.
- Sólo son posibles números no negativos y discretos: algunas personas tienen cero visitas, mientras que otras personas pueden tener incluso decenas de visitas.

Ejemplo 3

El efecto de las penas de prisión en la criminalidad.

- La variable dependiente, el número de arrestos, es un **recuento** del número total de arrestos que un hombre tiene en un año determinado.
- Sólo son posibles números no negativos y discretos: algunas personas tienen cero arrestos, mientras que otras personas pueden tener varias detenciones al año.

MCO con datos de recuento

- En principio, si la variable dependiente es positiva, $y_i > 0$, podríamos tomar su logaritmo y suponiendo que

$$\mathrm{E}(\ln(y_i)|x_i) = x'\beta$$

entonces MCO nos daría estimaciones del vector β consistentes.

- Sin embargo:
  - si $y = 0$ entonces no hay posibilidad de realizar la trasnformación logarítmica (soluciones adhoc: $\ln(y+1)$, usa $\ln(0.5)$ cuando $y = 0$).
  - para predicción queremos predecir $\mathrm{E}(y)$, pero $\exp(\mathrm{E}(\ln y)) \neq \mathrm{E}(y)$ incluso aunque $\exp(\ln y) = y$.
- En esta sesión, vamos a estudiar un modelo no lineal que evita estos dos problemas.

La distribución Poisson

Definición de la distribución de Poisson

- Una variable aleatoria discreta Z se distribuye con la distribución de Poisson con parámetro λ si la probabilidad de que $Z = z$ es

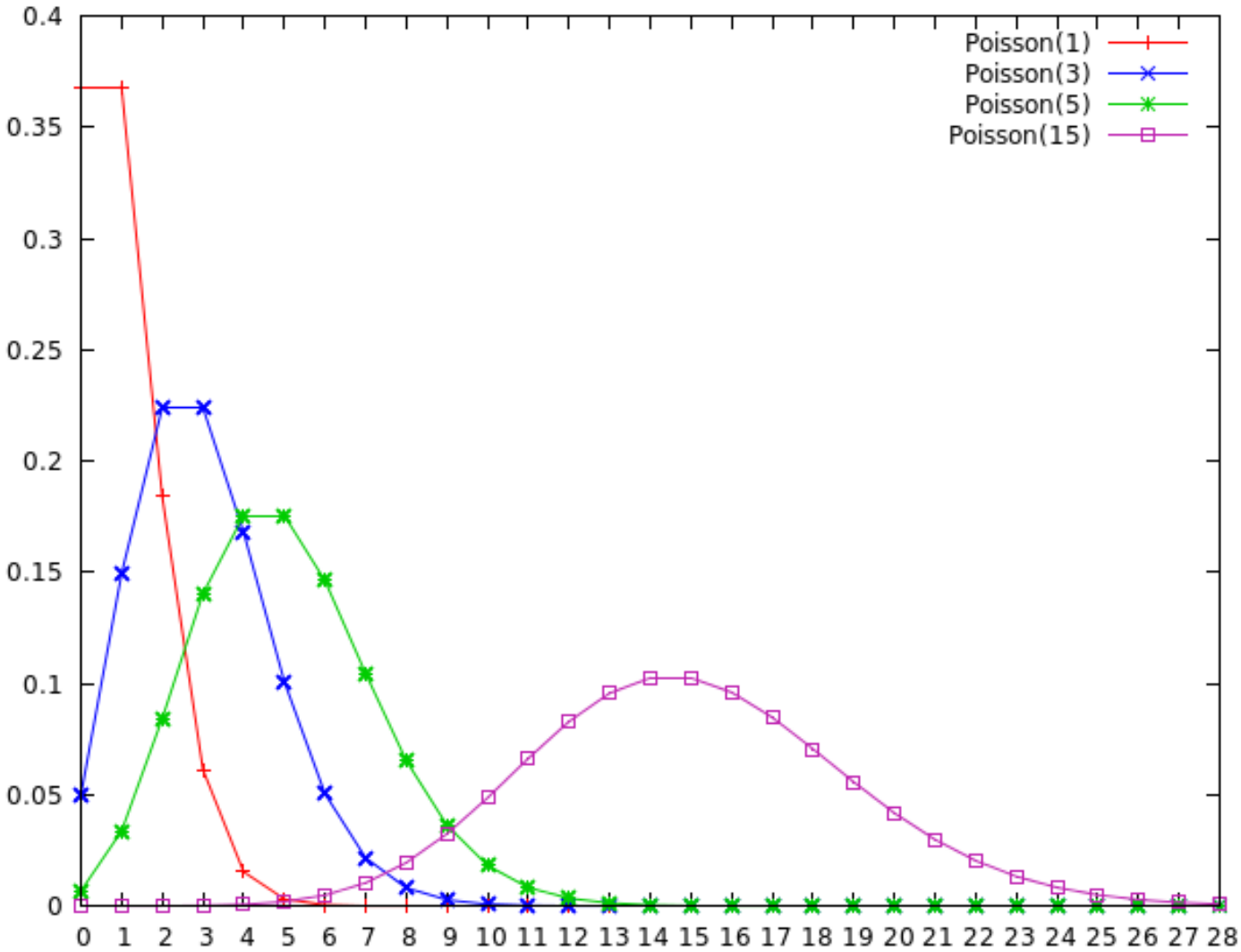
$$\Pr(Z = z) = \frac{\lambda^z}{z!} e^{-\lambda}$$

