Applied Statistical Modelling Lecture 6: Poisson regression

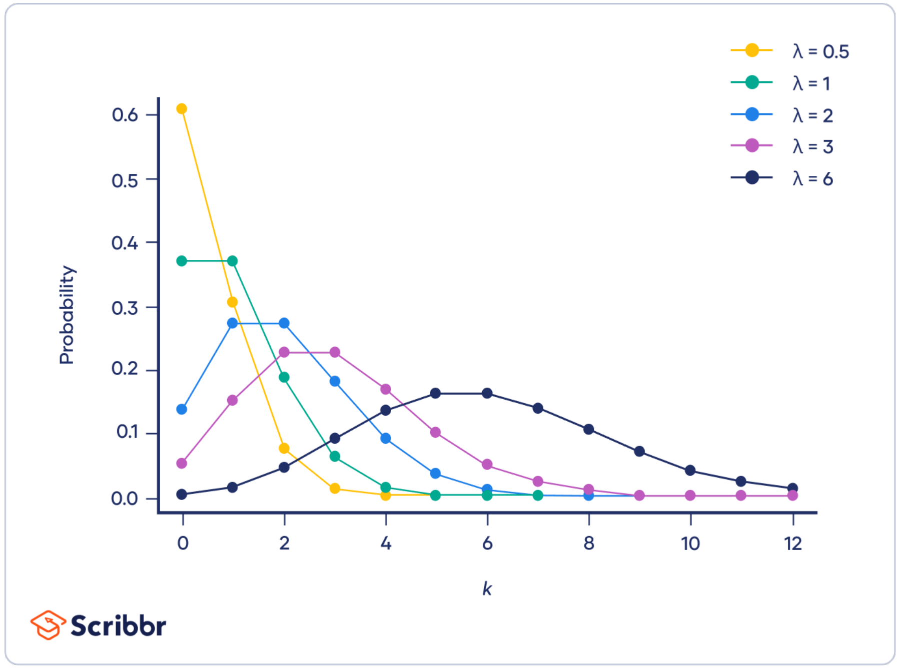
Prof.ssa Chiara Seghieri

Laboratorio di Management e Sanità, Istituto di Management, L'EMbeDS
Scuola Superiore Sant'Anna, Pisa
c.seghieri@santannapisa.it
c.tortu@santannapisa.it

Poisson Model

- Count response variable: occurrence count recorded for a particular measurement window (length of time, but it can also be a distance, area, etc). number of calls to a customer service in an hour or number of flaws in a tabletop of a certain area.

- Poisson distribution: its minimum value is zero and, in theory, the maximum is unbounded. Probability of 0, 1, 2, . . . Events. **The mean of the distribution is equal to the variance**.



$$P(Y = k) = \frac{e^{-\lambda}\lambda^{k}}{k!} \qquad \text{K=0,1,2,...}$$

$$\lambda = media\ e\ varianza\ di\ Y$$

We'd like to model our main parameter λ, the average number of occurrences per unit of time or space, as a function of one or more covariates.

For the random component, we assume that the response Y has a Poisson distribution: $Y_i \sim Poisson(\lambda_i)$ where the expected count of Y_i is $E(Y_i) = \lambda_i$. The link function is usually the (natural) log, but sometimes the identity function may be used. The systematic component consists of a linear combination of explanatory variables, identical to that for logistic regression. Thus, in the case of a single explanatory, the model is:

