Poisson Regression

Quantitative Microeconomics

R. Mora

Department of Economics
Universidad Carlos III de Madrid

Outline

Introduction

Example 1

Relation between Research and Development (R&D) expenditures of firms and the number of patents received by them.

- The dependent variable, number of patents, is a **count** of the total number of patents of a particular firm during a period of time, such as a year.
- Only non-negative and discrete numbers are possible: some firms have zero patents, while other firms may many patents.

Example 2

Use of health care services among the elderly.

- The dependent variable, number of doctor visits, is a **count** of the total number of visits to the doctor an elderly individual has made during last year.
- only non-negative and discrete numbers are possible: some individuals have zero visits, while other individuals may have more than ten visits a year.

Example 3

The effect of prison sentences on criminality.

- The dependent variable, number of arrests, is a **count** of the total number of arrests a man has on a given year.
- only non-negative and discrete numbers are possible: some individuals have zero arrests, while other individuals may have several arrests.

OLS with count data

- In principle, if the dependent variable is positive, $y_i > 0$, we could take its logarithm and assuming

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

  OLS would provide consistent estimates of vector β.
- However:
  - if $y = 0$ then there is no logarithmic transformation (adhoc solutions: $\ln\left(y+1\right)$, use $\ln\left(0.5\right)$ when $y = 0$).
  - for prediction we want to predict $\mathrm{E}\left(y\right)$, but $\exp\left(\mathrm{E}\left(\ln y\right)\right) \neq \mathrm{E}\left(y\right)$ even though $\exp\left(\ln y\right) = y$.
- In this session, we are going to study a nonlinear model which avoids these two problems.

The Poisson distribution

Definition of Poisson distribution

- A discrete random variable Z has a Poisson distribution with parameter λ if the probability that $Z = z$ is

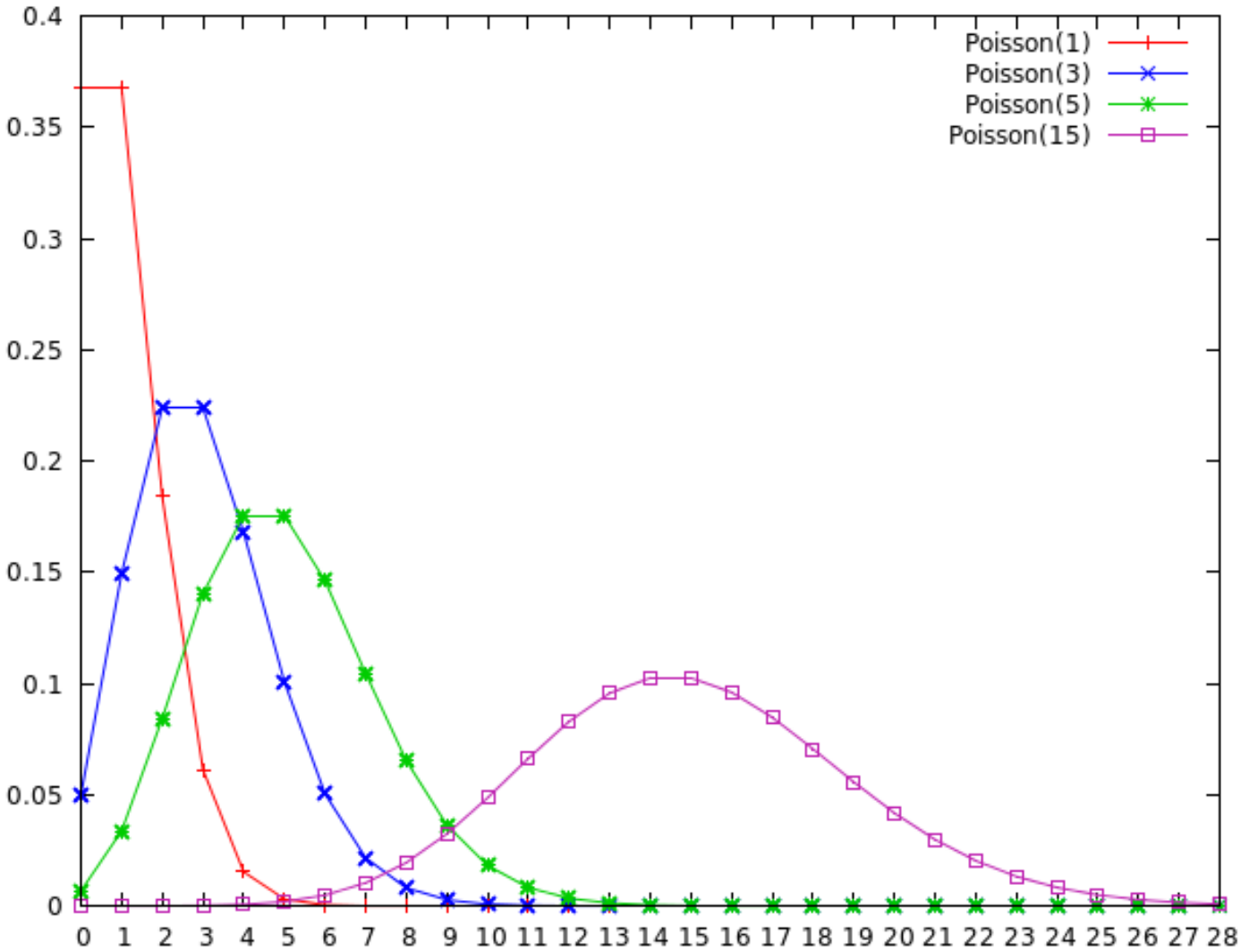
$$\Pr(Z = z) = \frac{\lambda^z}{z!} e^{-\lambda}$$

where

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

- Parameter λ is equal to both the expectation and the variance of Z (equidispersion)

