Introduction to the Generalized Linear Model: Logistic regression and Poisson regression

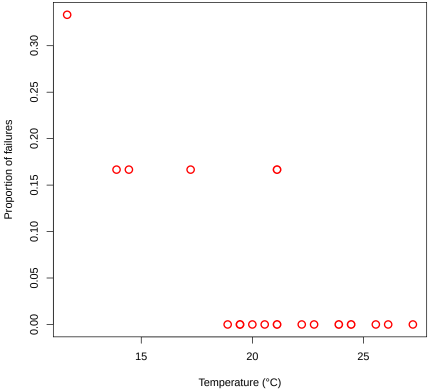
Statistical modelling: Theory and practice

Gilles Guillot

gigu@dtu.dk

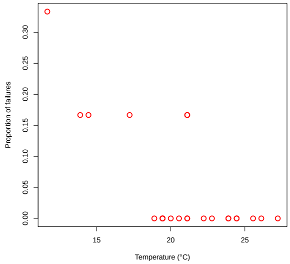
November 4, 2013

US space shuttle Challenger accident data



Proportion of O-rings experiencing thermal distress on the US space shuttle Challenger. Each point corresponds to a flight. Proportion obtained among 6 O-rings for each flight.

US space shuttle Challenger accident data (cont')



- Key question: what is the probability to experience a failure if the temperature is close to 0?
- For the previous example, we would need a model along the line of
  Number of failures $\sim$ temperature
  or
  Proportion of failures $\sim$ temperature

Targeted online advertising

The problem

- A person makes a search on Google on say *'cheap flight Yucatan'*
- Is it worthwhile to display an ad for say *'hotel del Mar Cancún* ?

Targeted online advertising (cont')

- Google knows the preferences of this person from before
  - Google knows on which ad this person as clicked or not clicked in the past
  - In stat. words: Google has vector of binary explanatory variables $(x_1, ..., x_p)$ that summarizes the prospect's preferences
  - Google knows the preferences of other persons from before who have been shown the ad for *'hotel del Mar Cancún* and the same ads as the prospect.
  - Denoting y_i ($\in \{0, 1\}$) the past
  - Google knows y_i and $(x_{i1}, ..., x_{ip})$ for these persons

Targeted online advertising (cont')

- In statistics word:
  - We have a dataset $y_i, x_{i1}, ..., x_{ip}$ with $i = 1, ..., n$
  - We have a prospect known through its vector $(x_1, ..., x_p)$
  - We want to predict the behavior y (0 or 1) of this person regarding the ad *'hotel del Mar Cancún*
- We need a model along the line of $y \sim x_1, ..., x_p$
- The novelty (compared to the linear model) is that y is a count
- What follows applies
  - to any type (quantitative or categorical) of explanatory variables
  - to any number ($p = 1$, $p > 1$) of explanatory variables

Hints about what is coming below

In a linear regression we write

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ and } \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

which we can rephrased as

$$y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

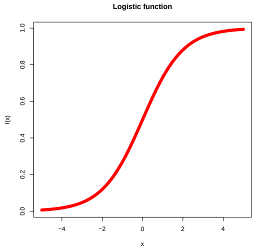
Idea of a logistic regression: twinkle the equation above to fit data that can not be considered Normal.

The logistic function

Definition of the logistic function

The logistic function is a function $l : \mathbb{R} \longrightarrow \mathbb{R}$ defined as

$$l(x) = \frac{\exp(x)}{1+\exp(x)} = \frac{1}{1+\exp(-x)}$$



This function l is a bijection (one-to-one), with inverse $g(x) = l^{-1}(x) = \ln(y/(1-y))$

Logistic regression for a 0/1 response variable

- We consider here the case of a single explanatory variable x and a binary response y.
- The data consist of
  - $y = (y_1, ..., y_n)$ observations of a binary (0/1) or two-way categorical response variable
  - $x = (x_1, ..., x_n)$ observations of a quantitative explanatory variable

Logistic regression for a 0/1 response variable (cont')

- We want a model such that $P(Y_i = 1)$ varies with x_i
- We assume that $Y_i \sim b(p_i)$
- We assume that $p_i = l(\beta_0 + \beta_1 x_i)$ (l logistic function)
- This is an instance of a so-called Generalized Linear Model

Formal definition of the logistic regression for a binary response

- $p_i = l(\beta_0 + \beta_1 x_i)$ for a vector of unknown deterministic parameters $\beta = (\beta_0, \beta_1)$
- $Y_i \sim b(p_i) = b(l(\beta_0 + \beta_1 x_i))$

- Note the analogy with the linear regression $y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$
- Under this model: $E[Y_i] = p_i = 1/(1 + e^{-(\beta_0+\beta_1 x_1)})$
- $V[Y_i] = p_i(1-p_i)$. The variance of Y_i around its mean is entirely controlled by p_i, hence by β_0 and β_1.
- The model enjoys the structure: "Response = signal + noise" but we never write it explicitly nor do we make any assumption or even refer to a residual ε. The structure "Response = signal + noise" is implicit.
- Straightforward extension for more than one explanatory variables: $p_i = l(\sum_{j=1}^{p} x_i^{(j)})$

Logistic regression for finite count data

We consider data $y = (y_1, ..., y_n)$ obtained as counts over several replicates denoted $N = (N_1, ..., N_n)$. [For example, each N_i is equal to 6 in the 0-rings data.]

