Probability and Statistics 2 (GCED)
Models for Poisson response

Marta Pérez-Casany and Jordi Valero Baya

Department of Statistics and Operations Research
Technicat University of Catalonia

Facultat d'Informàtica de Barcelona, First Semester

Poisson response

A r.v. $Y \sim \mathrm{Po}(\lambda)$, $\lambda > 0$, if and only if takes values in $\mathbb{Z}^+$ with probabilities:

$$\Pr\{Y = k\} = e^{-\lambda}\,\frac{\lambda^k}{k!}, \quad \forall k \in \mathbb{Z}^+.$$

It verifies:

$$E(Y) = \lambda, \quad \text{y} \quad Var(Y) = \lambda$$

From where the **Dispersion index**

$$I(Y) = \frac{Var(Y)}{E(Y)} = 1$$

The Poisson distribution appears as a limit of Binomial distributions.

If $X_n \sim Bin(n, p_n)$ and one assumes that:

1) $n \to +\infty$
2) $p_n \to 0$ when $n \to +\infty$
3) $n \cdot p_n \to \lambda$ when $n \to +\infty$

then,

$$X_n \to Y, \text{ where } Y \sim Po(\lambda)$$

Comment: That's why it is called *Law for rare events*

Relation with other distributions

1) **Multinomial** If $Y = (Y_1, Y_2, \cdots, Y_n)$, Y_i i.i.d $Y_i \sim Po(\lambda_i)$ then

$$Y|\sum_{i=1}^{n} Y_i = N \sim \mathrm{Multinomial}(N, p_1, p_2, \cdots, p_n)$$

with $p_i = \frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i}$.

Important: $Y_i|N \sim Bin(N, p_i)$ y $corr(Y_i|N, Y_j|N) = -Np_ip_j$.

2) **Exponencial** Fixing a time interval $(0, t)$, if T is the r.v. corresponding to time between independent events, and Y is lthe r.v. corresponding to number of events in $(0, t)$, then one has that

$$T \sim \mathrm{Exponencial}(\lambda) \Longleftrightarrow Y \sim \mathrm{Po}(\lambda\, t)$$

Poisson response with covariates

Question: Why it has no sense to consider that

$$E(Y_i) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{i(p-1)}\beta_{p-1}$$

when $Y_i \sim \mathrm{Po}(\lambda_i)$?

Three important reasons:

1) $Var(Y_i) = \lambda_i$,
2) we do not have normality,
3) $(X\beta)_i \in \mathbb{R}$ while $\lambda_i \in (0, +\infty)$.

The function $g(\lambda)$ reasonable to be linear with the covariates is

$$g(\lambda) = \log(\lambda);$$

since it transforms $(0, +\infty)$ in all the real line.

Coment: Models with Poisson response are also called **log-lineal models**

Parameter Estimation

Maximum likelihood function:

$$L(\mu; y) = exp\Big(- \sum_{i=1}^{n} \mu_i \Big) \Pi_{i=1}^{n} \frac{\mu_i^{y_i}}{y_i!}.$$

assuming $\log(\mu) = X\beta$, the log-likelihood function, with the exception of a constant, is equal to:

$$l(\beta; y) = \sum_{i=1}^{n} \Big(y_i \sum_{j=1}^{p} x_{ij}\beta_j - \exp\big(\sum_{j=1}^{p} x_{ij}\beta_j \big) \Big).$$

$$U = \frac{\partial l}{\partial \beta} = 0 \Longleftrightarrow X^t(Y - e^{\mu}) = 0.$$

Observation: $X^t y$ is a suficient statistic for β.

Goodness of fit I

The Pearson statistic for the Poisson is equal to:

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} = \sum_{i=1}^{N} r_i^2$$

If the model is correct, X^2 assymtotically follows a χ^2_{N-p}.

Thus, we can reject our model when $X^2 \geq \chi^2_{\alpha,N-p}$.

The values signed r_i are called Pearson residuals and when plotted they should follow approximately a standarized Normal distribution.

