Poisson regression

Jan Graffelman[1]

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

jan.graffelman@upc.edu

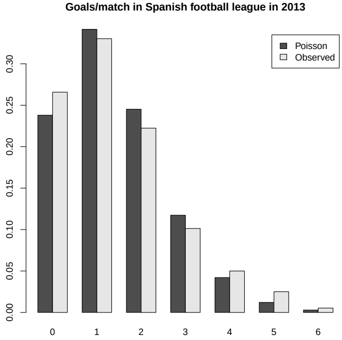
October 17, 2022

Contents

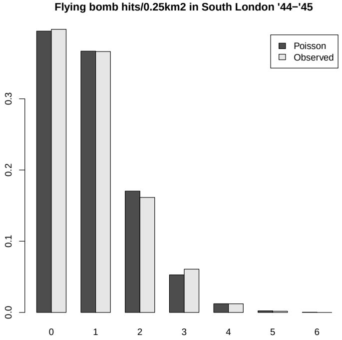
The Poisson distribution (Siméon-Denis Poisson, 1838)

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

$$E(X) = \lambda \qquad V(X) = \lambda$$

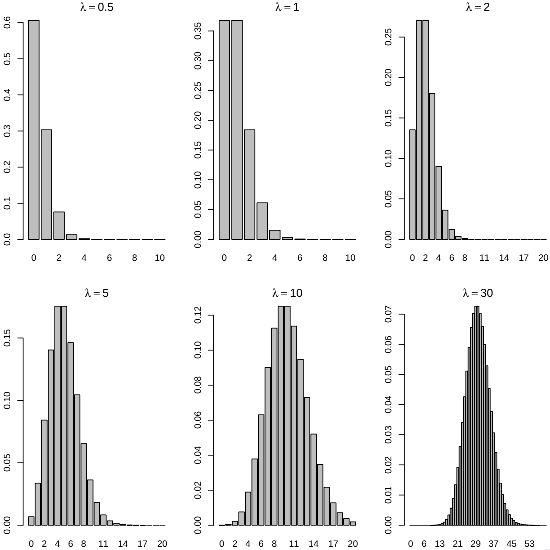


$\overline{x} = 1.436 \qquad s^2 = 1.688$

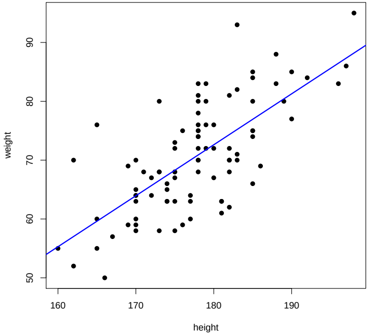


$\overline{x} = 0.929 \qquad s^2 = 0.936$

Some Poisson densities



Classical linear regression



Theoretical model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Usual assumptions:

- $E(\varepsilon_i) = 0$.
- $V(\varepsilon_i) = \sigma^2$ (constant variance).
- $Cov(\varepsilon_i, \varepsilon_j) = 0$ (independent observations).
- $\varepsilon_i \sim N(0, \sigma^2)$.

Summarized:

$$Y_i | X_i \sim N(\beta_0 + \beta_1 X_i, \sigma)$$

Generalized linear models

- In classical regression, $Y_i|X_i$ is required to be normally distributed.
- Regression theory has been generalized in order to deal with responses that are non-normal.
- The corresponding models are known as generalized linear models.
- Generalized linear models form a unifying framework for many statistical methods.
- In particular:
  - 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.
  - ....

Ingredients 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:

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

3. A monotone link function g, such that

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

The 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

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

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.

