Logistic regression and Poisson regression

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

October 6, 2017

Outline

- Logistic regression
- Poisson regression

Binary and count data

Linear mixed models very flexible and useful model for continuous response variables that can be well approximated by a normal distribution.

If the response variable is binary a normal distribution is clearly inappropriate.

For count response variables normal distribution may be OK approximation if counts are not too small. However this not so for small counts.

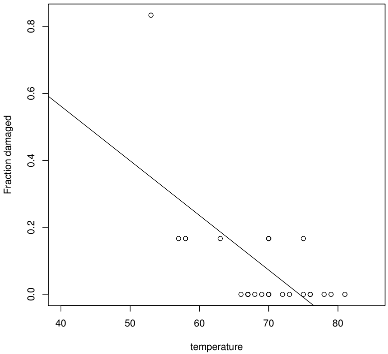
Also problem with variance heterogeneity: typically larger variances for larger counts.

This lecture: regression models for binary and count data.

Example: o-ring failure data

Number of damaged O-rings (out of 6) and temperature was recorded for 23 missions previous to Challenger space shuttle disaster.

Proportions of damaged O-rings versus temperature and least squares fit:

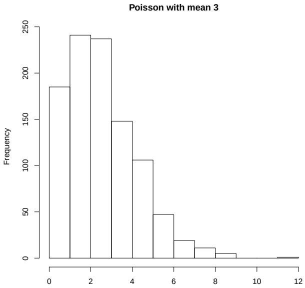


Problems with least squares fit:

- predicts proportions outside $[0, 1]$.
- assumes variance homogeneity (same precision for all observations).
- proportions not normally distributed.

Modeling of o-ring data

Number of damaged o-rings is a count variable but restricted to be between 0 and 6 for each mission. Hence Poisson distribution not applicable (a Poisson distributed variable can take any value $0, 1, 2, \ldots$).



To jth ring for ith mission we may associate binary variable I_{ij} which is one if ring defect and zero otherwise.

We assume the I_{ij} independent with $p_i = P(I_{ij} = 1)$ depending on temperature.

Then $Y_i = \sum_{j=1}^{6} I_{ij}$ follows a binomial $b(6, p_i)$ distribution.

Binomial model for o-ring data

Y_i number of failures and t_i temperature for ith mission.

$Y_i \sim b(6, p_i)$ where p_i probability of failure for ith mission.

Model for variance heterogeneity:

$$\mathbb{V}\mathrm{ar}\, Y_i = n_i p_i (1 - p_i)$$

How do we model dependence of p_i on t_i ?

Linear model:

$$p_i = \alpha + \beta t_i$$

Problem: p_i not restricted to $[0, 1]$!

Logistic regression

Consider logit transformation:

$$\eta = \text{logit}(p) = \log(\frac{p}{1-p})$$

where

$$\frac{p}{1-p}$$

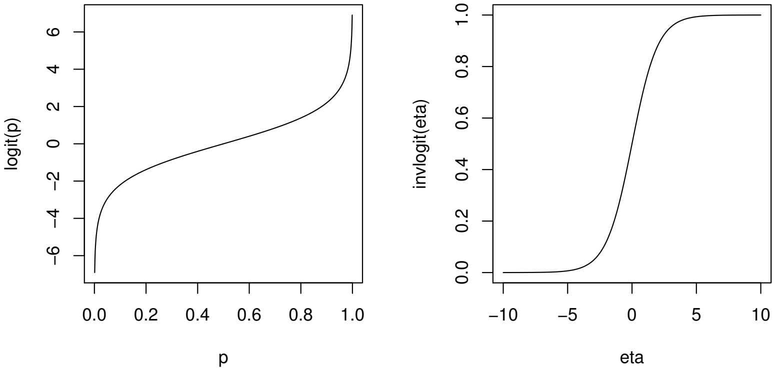
is the *odds* of an event happening with probality p.