donde

  - $z = 0, 1, 2, \ldots$
  - $z! = z \times (z-1) \times (z-2) \times \ldots \times 2 \times 1$

- El parámetro λ es igual tanto a la esperanza como a la varianza de Z (equi-dispersión)

$$\lambda = \mathrm{E}(Z) = \mathrm{Var}(Z)$$


Poisson(1)
Poisson(3)
Poisson(5)
Poisson(15)
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28

Intuición

- La distribución de Poisson es la probabilidad de que un determinado número de eventos ocurran en un intervalo fijo de tiempo y/o espacio si estos eventos
  - Se producen a una tasa media constante.
  - Independientemente del tiempo transcurrido desde el último evento.
- Ejemplos históricos famosos incluyen (de Wikipedia):
  - El número de condenas erróneas en un determinado país en un período determinado de tiempo.
  - El número de soldados en el ejército prusiano muertos accidentalmente por patadas de caballo.

Regresión de Poisson

Regresión de Poisson

- Es un tipo de análisis de regresión para modelar datos de recuento.

- Sea y la variable dependiente y x un vector de variables independientes.
- La regresión de Poisson asume
  - y tiene una distribución de Poisson
  - La expectativa (y varianza) de y dado x es

$$\lambda = \mathrm{E}(y|x) = e^{x'\beta}$$

  - Por lo tanto, $\log(\mathrm{E}(y|x)) = x'\beta$ (el modelo de regresión de Poisson se refiere a veces como el modelo log-lineal).

Estimación por MV

- La probabilidad viene dada por $\Pr(y|x) = \frac{e^{yx'\beta} e^{-e^{x'\beta}}}{y!}$
- Para una muestra de N observaciones,

  $\Pr(y_1, ..., y_N|x_1, ..., x_N) = \prod_{i=1}^{N} \frac{e^{y_i x_i'\beta} e^{-e^{x_i'\beta}}}{y_i!}$
- La log-verosimilitud es:

$$L(b) = \sum_{i=1}^{N} \left\{ y_i x_i' b - e^{x_i' b} - \log(y_i!) \right\}$$

(ya que $y_i!$ no depende de b, podemos ignorar este elemento)

- Las CPO sólo pueden resolverse numéricamente. Como la log-verosimilitud es cóncava, esto no es un gran problema.

Interpretación

- Tenemos
  $$\mathrm{E}(y|x) = \exp(x'\beta) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$$

- Los efectos marginales
  $$\frac{\partial \mathrm{E}(y_i|x_i)}{\partial x_{ji}} = \beta_j \times \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$$
  $$= \beta_j \times \mathrm{E}(y_i|x_i)$$

- Un cambio de una unidad en el regresor j conduce a un cambio en la media condicional de $\beta_j \times \mathrm{E}(y_i|x_i)$