The exponential family: the Poisson distribution

$$f(y, \lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}$$

$$f(y, \lambda) = e^{y \ln(\lambda) - \lambda - \ln(y!)}$$

- $a(y) = y$ (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

Poisson regression for **count** data

Poisson regression with a single predictor:

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

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

Poisson regression

- 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

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

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

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

  - Deviance residuals

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

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

Goodness-of-fit

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

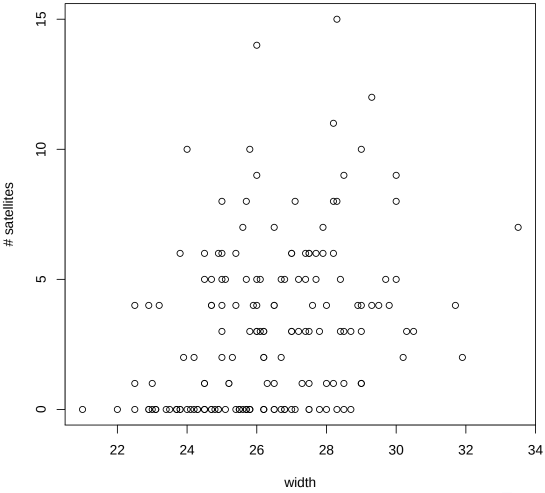
- The chi-square statistic $X^2 = \sum_{i=1}^{n} r_i^2$.
- The deviance $D = \sum_{i=1}^{n} d_i^2 = 2(\ln(L_{sat}) - \ln(L_{fit}))$.
- The pseudo R^2 statistic $R^2 \equiv 1 - D_{fitted}/D_{null}$
- Akaike's information criterion (AIC)

- 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.

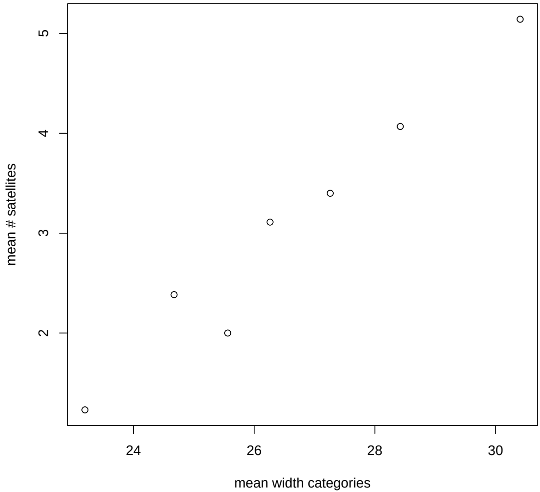
Example: female crab satellites

	color	spine	width	weight	satellites
1	2	3	28.30	3.05	8
2	3	3	26.00	2.60	4
3	3	3	25.60	2.15	0
4	4	2	21.00	1.85	0
5	2	3	29.00	3.00	1
6	1	2	25.00	2.30	3
7	4	3	26.20	1.30	0
8	2	3	24.90	2.10	0
9	2	1	25.70	2.00	8
10	2	3	27.50	3.15	6
11	1	1	26.10	2.80	5
12	3	3	28.90	2.80	4
13	2	1	30.30	3.60	3
14	2	3	22.90	1.60	4
15	3	3	26.20	2.30	3
16	3	3	24.50	2.05	5
17	2	3	30.00	3.05	8
18	2	3	26.20	2.40	3
19	2	3	25.40	2.25	6
20	2	3	25.40	2.25	4
21	4	3	27.50	2.90	0
22	4	3	27.00	2.25	3
23	2	2	24.00	1.70	0
24	2	1	28.70	3.20	0
25	3	3	26.50	1.97	1
26	2	3	24.50	1.60	1
27	3	3	27.30	2.90	1
28	2	3	26.50	2.30	4
29	2	3	25.00	2.10	2
30	3	3	22.00	1.40	0
⋮	⋮	⋮	⋮	⋮	⋮
173	2	2	24.50	2.00	0

Scatterplot



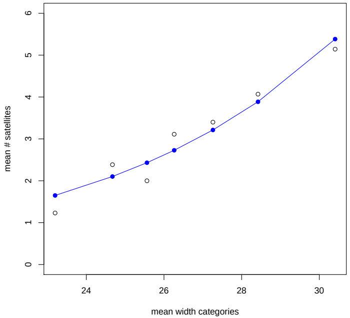
Scatterplot with categorized width



Fitting the model

```
> 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 ***
---


(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
```