Note: logit injective function from $[0, 1]$ to $\mathbb{R}$. Hence we may apply linear model to η and transform back:

$$\eta = \alpha + \beta t \Leftrightarrow p = \frac{\exp(\alpha + \beta t)}{\exp(\alpha + \beta t) + 1}$$

Note: p now guaranteed to be in $[0, 1]$

Plots of logit and inverse logit functions



Logistic regression and odds

Odds for a failure in ith mission is

$$o_i = \frac{p_i}{1 - p_i} = \exp(\eta_i)$$

and odds ratio is

$$\frac{o_i}{o_j} = \exp(\eta_i - \eta_j) = \exp(\beta(t_i - t_j))$$

Example: to double odds we need

$$2 = \exp(\beta(t_i - t_j)) \Leftrightarrow t_i - t_j = \log(2)/\beta$$

Example: $\exp(\beta)$ is increase in odds ratio due to unit increase in t.

Estimation

Likelihood function for simple logistic regression $\text{logit}(p_i) = \alpha + \beta x_i$:

$$L(\alpha, \beta) = \prod_i p_i^{y_i}(1 - p_i)^{n_i - y_i}$$

where

$$p_i = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}$$

MLE $(\hat{\alpha}, \hat{\beta})$ found by iterative maximization (Newton-Raphson)

More generally we may have multiple explanatory variables:

$$\text{logit}(p_i) = \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$

Logistic regression in R

```
> out=glm(cbind(damage,6-damage)~temp,family=binomial(logit))
> summary(out)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 11.66299    3.29626   3.538 0.000403 ***
temp        -0.21623    0.05318  -4.066 4.78e-05 ***
...
  Null deviance: 38.898  on 22  degrees of freedom
Residual deviance: 16.912  on 21  degrees of freedom
...
```

Residual deviance: see later slide.

Note response is a matrix with first rows numbers of damaged and second row number of undamaged rings.

If we had the separate binary variables I_{ij} in a vector `y`, say, this could be used as response instead: `y~temp`.

Hypothesis testing

Wald test:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	11.66299	3.29626	3.538	0.000403	***
temp	-0.21623	0.05318	-4.066	4.78e-05	***

Temperature highly significant.

Same conclusion using likelihood ratio test:

```
> out2=glm(cbind(damage,6-damage)~1,family=binomial(logit)
> anova(out2,out,test="Chisq")
Analysis of Deviance Table

Model 1: cbind(damage, 6 - damage) ~ 1
Model 2: cbind(damage, 6 - damage) ~ temp
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        22     38.898
2        21     16.912  1   21.985 2.747e-06
```

(log likelihood ratio approximately χ^2 distributed)

(alternatively you may use `drop1(out,test="Chisq")`)

Another example: radioactive decay

Intensity of radioactive decay: $\lambda(t) = A \exp(at)$

By theory of physics number of decays X_i in time interval $[t_i, t_{i+1}[$ is a Poisson variable with mean

$$\int_{t_i}^{t_{i+1}} \lambda(t)\mathrm{d}t \approx \Delta_i \lambda(t_i) = \exp(\log \Delta_i + logA + at_i)$$

where $\Delta_i = t_{i+1} - t_i$.

NB: X_i for disjoint intervals independent.

Simulated radioactive decay $x_0, \ldots, x_{14}$ within unit intervals $[t, t+1[$, $t = 0, 1, 2, \ldots$:

5 9 5 5 2 1 4 0 0 2 0 0 0 0 1

Naive approach:

$$\log \mathbb{E} X_i \approx \log 1 + \log A + at_i = \log A + at_i, \quad i = 0, 1, 2,$$

hence fit linear regression to $(t_i, \log x_i)$.

Problems:

- log transformation of zero counts ?
- variance heterogeneity: larger counts have large variance
- linear model fits model for $\mathbb{E} \log X_i$ but this is different from $\log \mathbb{E} X_i$

Better approach: Poisson regression with log link.

Poisson regression

Suppose $X_1, \ldots, X_n$ are Poisson distributed with associated covariates $z_1, \ldots, z_n$.