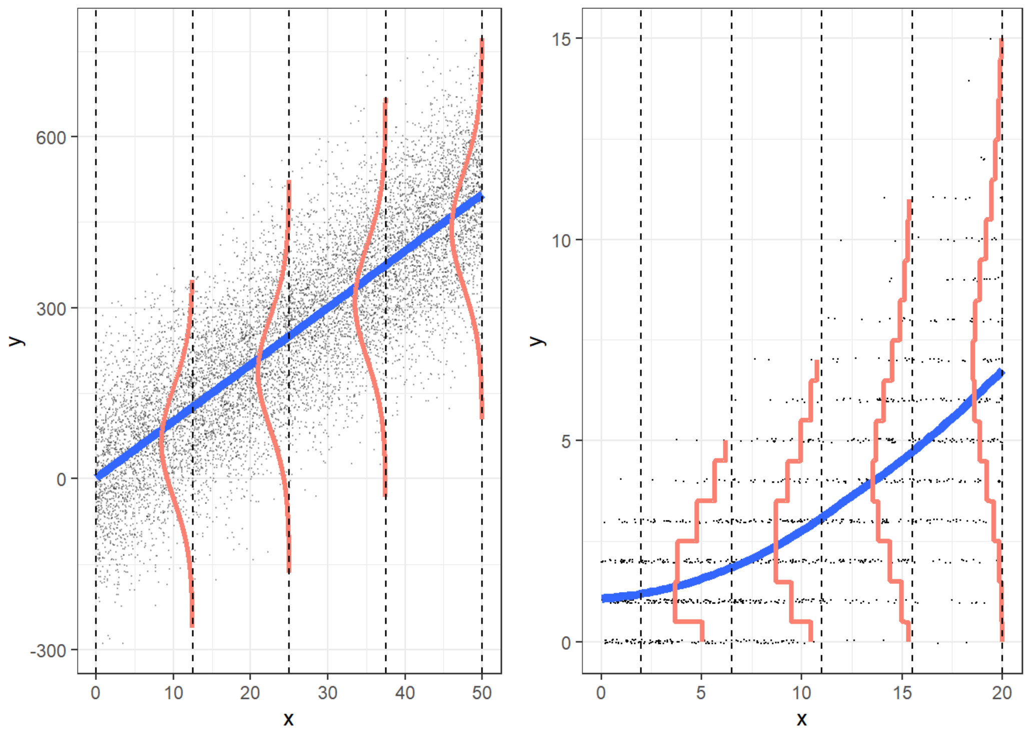
$$\log(\lambda) = \alpha + \beta x$$

This is equivalent to:

$$\lambda = \exp(\alpha + \beta x) = \exp(\alpha) \exp(\beta x)$$

Interpretations of these parameters are similar to those for logistic regression: $\exp(\alpha)$ is the effect on the mean of Y when x=0, and $\exp(\beta)$ is
the multiplicative effect on the mean of Y for each 1-unit increase in x.

Regression models: Linear regression (left) and Poisson regression (right)

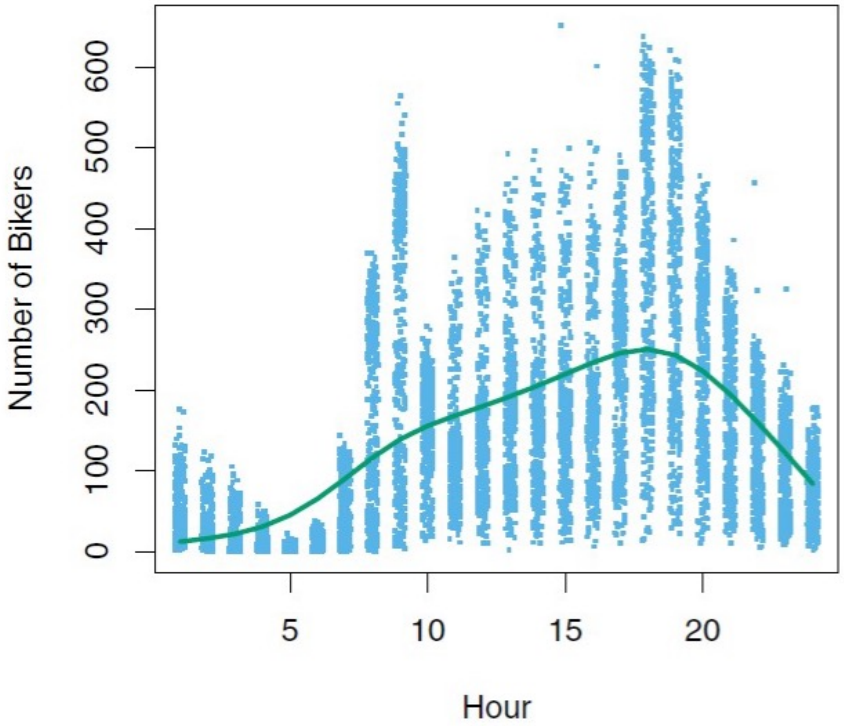


Linear regression: for each level of X, the responses are approximately normal.
Poisson regression: for each level of X, the responses follow a Poisson distribution.
For Poisson regression, small values of λ are associated with a distribution that is noticeably skewed with lots of small values and only a few larger ones. As λ increases the distribution of the responses begins to look more and more like a normal distribution

Number of bikers per hour. The variance of biker numbers changes as the mean number changes:

•during worse conditions, there are few bikers, and little variation in the number of bikers

•during better conditions, there are many bikers on average, but also larger variation in the number of bikers



From: An Introduction to Statistical Learning

If the observations recorded correspond to different measurement windows, a scale adjustment has to be made to put them on equal terms, and we model the rate or count per measurement unit.

