Generalized Linear Models

Lecture 8. Count data models I. Poisson model

Count data

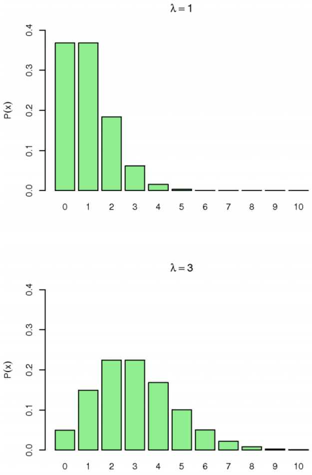
Count data – counting the number of events in a time interval

Common areas of use: insurance, reliability, medicine

- number of insurance claims/losses
- number of failures of a system
- number of patients arriving in an emergency room
- number of customer entering a shop
- number of raisins in a bun

- Poisson distribution occurs while counting events such that the probability of an event is quite small
- Typical histogram is asymmetric
- The smaller the parameter, the more skewed histogram
- The parameter of the distribution can be interpreted as the average number of events in a time unit
- Poisson distribution is also called **the law of small numbers**

Poisson distribution in case of different parameters



Poisson distribution (Simeon-Denis Poisson, 1781 – 1840)

Definition

Discrete r.v. Y has Poisson distribution, $Y \sim Po(\mu)$, with parameter μ ($\mu > 0$), if its pmf has the following form:

$$p(y) = \frac{e^{-\mu}\mu^y}{y!}, \quad y \in \{0, 1, 2 \dots\}$$

Well-known properties:

1. Equidispersion property: $\mathbf{E}Y = \mathbf{D}Y = \mu$
2. Additivity: if $Y_1, \dots, Y_m$ are independent, $Y_i \sim Po(\mu_i)$, then the sum $Y = \sum_{i=1}^m Y_i$ is also Poisson distributed, $Y \sim Po(\mu)$, where $\mu = \sum_{i=1}^m \mu_i$
3. Poisson limit theorem (law of rare events): if the sample size $n \to \infty$ and the probability of an event p is small, $np \to const$, $B(n, p) \to Po(np)$
4. In case of big samples and big μ, $Po(\mu) \to N(\mu, \mu)$

Property 2 implies that we can consider grouped and ungrouped data in a similar way

A classical example of Poisson model

Prussian army horse kick data (Vladislav Bortkiewicz, 1898)

Contains deaths by year and corp from horse kicks in Prussian army during 1875–1894.

Data (grouped by year):

Year	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
Deaths	3	5	7	9	10	18	6	14	11	9	5	11	15	6	11	17	12	15	8	4

In R, the dataset is available in library `pscl`:

```
library(pscl)
data(prussian)
```

Poisson distribution as a member of exponential family

Let us rewrite the pmf to match the form of exponential family:

$$
\begin{aligned}
p(y_i) &= \exp(\log \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}) = \exp[\log(\exp(-\mu_i)) + \log \mu_i^{y_i} - \log y_i!] \\
&= \exp[y_i \log \mu_i - \mu_i - \log y_i!]
\end{aligned}
$$

- $\theta_i = \log(\mu_i)$
- $b(\theta_i) = \mu_i = \exp(\theta_i)$
- $\varphi = 1$
- $\mathbf{E} Y_i = b'(\theta_i) = \mu_i$
- $\mathbf{D} Y_i = \varphi b''(\theta_i) = \mu_i$

Prove it!

$\Rightarrow$ Poisson distribution belongs to exponential family

GLM with Poisson distributed response

Sample: n observations, $y_1, \ldots, y_n$, are considered as realizations from $Y_i \sim Po(\mu_i)$

Mean μ_i (and thus also the variance!) depends on arguments $\mathbf{x}_i$

Canonical link for Poisson model is log-link, which produces log-linear model:

(1) Log-linear (multiplicative) Poisson model

$$g(\mu_i) = \ln(\mu_i), \mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta}), \quad \text{the effect of arguments is multiplicative}$$

Coefficient β_j corresponds to the change in natural logarithm of the mean of the response variable if there is a unit change in j-th argument
Another possible choice is to use the identity link:

(2) Linear (aditive) Poisson model

$$\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}, \quad \text{the effect of arguments is additive}$$

Problems with the linear model:

- the range of values of the right is not restricted
- the left hand side (the response) can only take positive values

Multiplicative Poisson model

Link function **Log**: $g(\mu_i) = \ln(\mu_i), \quad \ln \mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$

GLM for the mean: $\mu_i = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$

$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}) = \exp(\beta_0) \cdot \exp(\beta_1 x_{i1}) \cdot \ldots \cdot \exp(\beta_k x_{ik})$$