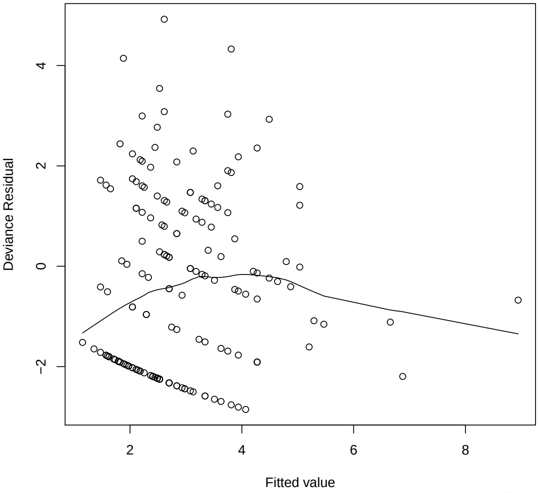
Graphing the fitted model (by interval)



$$\ln(\hat{\mu}) = -3.305 + 0.164 \mathit{width}$$

$$e^{0.16405} = 1.178$$

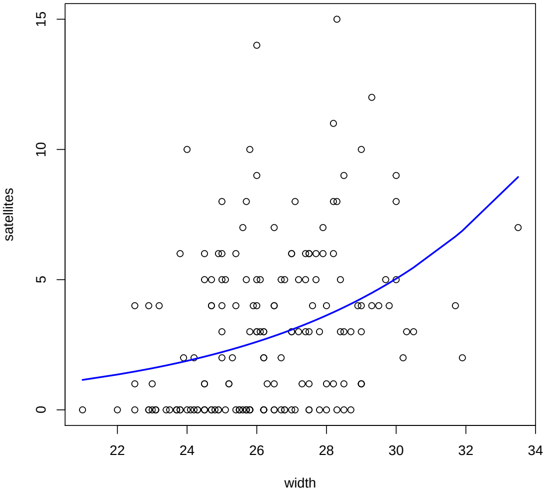
Plotting residuals



Some R instructions

```
model <- glm(satellites~width,family=poisson(link="log"))
summary(model)
yh1 <- predict(model,type="resp")
yh2 <- predict(model)
e  <- resid(model)
ii <- order(width)
plot(width,satellites)
points(width[ii],yh1[ii],type="l",col="blue",lwd="2")

plot(yh2,e,xlab="linear predictor",ylab="residual")
```

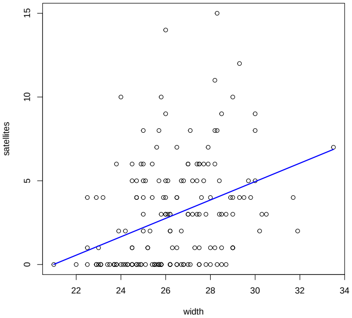
Graphing the fitted model on the original data



Poisson regression with identity link

```
Call:
glm(formula = satellites ~ width,
    family = poisson(link = "identity"),
    start = c(-3, 0.2))

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.52513    0.67877  -16.98   <2e-16 ***
width         0.54923    0.02972   18.48   <2e-16 ***
---

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 557.71  on 171  degrees of freedom
AIC: 917.01

Number of Fisher Scoring iterations: 22
```

$$\hat{\mu} = -11.53 + 0.549 width$$



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.
- 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 $V(Y_i) > E(Y_i)$
  - ...