Example: Gardner, Mulvey, & Shaw (1995), Psychological Bulletin, 118.

Y = Number of violent incidents exhibited over a 6 month period by patients who had been treated in the ER of a psychiatric hospital.

During the 6 months period of the study, the individuals were primarily residing in the community. The number of violent acts depends on the opportunity to commit them; that is, the number of days out of the 6 month period in which a patient is in the community (as opposed to being locked up in a jail or hospital).

Y = count (number violent acts).
t = index of the time or space (i.e days in the community).

The sample rate of occurrence is Y/t.

The sample rate of occurrence is Y/t.

The expected value of the rate is:

$$E\left({}^{Y}\!/_{t}\right) = \frac{1}{t}\mathrm{E(Y)} = {}^{\lambda}\!/_{t}$$

Therefore, the Poisson log-linear regression model for the expected rate is:

$$\log(\lambda) = \alpha + \beta x + \log(t)$$

The term "$-\log(t)$" is an adjustment term and each individual may have a different value of t. $-\log(t)$ is referred to as an "offset".

$$\log\left({}^{\lambda}\!/_{t}\right) = \alpha + \beta x$$

The expected value of counts depends on both t and x

$$\lambda = te^{\alpha}e^{\beta x}$$

If the variance doesn't behave as the mean structure suggests for that distribution (i.e.many times data admit more variability than expected under the assumed distribution) then we have a problem with **overdispersion** (or occasionally, underdispersion).

Situations of overdispersion:

✓ Dispersion up to a constant term: A simple situation when the true variance is proportional to what the model predicts. For Poisson model: $var(Y)=\theta\lambda$, $\theta>1$

✓ Structural zeros: some subjects are not subject to any risk. E.g., number of cigarettes smoked per day among the non-smokers. The probability of having outcomes greater than 0 for non-smokers is 0, but it could follow a Poisson for smokers.

All the above situations will lead to a problem of over-dispersion.

Issues related to the Poisson regression

B) Quasi Maximum Likelihood estimation

"quasi-maximum likelihood" (quasi-ML) is a method used to estimate parameters in situations where the likelihood function is not strictly adhering to the assumptions of standard maximum likelihood estimation.

If you face overdispersion then quasi-ML can be used to account for this by using a different likelihood function or incorporating extra parameters into the model.

Consequence of overdispersion:

- Standard errors will be underestimated
- Reject H0 when we shouldn't

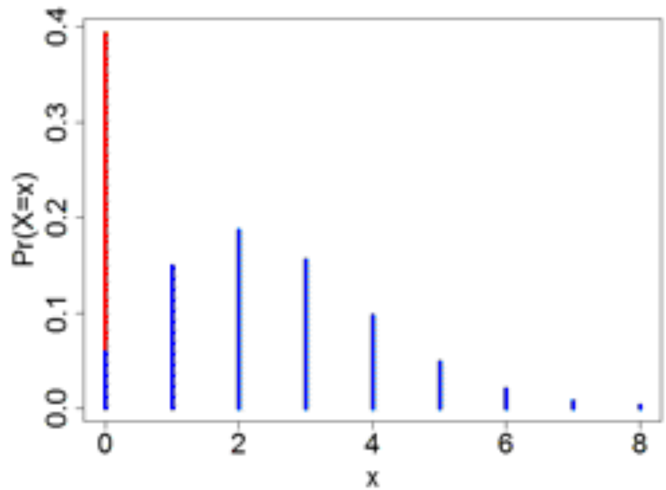
If zero mass and overdispersion (for the non-zero observations): ZINB might help

Issues related to the Poisson regression

A) Issues with count data

- A1) Zero Inflated Distribution: distribution of the Yis characterized by a peak on the 0
→ Zero Inflated Poisson (ZIP): mixture distribution that separately models the zeros.

Two groups of people: always zero and not always zero. Mixture of probability from the two groups. The probability Pr(X=0) has two components. The blue part is from the underlying Poisson distribution, the red part is due to zero-inflation.



there is a probability θ of drawing a zero, and a probability 1−θ of drawing from Poisson(λ). The probability function is thus:

$$\mathrm{P}(y_n \,|\theta, \lambda) = \begin{cases} \theta + \ (1-\theta) * Poisson(0|\lambda) \ \ if\ y_n = 0 \\ \quad (1-\theta) * Poisson(y_n|\lambda) \ \ if\ y_n > 0 \end{cases}$$

Alternative models:

Zero-inflated negative binomial (ZINB)
Hurdle models (two-parts)

Lambert, Diane. 1992. "Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing." *Technometrics* 34 (1).