Ricard Garcia Isern
2023
Statistical Methods & Stochastic Processes

POISSON REGRESSION

1. INTRODUCTION

- Used when there are **COUNTS** or **RATES.**

*Model allows

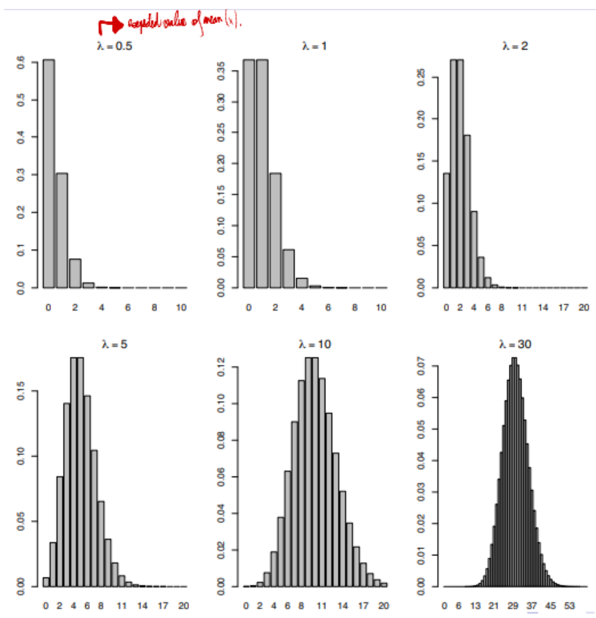
$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad x = 0, 1, 2, \ldots$$

Poisson

$$\underbrace{E(X) = \lambda}_{\text{expected}} \qquad \underbrace{V(X) = \lambda}_{\text{variance}}$$

$$P \rightarrow N(\overset{\text{Mean}}{\lambda}, \underset{\text{Variance}}{\lambda})$$

- **Poisson Densities:**



    - We can observe that the values tend to cluster together to the mean (expected) value.

2. GENERALIZED LINEAR MODELS

- We **know that generalized linear models:**

- Classical linear regression is a particular case of a generalized linear model for normally distributed response variables.
- Logistic regression, studied in the previous module, is also a particular case of a generalized linear model, for binary response variables with a Bernoulli distribution.
- Poisson regression is a statistical method for the modeling of count data, where the response is assumed to follow a Poisson distribution, is another particular case of a generalized linear model.

1. A random component, $Y_i|x_i$, assumed to be distributed according to a member of the exponential family.
2. A linear predictor, with explanatory variables $\mathbf{x}_i$, whose effects are modeled by coefficients $\boldsymbol{\beta}$, given by:

   y

   $$\boxed{\mathbf{x}_i'\boldsymbol{\beta}} = \beta_0 + \beta_1 x_{i1} + \cdots \beta_m x_{im}$$

3. A monotone link function g, such that → if formula is not linear

?

$$\boxed{g(u_i) = \mathbf{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots \beta_m x_{im}} \quad \text{where} \quad \mu_i = E(Y_i|x_i)$$

mean

- **EXPONENTIAL FAMILY:**

A random variable Y with a pdf depending on parameter θ belongs to the exponential family if the pdf can be written as

$$\boxed{f(y|\theta) = s(y)t(\theta)e^{a(y)b(\theta)}}$$

with known function a, b, s and t. Alternatively, this can be written as

$$f(y|\theta) = e^{a(y)b(\theta)+c(\theta)+d(y)}$$

with $s(y) = e^{d(y)}$ and $t(\theta) = e^{c(\theta)}$.

- if $a(y) = y$ the distribution is in canonical form
- $b(\theta)$ is called the natural parameter.
- potentially additional parameters are called nuisance parameters.

- **EXPONENTIAL FAMILY → POISSON DISTRIBUTION**

$$\boxed{\begin{gathered} f(y,\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} \\ f(y,\lambda) = e^{y\ln(\lambda)-\lambda-\ln(y!)} \end{gathered}}$$

- $a(y) = y$ (distribution is in canonical form)
- $b(\lambda) = \ln(\lambda)$ (the natural parameter)
- $c(\lambda) = -\lambda$
- $d(y) = -\ln(y!)$
- The Poisson distribution pertains to the exponential family

3. POISSON REGRESSION → COUNT DATA

Poisson regression for **count** data

ALWAYS if not linear, linearize it.

Poisson regression with a single predictor:

- Y_i is the number of events, with $Y_i|x_i \sim Poisson(\mu_i)$
- $E(Y_i) = \mu_i = e^{\beta_0+\beta_1 x_i} = e^{\beta_0}\left(e^{\beta_1}\right)^{x_i}$

original formula, but not linear

- A one-unit increase in x multiplies the mean of the response by e^{β_1}
- $\ln(\mu_i) = \beta_0 + \beta_1 x_i$
- The link function, $g(\mu_i)$, is usually the natural log, $g(\mu_i) = \ln(\mu_i)$.
- The identity function is sometimes also used as a link function.