Multiplicative effect of an argument x_{ij} (if all the other conditions remain the same): change of x_{ij} by one unit corresponds to change of μ_i e^{β_j} times
The empirical studies also suggest the use of log-link:

- In case of count data, the effect of arguments is more often multiplicative than additive: typically the effect to bigger counts is big and to smaller counts is small.
- The effects of arguments tend to be proportional to the number of events and the use of *log-scale* produces a simpler model and is justified

In case of small values ($y_i \approx 0$), there can occur problems with the log-transform and in that case a small adjustment is suggested: $y_i = y_i + c$

Model fitting

Log-likelihood for observation i:

$$l_i(\beta) = y_i \log(\mu_i) - \mu_i - \log(y_i!)$$

Sample log-likelihood:

$$l(\beta) = \sum_i l_i(\beta) = \sum_i y_i \ln(\mu_i) - \sum_i \mu_i + c$$

Score equations (assuming canonic link, i.e. $\mu_i = \mu(\mathbf{x}_i, \beta) = \exp(\mathbf{x}_i^T \beta)$):

$$s_j(\beta) = \frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - \mu_i) x_{ij} = 0, \quad j = 1, \ldots, p$$

$\Rightarrow$ in general, $s(\beta) = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\mu}) = 0$, which yields $\boldsymbol{X}^T \boldsymbol{y} = \boldsymbol{X}^T \boldsymbol{\mu}(\hat{\beta})$

An important corollary: in a model with intercept term

$$\sum_i (y_i - \hat{\mu}_i) = 0$$

Why?

Deviance

Let us recall that the deviance is defined as 2 times the difference between saturated and current model: $D = 2(l(\boldsymbol{y}, \boldsymbol{y}) - l(\boldsymbol{y}, \hat{\boldsymbol{\mu}}))$

For each observation i, the log-likelihoods are:

- For current model: $l_i(\beta) = y_i \ln(\hat{\mu}_i) - \hat{\mu}_i - \ln(y_i!)$
- For saturated model: $l_i(\beta) = y_i \ln(y_i) - y_i - \ln(y_i!)$

The deviance is thus:

$$D = 2\sum_i (y_i \ln \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)), \quad D \overset{a}{\sim} \chi^2_{n-p}$$

Note that because of the corollary from previous slide we can express the deviance in a model with intercept term similarly to binomial:

$$2\sum o_i \ln \frac{o_i}{e_i}$$

Difference of deviances is asymptotically χ^2-dist. (even if the deviance itself is not)
For two models: M_r – r params, deviance D_r, and M_s – s params, D_s, $\quad s > r$
$D_s - D_r \sim \chi^2_{s-r}$, i.e. we can decide whether to include these $s - r$ arguments

Pearson χ^2-statistic

$$\chi_P^2 = \sum \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

Asymptotic (for both D and χ_P^2) holds if $\mu_i \to \infty$

Note that in case of grouped data, the formulas implicitly contain n_i ($\hat{\mu}_i = n_i \tilde{\mu}_i$ and we can think of fixed cell asymptotic $n_i \to \infty$)

- The use of deviance and the Pearson statistic depends on whether asymptotic results apply
- Usually one expects all of the means to be larger than three
- If D and χ_P^2 are quite different, one might suspect that the approximation is inadequate

More details: Tutz, p 187

Overdisperion

Overdispersion is a serious problem for Poisson model

If the model fits, it must hold that $EY_i = DY_i = \mu_i$

Overdispersion is usually modelled via a scale parameter φ: $DY_i = \varphi EY_i$,

- $\varphi = 1$, no problem
- $\varphi > 1$, i.e. $DY_i > EY_i \Rightarrow$ overdispersion
- $\varphi < 1$, i.e. $DY_i < EY_i \Rightarrow$ underdispersion

If the model is correct (deviance and Pearson's χ^2-statistic are asymptotically χ^2-distributed), the following holds:

$$\frac{D}{df} \approx 1 \qquad \frac{\chi^2}{df} \approx 1$$

If the ratio > 2, overdispersion needs to be addressed

Simplest option: estimate the scale and take it into account

$$\hat{\varphi} = \frac{D}{n-p}, \quad \hat{\varphi} = \frac{\chi^2}{n-p}$$

Example. Overdispersion (Prussian army data), 1

```
> prussian2=sqldf("select sum(y) as y,
                   year from prussian group by year")
> modelP=glm(y~year,family="poisson",data=prussian2)
> summary(modelP)
...
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.69095    1.06056   0.651    0.515
year         0.01876    0.01243   1.509    0.131

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 38.503  on 19  degrees of freedom
Residual deviance: 36.216  on 18  degrees of freedom
```