Goodness of fit II

The Deviance for the Poisson (scaled deviance, because $\phi = 1$) is equal to:

$$D = 2[l(\hat{p}_{i,fullm}; y) - l(\hat{p}_{i,ourm}, y)] = 2\sum_{i=1}^{N} y_i \log \frac{y_i}{\hat{\mu}_i} = \sum_{i=1}^{N} d_i^2$$

Obs: if for some i $y_i = 0$, the corresponding term in D is taken to be equal to zero.

Under the hypothesis that our model is correct, $D \sim \chi^2_{N-p}$, and we reject our model then $X^2 \geq \chi^2_{\alpha,N-p}$.

The values signed d_i are known as deviance residuals and asymptotically follow a standarized normal distribution.

Concept of *offset* variable

Offset: *A term used in GLM to indicate a known regression coefficient that is to be included in the model, i.e. one that does not have to be estimated.* (Según *The Cambridge Dictionary of Statistics* de Everitt).

Assume that $Y_{ij} \sim \mathrm{Po}(\lambda_{ij})$, where $\lambda_{ij} = N_{ij}\, p_{ij}$.

Examples

EXAMPLE 1

Let Y_{ij} be the r.v. corresponding to the **Number of deaths** in a given period of time, in the i-thm district of a given city of people in a given range of age denoted by j. Assume that the city has a different districts and that b different ranges of age are considered.

N_{ij} would be the number of individuals that live in the i-thm district and which age belongs to the j-thm range,

p_{ij} is the probability that someone in the i-thm distric and j-thm range of age die.

The appropiate model could be:

$$\log(E(Y_{ij})) = \log(\mu_{ij}) = \log N_{ij} + \tau_i + \beta_j \quad i = 1, \cdots, a, \; j = 1, \cdots, b$$

and the coefficient of $\log(N_{ij})$ should not be estimated since it is assumed to be equal to one.

EXAMPLE 2

Let Y_{ijk} be the **Number of claims** received in an insurance company, corresponding to car policyholders.

Factors that has sense to consider:

1) Distance (5 levels)
2) Type of Bonus of the policyholder (7 levels)
3) Car brant (9 levels)

If T_{ijk} is the population at risk corresponding to the factors level combination (i, j, k),

$$\log(E(Y_{ijk})) = \log(\mu_{ijk}) = \mu + \tau_i + \beta_j + \gamma_k + \log(T_{ijk})$$

$i = 1, \cdots, 5$, $j = 1, \cdots, 7$ y $k = 1, \cdots, 9$.

Analyzing a contingence table by a Poisson GLM

Assume that one is interested in classifiying N experimental units depending on two (or more) cathegorical variables with a and b levels respectively.

This gives place to a contingence table of the form:

	1	2	$\cdots$	b	
1	n_{11}	n_{12}	$\cdots$	n_{1b}	$n_{1.}$
2	n_{21}	n_{22}	$\cdots$	n_{2b}	$n_{2.}$
$\vdots$					
a	n_{a1}	n_{a2}	$\cdots$	n_{ab}	$n_{a.}$
	$n_{.1}$	$n_{.2}$		$n_{.b}$	N

Observation: In this situation the response variables are the cathegorical variables.

If N **is known**, the vector $Y = (Y_{ij})$ follows a Multinomial(N, p_{ij}), and we are not going to consider how to analize the table.

If N **is unknown** and we denote by Y_{ij} the r.v. corresponding to the number of observations in the cell (i, j), $Y_{ij} \sim \mathrm{Po}(\mu_{ij})$ where $\mu_{ij} = N\, p_{ij}$, the table can be analyzed using a GLM with a Poisson response

If we consider a model with two categorical explanatory variables we have:

$$\log(E(Y_{ij})) = \log N + \beta_0 + \tau_i + \gamma_j \quad i,j \tag{1}$$

and the number of parameters of the model is:

$$1 + (a-1) + (b-1) = a + b - 1.$$

If we add the interaction term, and instead consider the model:

$$\log(E(Y_{ij})) = \log N + \beta_0 + \tau_i + \gamma_j + \gamma_{ij} \quad i,j \tag{2}$$

the number of parameters is:

$$a + b + a + (a-1)(b-1) = ab$$

and thus, it s equivalent to the full model.