- $p_i = l(\beta_0 + \beta_1 x_i)$ for a vector of unknown deterministic parameters $\beta = (\beta_0, \beta_1)$
- $Y_i \sim \text{Binom}(N_i, p_i)$

Under this model:
$E[Y_i] = N_i p_i = N_i/(1 + e^{-(\beta_0+\beta_1 x_1)})$
$V[Y_i] = N_i p_i(1 - p_i)$

Maximum likelihood inference for logistic regression

Data: $y = (y_1, ..., y_n)$
$x = (x_1, ..., x_n)$ and $N = (N_1, ..., N_n)$
Parameters: $\beta = (\beta_0, \beta_1)$. NB: there is no parameter controlling the dispersion of residuals
Likelihood:

$$
\begin{aligned}
L(y_1, ..., y_n; \beta_0, \beta_1) &= \prod_{i=1}^{n} \binom{N_i}{y_i} p_i^{y_i} (1 - p_i)^{N_i - y_i} \\
&\propto \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{N_i - y_i} \qquad (1)
\end{aligned}
$$

$$
\ln L(y_1, ..., y_n; \beta_0, \beta_1) = \sum_{i=1}^{n} y_i \ln p_i + (N_i - p_i) \ln(1 - p_i) \qquad (2)
$$

ML inference cont'

Maximizing the log-likelihood

In constrast with the linear model, the maximum log-likelihood of the logistic regression does not have a closed-from solution.
It has to be maximised numerically.

Model fitting in R

Binary response

With x, y two vectors of equal lengths, y being binary:

```
glm(formula = (y~x),
    family=binomial(logit))
```

Returns parameters estimates and loads of diagnostic tools (e.g. AIC).

Example of output of glm function

```
Call:
glm(formula = (y ~ x), family = binomial(link = logit))
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1886  -1.0344   0.5489   0.9069   2.1294
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.54820    0.07364   7.444 9.75e-14 ***
x            1.08959    0.08822  12.351  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 1342.7  on 999  degrees of freedom
Residual deviance: 1138.1  on 998  degrees of freedom
AIC: 1142.1
Number of Fisher Scoring iterations: 4
```

General binomial reponse

With `x`, `N`, `y` three vectors of equal lengths

```
glm(formula = cbind(y,N-y)~x,
   family=binomial(logit))
```

Poisson regression what for?

- In many statistical studies, one tries to relate a count to some environmental or sociological variables.
- For example:
  - Number of cardio-vascular accidents among people over 60 in a US state $\sim$ Average income in the state?
  - Number of bicycles in a Danish household $\sim$ Distance to the city centre
- When the response is a count which does not have any natural upper bound, the logistic regression is not appropriate.
- Even if there is a natural upper bound, the Binomial disitrbution may give a poor fit

The Poisson regression is a natural alternative to the logistic regression in these cases.

Poisson process, Poisson equation, Poisson kernel, Poisson distribution, Poisson bracket, Poisson algebra, Poisson regression, Poisson summation formula, Poisson's spot, Poisson's ratio, Poisson zeros....

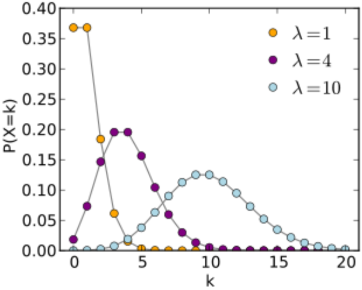
Simeon Poisson 1781-1840

The Poisson distribution

Definition: Poisson distribution

The Poisson distribution is a distribution on $\mathbb{N}$ with probability mass function: $P(X = k) = \dfrac{e^{-\lambda}\lambda^k}{k!}$ for $k \in \mathbb{N}$, and some parameter $\lambda \in (0, +\infty)$

Notation: $X \sim \mathcal{P}(\lambda)$



The Poisson distribution (cont)

If $X \sim \mathcal{P}(\lambda)$

- $E[X] = \lambda$
- $Var[X] = \lambda$

If $X \sim \mathcal{P}(\lambda)$ and $Y \sim \mathcal{P}(\mu)$, X, Y independent

- $X + Y \sim \mathcal{P}(\lambda + \mu)$

Set up for the Poisson regression

- $y_1, ..., y_n$ a discrete, positive variable, e.g. number of bicycles in n households
- $x_1, ..., x_n$ a quantitative variable, e.g. distance of household to city centre or household income

Poisson regression, definition

A Poisson regression is a model relating an explanatory variable x to a positive count variable y with the following assumptions:

Definition: Poisson regression

- $y_1, ..., y_n$ are independant realizations of Poisson random variables $Y_1, ..., Y_n$ with $E[Y_i] = \mu_i$
- $\ln \mu_i = \alpha x_i + \beta$

For short $Y_i \sim \mathcal{P}(\exp(\alpha x_i + \beta))$

- $\alpha x_i + \beta$ is called the linear predictor
- the function that relates the linear predictor to the mean of Y_i is called the link function

Poisson distribution fitting in R

```
glm.res = glm(formula = (nb.cycles ~ dist),
              family = poisson(link="log"),
              data = dat)
```

Output of glm() function

```
summary(glm.res)
Call:
glm(formula = (nb.cycles ~ dist), family = poisson(link = "log"),
    data = dat)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9971  -0.7196  -0.2033   0.4084   3.3813
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.09426    0.24599  -4.448 8.65e-06 ***
dist         0.63211    0.06557   9.640  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 198.77  on 99  degrees of freedom
Residual deviance: 101.55  on 98  degrees of freedom
AIC: 356.91
```

Reading

A Modern Approach to Regression with R, Chapter 8 Logistic Regression, Sheather, Springer 2009.