Female crab satellites

width	$\bar{y}$	s_y^2
(20,24.2]	1.23	6.02
(24.2,25.1]	2.38	6.65
(25.1,25.8]	2.00	9.04
(25.8,26.7]	3.11	10.10
(26.7,27.7]	3.40	6.42
(27.7,29]	4.07	15.00
(29,34.5]	5.14	8.29

$$\hat{\phi} = \frac{567.8786}{171} = 3.320927$$

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 $H_1 : \phi > 1$

```
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

>
```

Adding covariates

```
> model.col <- glm(satellites~width+colf,family=quasipoisson(link="log"))
> summary(model.col)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0415  -1.9581  -0.5575   0.9830   4.7523

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.65004    1.05740  -2.506   0.0132 *
width        0.14934    0.03748   3.985   0.0001 ***
colf2       -0.19969    0.27628  -0.723   0.4708
colf3       -0.43636    0.31713  -1.376   0.1707
colf4       -0.44736    0.37604  -1.190   0.2359
---

(Dispersion parameter for quasipoisson family taken to be 3.233628)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 559.34  on 168  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6

> anova(model.col)
Analysis of Deviance Table

Model: quasipoisson, 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
colf   3    8.534       168     559.34
> qchisq(0.95,df=3)
[1] 7.814728
```

Poisson regression for **rate** data

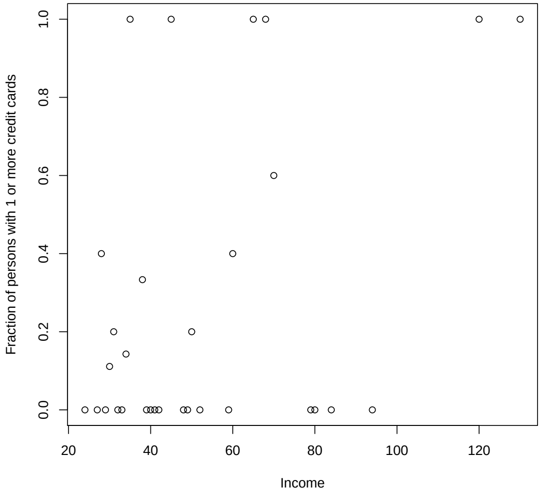
- Typically counts are registered over units of time or space (e.g. # births per village, # goals per match, etc.)
- If the unit of time or space is the same for all observations (e.g. all observations are per day or per square meter) then Poisson regression of count data applies.
- If the observations 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.
- $E(Y_i) = \mu_i = n_i e^{\beta_0 + \beta_1 x_i}$

Credit card data

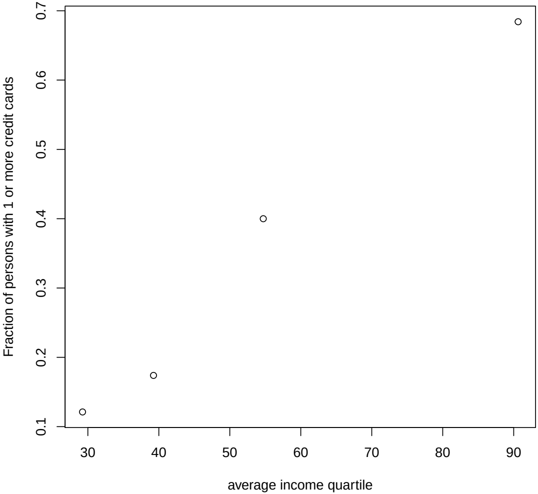
- Random sample of 100 Italians
- Annual income (millions of Lira)
- Number of individuals in each income class
- Number of individuals with one or more credit cards
- Do people with a high income have more credit cards?

	income	cases	creditcards
1	24	1	0
2	27	1	0
3	28	5	2
4	29	3	0
5	30	9	1
6	31	5	1
7	32	8	0
8	33	1	0
9	34	7	1
10	35	1	1
11	38	3	1
12	39	2	0
13	40	5	0
14	41	2	0
15	42	2	0
16	45	1	1
17	48	1	0
18	49	1	0
19	50	10	2
20	52	1	0
21	59	1	0
22	60	5	2
23	65	6	6
24	68	3	3
25	70	5	3
26	79	1	0
27	80	1	0
28	84	1	0
29	94	1	0
30	120	6	6
31	130	1	1