Poisson regression with a multiple predictors in vector notation:

- $E(Y_i) = \mu_i = e^{\mathbf{x}_i'\beta}$
- A one-unit increase in x_i multiplies the mean of the response by e^{β_i}, conditional on the other variables.
- $\ln(\mu_i) = \mathbf{x}_i'\beta$

- The Poisson regression model is estimated iteratively by numerical methods.
- Inference on the parameters of the model can be done in several ways. Of common use is the Wald statistic → check if $\beta_i = 0$

coefficient estimate

$$\boxed{Z = \frac{b_j - \beta_j}{s_{b_j}} \sim N(0,1)}$$

theoretical coefficient

- Several kinds of residuals are in use for Poisson regression
  - Let o_i and e_i be observed and expected (fitted) values
  - Pearson residuals

$$\boxed{r_i = \frac{o_i - e_i}{\sqrt{e_i}}}$$

→ with each value

  - Deviance residuals

$$\boxed{d_i = sign(o_i - e_i)\sqrt{o_i \ln(o_i/e_i) - (o_i - e_i)}}$$

- Different models can be compared using likelihood ratio tests, which are typically performed by looking at the deviance.

generalized linear model — linearize data

```
> model <- glm(satellites~width,family=poisson(link="log"))
> summary(model)

Call:
glm(formula = satellites ~ width, family = poisson(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8526  -1.9884  -0.4933   1.0970   4.9221

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.30476    0.54224  -6.095  1.1e-09 ***
width        0.16405    0.01997   8.216  < 2e-16 ***
---
```

β_0 (for the intercept estimate −3.30476); sig (for width)

mul $e^{0.16405} = \beta_1$

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 567.88  on 171  degrees of freedom
AIC: 927.18

Number of Fisher Scoring iterations: 6

>

> anova(model)
Analysis of Deviance Table

Model: poisson, link: log

Response: satellites

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev
NULL                    172     632.79
width  1   64.913       171     567.88
```

Goodness-of-fit

Several criteria can be used to assess the goodness-of-fit of a Poisson regression model

- The chi-square statistic $X^2 = \sum_{i=1}^{n} r_i^2$.
- The deviance $D = \sum_{i=1}^{n} d_i^2 = 2(\ln(L_{sat}) - \ln(L_{fit}))$.
- The pseudo R^2 statistic $R^2 \equiv 1 - D_{fitted}/D_{null}$
- Akaike's information criterion (AIC)

- Chi-square statistics and Deviance allow comparison of nested models.
- ???? With two nested models M_0 (with fewer parameters) and M_1, an LR test is provided by $G^2 = D_0 - D_1 \sim \chi^2_{(k)}$
- AIC allows the comparison of all models, even if these are not nested models.

Overdispersion

- A common problem in Poisson regression is overdispersion.
- Overdispersion refers to the fact that the variance exceeds the mean.
- Underdispersion can also occur, but is less common. → mean > variance
- Overdispersion can be due to various factors such as
  - data heterogeneity (fluctuating covariates)
  - correlation between observations
  - ...
- There are several ways to deal with overdispersion.
  - modeling overdispersion with $V(Y_i) = \phi E(Y_i)$, where ϕ is the overdispersion parameter (typically $\phi > 1$).
  - ϕ can be estimated as $\hat{\phi} = \frac{X^2}{df}$.
  - This can be done by quasi-poisson regression.
  - using negative binomial regression, which allows for $V(Y_i) > E(Y_i)$
  - ...

Testing for overdispersion

- It is possible to formally test for overdispersion (or underdispersion) by a hypothesis test on ϕ
- Typically by testing $H_o : \phi = 1$ against $H_1 : \phi > 1$

→ if multiply = no difference

```
library(AER)
model <- glm(satellites~width,family=poisson(link="log"))
dispersiontest(model)

Overdispersion test

data:  model
z = 5.558, p-value = 1.364e-08
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  3.157244
```

Accounting for overdispersion

```
Call:
glm(formula = satellites ~ width, family = quasipoisson(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8526  -1.9884  -0.4933   1.0970   4.9221

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.30476    0.96729  -3.417 0.000793 ***
width        0.16405    0.03562   4.606 7.99e-06 ***
---

(Dispersion parameter for quasipoisson family taken to be 3.182205)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 567.88  on 171  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

we reduce error → improved model

4. POISSON REGRESSION → RATE DATA

- Typically counts are registered over units of time or space (e.g. # births per village, # goals per match, etc.)
- If the unit of time or space is the same for all observations (e.g. all observations are per day or per square meter) then Poisson regression of count data applies.
- If the observations are made for units of varying size, then it is natural to calculate rates, obtained by dividing counts by the time lapse or population size (n_i).
- Y_i number of events with $Y_i|x_i \sim Poisson(\mu_i)$
- $\ln(\mu_i/n_i) = \beta_0 + \beta_1 x_i$
- $\ln(\mu_i) = \ln(n_i) + \beta_0 + \beta_1 x_i$
- $\ln(n_i)$ is called the offset. This is a fixed term without parameter.
- $E(Y_i) = \mu_i = n_i e^{\beta_0 + \beta_1 x_i}$

$e^{\ln(n_i) + \beta_0 + \beta_1 x} = n_i e^{\beta_0 + \beta_1 x}$

different units

```
Call:
glm(formula = creditcards ~ income + offset(log(cases)), family = poisson(link = "log"))

Deviance Residuals:
    Min       1Q    Median      3Q      Max
-1.6907  -0.9329  -0.5675  0.2186   2.1681

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.386586β0 0.399655  -5.972 2.35e-09 ***
income       0.020758β1 0.005165   4.019 5.84e-05 ***
---

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 42.078  on 30  degrees of freedom
Residual deviance: 28.465  on 29  degrees of freedom
AIC: 67.604

Number of Fisher Scoring iterations: 5
```

Allowing for overdispersion

```
Call:
glm(formula = creditcards ~ income + offset(log(cases)), family = quasipoisson(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6907  -0.9320  -0.5675   0.2186   2.1681

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.386586   0.387408  -6.160 1.03e-06 ***
income       0.020758   0.005006   4.146 0.000269 ***
---

(Dispersion parameter for quasipoisson family taken to be 0.9396513)

    Null deviance: 42.078  on 30  degrees of freedom
Residual deviance: 28.465  on 29  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5

>
```

quasi when is very dif from 1