Logistic regression and generalized linear models

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

April 3, 2012

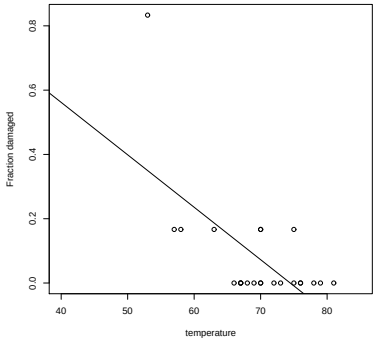
Topics of the day

- Logistic regression
- Generalized linear models
- Poisson regression

O-ring failure data

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

Fractions 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.

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.

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:

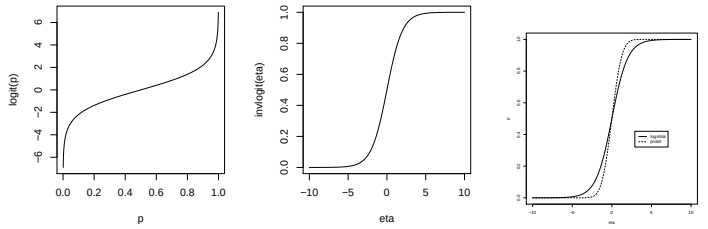
$$\eta = \text{logit}(p) = \log(\frac{p}{1-p})$$

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 guaranteed to be in $[0,1]$

Plots of logit, inverse logit, and probit



Probit transformation: $p_i = \Phi(\eta_i)$ where Φ cumulative distribution function of standard normal variable ($\Phi(u) = P(U \leq u)$.)

Regression parameter for logistic roughly 1.8 times regression parameter for probit since Φ more steep than inverse logit.

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

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

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:

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

Deviance

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}.

Logistic regression in R

```
> out=glm(cbind(damage,6-damage)~temp,family=binomial(logi
> 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 not large compared with numbers of degrees of freedom.

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")`)

Generalized linear models

Suppose Z is random variable with expectation $\mathbb{E}Z = \mu \in M$ where $M \subset \mathbb{R}$. Idea: use invertible link function $g : M \to \mathbb{R}$ and apply linear modelling to $\eta = g(\mu)$.

Binomial data: $Z = Y/n$, $Y \sim b(n, p)$. $\mu = p \in M =]0, 1[$. $g(\cdot)$ e.g. logistic or probit.

Poisson data: $Z \sim \text{pois}(\lambda)$. $\mu = \lambda > 0$. g e.g. log.

Many other possibilities (McCullagh and Nelder, Faraway, Dobson) e.g. gamma distribution and inverse Gaussian for positive continuous data.

For binomial and Poisson, $\mathbb{V}\text{ar}Z = V(\mu)$ determined by μ: $V(\mu) = \mu(1-\mu)/n$ and $V(\mu) = \mu$, respectively.

Radioactive decay

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

Probability of decay within infinitesimally small interval $[t, t + \mathrm{d}t[$ is $\lambda(t)\mathrm{d}t$ and independence between disjoint intervals implies that times of decays $0 < T_1 < T_2 < T_3 < \ldots$ form an inhomogeneous Poisson process with intensity function $\lambda(t)$.

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

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

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:

$$\log \mathbb{E}X_t \approx \log 1 + \log A + at = \log A + at, \quad t = 0, 1, 2,$$

hence fit linear regression to $(t, \log x_t)$.

Problems:

- log transformation of zero counts ?
- variance heterogeneity
- $\mathbb{E}\log X_t < \log \mathbb{E}X_t \Rightarrow \exp(\mathbb{E}\log X_t) < \mathbb{E}X_t$.

Right approach: Poisson regression with log link.

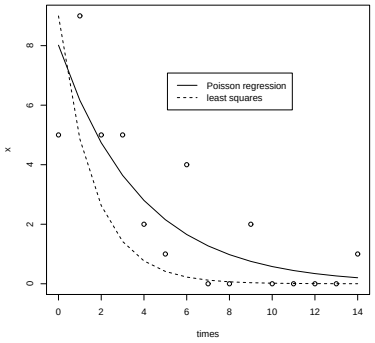
Implementation in R

```
> radiofit=glm(x~offset(log(deltat))+times,family=poisson(log))
> summary(radiofit) #offset to take into account lengths of time
...                 #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
> 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.

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 !

Model assessment for logistic and Poisson regression

- consider residual deviance or Pearsons statistic

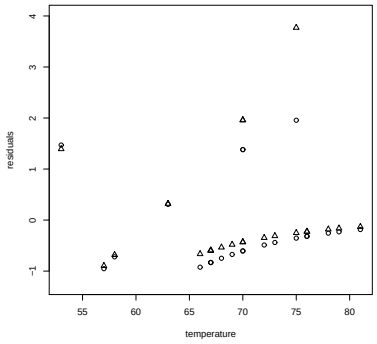
$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

  where $V(\mu)$ is variance of observation with mean μ.
- plot deviance or Pearson residuals against predicted values and covariates
  1. Pearson:
$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$
  2. deviance: residual deviance is $\sum_i (r_i^D)^2$ where r_i^D is contribution from ith observation.

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

Residuals for o-rings

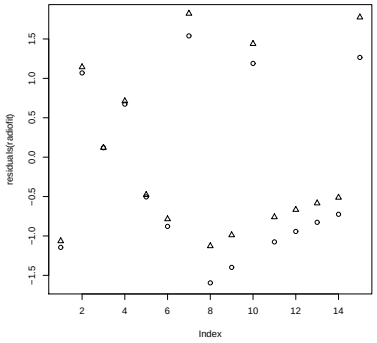
```
devres=residuals(out)
plot(devres~temp,xlab="temperature",ylab="residuals",ylim=c(-1.2
pearson=residuals(out,type="pearson")
points(pearson~temp,pch=2)
```



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.

Overdispersion

Suppose residual deviance or Pearsons X^2 is large relative to degrees of freedom.

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.

Exercises

- Consider the wheezing data (available as data set `ohio` in the `faraway` package). Fit a logistic regression model with age and smoke as factors. Check the significance of the different effects using likelihood ratio test and omit nonsignificant effects. Compare with a model with age as a covariate (i.e. a single slope parameter for age). Compare the fit of this model with the previous model using a likelihood ratio test. Take a look at deviance and Pearson residual plots. Can you use the residual deviance for goodness-of-fit testing ?
- The wheezing data may be aggregated according to the groups given by age and smoke (the aggregated data set is available at the web-page). Repeat the preceeding exercise but now with the aggregated data.

More exercises

- (Exercise from Faraway) The `ships` dataset found in the `MASS` package gives the number of damage incidents and aggregate months of service for different types of ships broken down by year of construction and period of operation. Develop a model for the rate of incidents, describing the effects of important predictors.

Optional exercises:

- show that the mean and variance of a binomial variable $Y \sim b(n, p)$ are np and $np(1-p)$, respectively.
- What is the formula for the residual deviance in case of a Poisson regression ?
- Show that the probit model for binary data may be viewed as a latent variable model where $Y = 1[U < a + bx]$ for a latent standard normal variable U. The latent variable could e.g. correspond to susceptibility to an insecticide if Y represents dead/alive for an insect subjected to an insecticide dose x.