- Otra forma de decir esto es que un cambio de una unidad en el regresor j conlleva un cambio **proporcional** de $\mathrm{E}(y_i|x_i)$ de β_j.
- Si el regresor está en logaritmos entonces β_j es una elasticidad.
- Supón que $\mathrm{E}(y_i|x_{1i}) = \exp(\beta_0 + \beta_1 \log(x_{1i}))$ y x_1 mide exposición (por ejemplo, el tiempo en el que se ha medido el recuento): esperamos que $\beta_1 = 1$.
- Por ejemplo, toma y_i =número de accidentes de tráfico en una semana, x_{1i} =volumen de tráfico, y esperaríamos que $\beta_1 = 1$:

$$\mathrm{E}(y_i|x_{1i}) = e^{\beta_0} x_{1i}$$

Algunos problemas cuando y no es una Poisson

- Mientras $\mathrm{E}(y_i|x_i) = e^{x_i'\beta}$, entonces MV seguirá siendo consistente (pero los errores estándar deben ser computados de forma diferente).
- Por ejemplo, si la varianza es mayor que la media (**sobre-dispersión**)
  - Se puede contrastar: $H_0 : \mathrm{E}(y) = \mathrm{Var}(y)$.
  - `gretl` proporciona un test de sobre-dispersión.
  - En caso de rechazo, debe usarse la opción `--robust`.
- A veces podemos evitar la sobre-dispersión incluyendo controles adicionales.
- Alternativamente podemos estimar un modelo que es más flexible, como el **modelo de distribución binomial negativa** (en `gretl`: `negbin`—fuera del objetivo de este curso).

Inflación de ceros

Otro problema común con la regresión de Poisson es el exceso de ceros:

- Supongamos que hay dos procesos:
  - Uno para determinar si hay cero eventos o cualquier evento,
  - Un proceso de Poisson que determina cuántos eventos existen.
- Aquí habrá más ceros que lo que predeciría una regresión de Poisson.
- Un ejemplo sería la distribución de cigarrillos fumados en una hora por miembros de un grupo en el que algunos individuos son no fumadores.
- Un modelo alternativo, **el modelo de Poisson con inflación de ceros**, es mejor en estos casos (en gretl: ver sección 20.7 en el manual—más allá del alcance de este curso)

Regresión de Poisson en gretl

poisson *depvar indepvars* [; *offset* --robust]

- *depvar* debe tomar valores enteros no negativos.
- Se puede opcionalmente añadir una variable estrictamente positiva a la especificación (*offset*) si se espera que el número de ocurrencias del evento sea proporcional a la misma.
  - Por ejemplo, el número de accidentes de tráfico es, ceteris paribus, proporcional al volumen de tráfico.
- Automáticamente se muestra un test de sobre-dispersión. La hipótesis nula es ausencia de sobre-dispersión. En caso de rechazo:
  - La inferencia en el modelo debe llevarse a cabo usando la opción `--robust`
  - Un modelo binomial negativo sería más apropiado (`gretl`: comando `negbin`).

poisson DVISITS const SEX age income --robust

Model 4: Poisson, using observations 1–5190 ($n = 5111$)
Missing or incomplete observations dropped: 79
Dependent variable: DVISITS
QML standard errors

	Coefficient	Std. Error	z	p-value
const	−0.954136	0.112619	−8.4723	0.0000
SEX	0.229700	0.0869432	2.6419	0.0082
age	0.524717	0.0753856	6.9604	0.0000
income	−0.141529	0.0571679	−2.4757	0.0133

Mean dependent var	0.300724	S.D. dependent var	0.793169
Sum squared resid	3152.281	S.E. of regression	0.785651
McFadden R^2	0.027979	Adjusted R^2	0.026957
Log-likelihood	−3803.936	Akaike criterion	7615.873
Schwarz criterion	7642.029	Hannan–Quinn	7625.030

Overdispersion test: $\chi^2(1) = 124.089$ [0.0000]

Resumen

- Con datos de recuento, OLS puede no ser implementable o dar resultados no satisfactorios.
- Una buena alternativa es el modelo de regresión de Poisson.
- Este modelo puede ser estimado por MV y la interpretación de los resultados es relativamente sencilla.
- Si los datos presentan sobre-dispersión, entonces la inferencia debe llevarse a cabo con errores estándar robustos.
- Otros modelos alternativos pueden ser más adecuados en presencia de sobre-dispersión o con muchos ceros.