Scatterplot credit card data



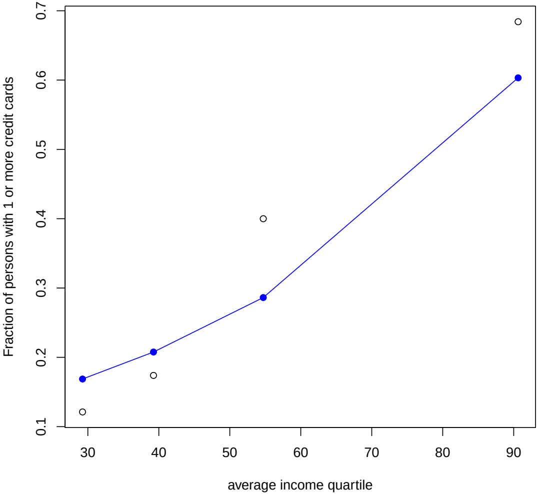
Grouped credit card data



Fitting the Poisson regression

```
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) -2.386586   0.399655  -5.972 2.35e-09 ***
income       0.020758   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
```

Grouped credit card data



Allowing for overdispersion

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

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

>
```

Extensions

Count data may present:

- Zero truncation (zero is not a possible outcome)
- Zero inflation (a zero is more likely than expected under a Poisson model)

Zero inflation

- Sometimes count data has a higher probability for a zero than expected under a Poisson distribution
- E.g. daily number of cigarettes smoked when there are non-smokers.
- In these cases a zero-inflated poisson model can be fitted.
- Zero-inflated Poisson (ZIP) distribution:

$$f(x) = \begin{cases} \theta + (1-\theta)e^{-\lambda} & \text{if } x = 0 \\ (1-\theta)\frac{e^{-\lambda}\lambda^x}{x!} & \text{if } x = 1, 2, \ldots \end{cases}$$

$$E(X) = (1-\theta)\lambda \qquad V(X) = (1-\theta)(1+\theta\lambda)\lambda$$

- A ZIP variable can also be used as a response in a linear model.

References

- Agresti, A. (2013) Categorical data analysis. Third edition. John Wiley & Sons, New York.
- Dobson, A.J & Barnett A.G. (2008) An introduction to generalized linear models. Third edition. Chapman & Hall/CRC, Boca Raton, FL.

Exercise (Poisson regression with count data)

We consider a dataset of 316 highschool students. For each student, school (1 or 2), sex (0 = female, 1 = male), a standardized test score for math, standardized test score for language arts and the number of days of absence at school has been registered.

- Load the data into R by installing package `rsq` and using the instruction `data(hschool)`.
- Perform an exploratory data analysis, and try to identify factors that potentially could affect days of absence.
- Show the relationship between `daysabs` and `langarts` in a scatter plot. Can you think of a way to better visualize the relationship?
- Do you think `daysabs` follows a Poisson distribution? Do a chi-square test for goodness of fit. Is it necessary that `daysabs` follows a Poisson distribution in order to apply Poisson regression?
- Investigate the relevance of the different predictors by doing simple Poisson regressions.
- Do a Poisson regression with all predictors. Do you consider all predictors to be relevant? Simplify the model as you deem convenient.
- Quantify the effect of the different predictors on the response, and give a 95% confidence interval for their true effect.
- Is there evidence for overdispersion? Calculate the mean and variance of `daysabs` for each decile of `langarts`.
- Which factors do, in your opinion, affect the `daysabs`?

Exercise (Poisson regression with rate data)

We consider a dataset on melanoma. This data set from Koch et al. (1986) contains the number of new melanoma cases in 1969-1971 among white males in two areas for various age groups. The size of the estimated population at risk is given in the variable Population

- Load the data file `koch.dat` into R.
- Investigate the relevance of `Area` and `Agegroup` by doing simple Poisson regressions.
- Do a Poisson regression with both predictors. Do you consider both predictors to be relevant? Simplify the model as you deem convenient.
- Quantify the effect of the different predictors on the response, and give a 95% confidence interval for their true effect.
- Is there evidence for overdispersion? Re-fit a model accounting for overdispersion if needed.
- Is there evidence for interaction between `Area` and `Agegroup`? Fit the corresponding model(s) to address this issue.