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


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

Intuition

- The Poisson distribution is the probability of a given number of events occurring in a fixed interval of time and/or space if these events
  - Occur with a known average rate.
  - Independently of the time since the last event.
- Famous historical examples include (from Wikipedia):
  - The number of wrongful convictions in a given country in a given period of time.
  - The number of soldiers in the Prussian army killed accidentally by horse kicks.

Poisson regression

- It is a form of regression analysis to model count data.
- Let y be the dependent variable and x a vector of independent variables.
- Poisson regression assumes:
  - y has a Poisson distribution,
  - the expectation (and variance) of y given x is

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

  - Thus, $\log(\mathrm{E}(y|x)) = x'\beta$ (the Poisson regression model is sometimes referred to as the log-linear model).

ML estimation

- The Poisson distribution's probability mass function is given by
  $\Pr(y|x) = \frac{e^{yx'\beta}e^{-e^{x'\beta}}}{y!}$
- For a sample with N observations,
  $\Pr(y_1,...,y_N|x_1,...,x_N) = \prod_{i=1}^{N} \frac{e^{y_i x_i'\beta}e^{-e^{x_i'\beta}}}{y_i!}$
- The log-likelihood is:

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

(since $y_i!$ does not depend on b, we may drop it)

  - the FOCs can only be solved numerically. Fortunately, the log likelihood is concave, so standard numerical methods work well.

Interpretation

- We have
  $$\mathrm{E}(y|x) = \exp(x'\beta) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$$
- Marginal effects
  $$\begin{aligned}\frac{\partial \mathrm{E}(y_i|x_i)}{\partial x_{ji}} &= \beta_j \times \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k) \\ &= \beta_j \times \mathrm{E}(y_i|x_i)\end{aligned}$$
- A one unit change in the jth regressor leads to a change in the conditional mean by $\beta_j \times \mathrm{E}(y_i|x_i)$

- Another way of saying this is that a one unit change in the jth regressor leads to a **proportionate change** in $\mathrm{E}(y_i|x_i)$ of β_j.
- If the regressor is the log of a variable, then β_j is an elasticity.
- Suppose that $\mathrm{E}(y_i|x_{1i}) = \exp(\beta_0 + \beta_1 \log(x_{1i}))$ and x_1 measures exposure (such as time): we expect $\beta_1 = 1$.
- For example, take y_i =number of traffic accidents, x_{1i} =volume of traffic, and we would expect $\beta_1 = 1$:

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

Some issues

- If $\mathrm{E}(y_i|x_i) = e^{x_i'\beta}$, then the Poisson MLE estimate is consistent even if y_i is not Poisson distributed. MLE standard errors and t-statistics need to be adjusted.
- For example, if the variance is larger than the mean (**over-dispersion**).
  - this can be tested: $H_0 : \mathrm{E}(y) = \mathrm{Var}(y)$.
  - `gretl` does one test of over-dispersion automatically.
  - in case of rejection, the `--robust` option must be used for the `poisson` command.
- Sometimes we can avoid over-dispersion by including additional explanatory variables.
- Alternatively, one can try to estimate a more flexible model, such as the **negative binomial model** (in `gretl`: `negbin`—beyond the scope of this course).

Zero inflation

Another common problem with Poisson regression is excess zeros:

- Assume there are two processes at work:
  - One determining whether there are zero events or any events.
  - A Poisson process determining how many events there are.
- Here there will be more zeros than a Poisson regression would predict.
- An example would be the distribution of cigarettes smoked in an hour by members of a group where some individuals are non-smokers.
- An alternative model, the **zero-inflated Poisson model**, functions better in these cases (in `gretl`: see section 20.7 in manual—beyond the scope of this course).

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

- depvar must take on only non-negative integer values.
- Optionally, you may add an strictly positive variable to the specification (*offset*) if you expect the number of occurrences of the event to be proportional to it.
  - For example, the number of traffic accidents might be proportional to traffic volume, other things equal.
- An over-dispersion test is automatically computed and displayed. The null of the test is absence of over dispersion. In case of rejection:
  - Inference in the model should be carried out using the --robust option
  - A negative binomial model would be more appropriate (gretl: negbin command).

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]

Summary

- With count data, OLS may not give adequate results.
- A good alternative is the Poisson regression model.
- This model can be estimated by ML, and interpretation of the results is relatively straightforward.
- If the data presents over dispersion, then inference must be carried out with robust standard errors.
- Other alternative models may be better in the presence of over-dispersion or with many zeros.