Let $\lambda_i > 0$ denote expectation of X_i. We might try linear model

$$\lambda_i = \alpha + \beta z_i$$

but this may conflict with the requirement $\lambda_i > 0$.

Better alternative is log-linear model

$$\lambda_i = \exp(\alpha + \beta z_i)$$

since this guarantees $\lambda_i > 0$.

Variance heterogeneity: for a Poisson variable, the variance is equal to the expectation:

$$\mathbb{V}\mathrm{ar} X_i = \mathbb{E} X_i = \lambda_i.$$

Implementation in R - linear model

```
> radiols=lm(log(x+0.001)~offset(log(deltat))+times)
> summary(radiols)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1969     1.5489   1.418  0.17961
times        -0.6152     0.1883  -3.267  0.00612 **
```

True $\log A$ and a are 2.08 and -0.3.

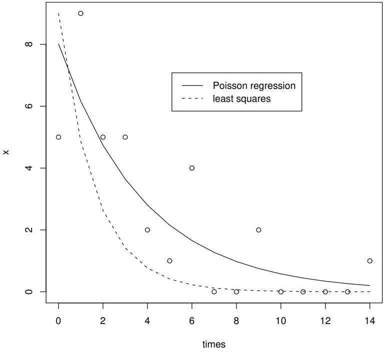
Implementation in R - Poisson regression model

```
> radiofit=glm(x~offset(log(deltat))+times,family=poisson(l
> summary(radiofit) #offset to take into account lengths of
...                  #which may in general differ from 1
    Min        1Q    Median       3Q       Max
-1.5955   -1.0093   -0.7251   0.8709   1.5391
...
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.08130    0.23835   8.732  < 2e-16 ***
times       -0.26287    0.05464  -4.811  1.5e-06 ***
...
Residual deviance: 17.092  on 13  degrees of freedom
```

True $\log A$ and a are 2.08 and -0.3.

Data and fitted values

```
plot(times,x)
lines(times,fitted(radiofit))
lines(times,exp(fitted(radiols)),lty=2)
legend(locator(1),lty=c(1,2),legend=c("Poisson regression","leas
```



Note problems with least squares fit: follows zeros too closely !

Model assessment for logistic and Poisson regression

- Pearson's statistic

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

  where $V(\mu)$ is variance of observation with mean μ ($\mu = p$ or $\mu = \lambda$, $V(p) = np(1-p)$ or $V(\lambda) = \lambda$).
- Plot Pearson residuals against predicted values and covariates

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

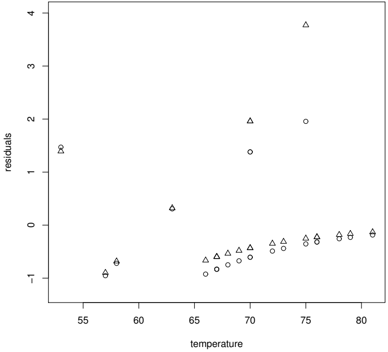
NB: Pearson's statistic approximately $\chi^2(n-p)$ where p number of parameters - if μ_i's not too small (larger than 5 say).

NB: Pearson residuals not normal - can make interpretation difficult.

Deviance closely related to Pearson's statistic but more technical. Deviance residuals similar to Pearson residuals.

Residuals for o-rings

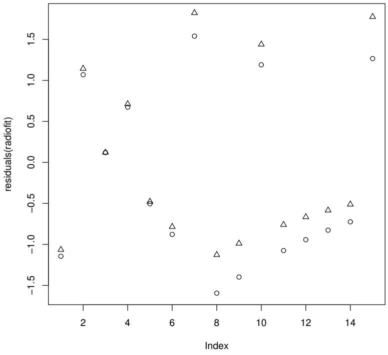
```
devres=residuals(out)
plot(devres~temp,xlab="temperature",ylab="residuals",ylim=c(-1.2
pearson=residuals(out,type="pearson")
points(pearson~temp,pch=2)
```



Much spurious structure due to discreteness of data.