Example. Overdispersion (Prussian army data), 2

```
> modelP_orig=glm(y~year+corp,family="poisson",data=prussian)
> summary(modelP_orig)
...
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.815e+00  1.087e+00  -1.669   0.0951 .
year         1.876e-02  1.243e-02   1.510   0.1312
corpI        3.850e-09  3.535e-01   0.000   1.0000
...
corpVIII    -8.267e-01  4.532e-01  -1.824   0.0681 .
corpX       -6.454e-02  3.594e-01  -0.180   0.8575
corpXI       4.463e-01  3.202e-01   1.394   0.1633
corpXIV      4.055e-01  3.227e-01   1.256   0.2090
corpXV      -6.931e-01  4.330e-01  -1.601   0.1094
...
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 323.23  on 279  degrees of freedom
Residual deviance: 294.81  on 265  degrees of freedom
```

Reasons of overdispersion

Main reasons (small/apparent overdispersion):

- Systematic component of the model is not correctly estimated (some significant argument or interaction is missing). Think of the Prussian army example!
- Scale of the arguments is not the best for the exercise, a scale transform can help (e.g. *log*)
- Outliers in data

If overdispersion is small (< 5), the first step is to check the model structure $\Rightarrow$ if the structure is ok, overdispersion needs to be taken into account

If overdispersion is big (> 5), something must be wrong... $\Rightarrow$ Poisson distribution does not fit

Big (real) overdispersion indicates extra randomness in data. That can be caused by:

- Poisson process in an interval with random length
- excess zeros or missing zeros in data

Solving the overdispersion problem (small overdispersion)

Things to try:

1. Check for outliers
2. Modify the systematic component of the model (incl. interaction terms), scale transforms of arguments
3. Use variance stabilizing transforms $y^{1/2}$, $y^{2/3}$ (*Anscombe*, 1953)
4. Use *quasi-likelihood*, assume that $\mathbf{D}Y_i = \varphi\mu_i$, estimate $\hat{\varphi}$ and adjust the covariance matrix of arguments accordingly

Possibilities to check for overdispersion:

- analysis of generalized residuals $\rightarrow$ outliers
- visualization, e.g. plot $\bar{y}_i$ vs s_i^2
  in ideal case the points should lie close to bisector
- different tests

Solving the overdispersion problem (big overdispersion)

Poisson distribution does not fit, possible reasons:

1. no zeros
2. too many zeros
3. mixture of distributions
4. censored data
5. truncated data
6. counting depends on additional argument (which is not used)

Solving the overdispersion problem (big overdispersion)

In general, the solution is to apply another (more complex) model:

- ZIP, ZTP, ZAP models
- Negative binomial model
- ZINB, ZTNB, ZANB models
- Generalized Poisson distribution
- Mixtures of distributions

Grouped data

Y_{ij} – events for observation j in group i

$Y_i = \sum_j Y_{ij}$ – total number of events in group i

Assuming the independence, we get that from $Y_{ij} \sim Po(\mu_i), \quad j = 1, \dots, n_i$ follows $Y_i \sim Po(n_i \mu_i)$

The same likelihood function is used for both grouped and ungrouped data

Model for individual means has the following form:

$$\ln \mathbf{E}(Y_{ij}) = \ln \mu_i = \boldsymbol{x}_i^T \beta$$

Thus, for the whole ith group

$$\ln \mathbf{E}(Y_i) = \ln(n_i \mu_i) = \ln n_i + \ln \mu_i = \ln n_i + \boldsymbol{x}_i^T \beta$$

- The estimates for β are the same for grouped and ungrouped case
- In case of grouped data, there is an extra term ($\ln n_i$) called **offset**

Rate data as grouped data

Parameter of Poisson distribution, μ_i, is considered as the number of events in a time unit

Let Y_i be the number of events in a time interval t_i

Rate data

Number of events in a time unit (incidence rate) can be obtained by:

$$IR_i = \frac{Y_i}{t_i}$$

The mean of the rate is

$$\mathbf{E}(\frac{Y_i}{t_i}) = \frac{1}{t_i}\mathbf{E}(Y_i) = \frac{\mu_i}{t_i}$$

Thus the model is

$$\ln(\frac{\mu_i}{t_i}) = \mathbf{x}_i^T\boldsymbol{\beta}, \quad \ln(\mu_i) = \ln(t_i) + \mathbf{x}_i^T\boldsymbol{\beta},$$

where $\ln(t_i)$ is offset