Residuals for radioactive decay

```
plot(residuals(radiofit),ylim=c(-1.6,1.8))
points(residuals(radiofit,type="pearson"),pch=2)
```



Much spurious structure due to discreteness of data.

Generalized linear models

Logistic and Poisson regression special cases of wide class of models called *generalized linear models* that can all be analyzed using the `glm`-procedure.

We need to specify distribution family and link function.

In practice Binomial/logistic and Poisson/log regression are the most commonly used examples of generalized linear models.

SPSS: Analyze $\rightarrow$ Generalized linear models $\rightarrow$ etc.

Overdispersion

Suppose Pearsons X^2 is large relative to degrees of freedom $n - p$.

This may either be due to systematic defiency of model (misspecified mean structure) or *overdispersion*, i.e. variance of observations larger than model predicts.

Overdispersion may be due e.g. to unobserved explanatory variables like e.g. genetic variation between subjects, variation between batches in laboratory experiments, or variation in environment in agricultural trials.

There are various ways to handle overdispersion - we will focus on a model based approach: generalized linear mixed models.

Deviance for logistic regression

Predicted observation for current model:

$$\hat{y}_i = n_i \hat{p}_i \quad \text{logit}\hat{p}_i = \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_p x_{pi}$$

Saturated model: no restrictions on p_i so $\hat{p}_i^{\text{sat}} = y_i / n_i$ and $\hat{y}_i^{\text{sat}} = y_i$ (perfect fit).

Residual deviance D is -2 times the log of the ratio between $L(\hat{\beta}_1, \ldots, \hat{\beta}_p)$ and likelihood L_{sat} for the saturated model.

$$D = 2 \sum_{i=1}^{n} [y_i \log(y_i / \hat{y}_i) + (n_i - y_i) \log((n_i - y_i)/(n_i - \hat{y}_i))]$$

If n_i not too small $D \approx \chi^2(n - p)$ where p is the number of parameters for current model. If this is the case, D may be used for goodness-of-fit assessment.

Null deviance is log ratio between maximum likelihood for model with only intercept and L_{sat}.

Exercises

1. Suppose the probability that the race horse Flash wins is 10%. What are the odds that Flash wins ?
2. Suppose the that the logit of the probability p is 0, $\text{logit}(p) = 0$. What is then the value of p ?
3. Consider a logistic regression model with $P(X = 1) = p$ and $\text{logit}(p) = 3 + 2z$. What are the odds for the event $X = 1$ when $z = 0.5$? What is the increase in odds if z is increased by one ?
4. Show that the mean and variance of a binomial variable $Y \sim b(n, p)$ are np and $np(1 - p)$, respectively.

   Hint: use that $Y = I_1 + I_2 + \ldots, I_n$ where the I_i are independent binary random variables with $P(I_i = 1) = p$.

5. Consider the wheezing data (available as data set `ohio` in the `faraway` package or ohio.sav at the course web page).

   The variables in the data set are resp (an indicator of wheeze status, 1=yes, 0=no), id (a numeric vector for subject id), age (a numeric vector of age, 0 is 9 years old), smoke (an indicator of maternal smoking at the first year of the study).

   Fit a logistic regression model for the binary resp variable with age and smoke as factors. Check the significance of age and smoke. Compare with a model with age as a covariate (i.e. a single slope parameter for age).

6. Consider the epilepsy data (available in the `faraway` package or as faraway.sav). The data are from a clinical trial of 59 epileptics. For a baseline, patients were initially observed for 8 weeks and the number of seizures recorded. The patients were then randomized to treatment by the drug Progabide (31 patients) or to the placebo group (28 patients). They were then observed for additionally four 2-week periods and the number of seizures in each period was recorded.

   The variables in the data are seizures (number of seizures), id (identifying number), treat (1=treated group, 0=placebo group), expind (0=baseline period, 1=treatment period), timeadj (length of observation period in weeks), age in years.

   Fit a Poisson regression to the seizures data in order to investigate the effect of treatment on the number of seizures. Use log(timeadj) as an offset to adjust for the different observation periods (8 or 2 weeks) for the counts. Also investigate the effect of age.