Logistic regression

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

October 12, 2021

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.

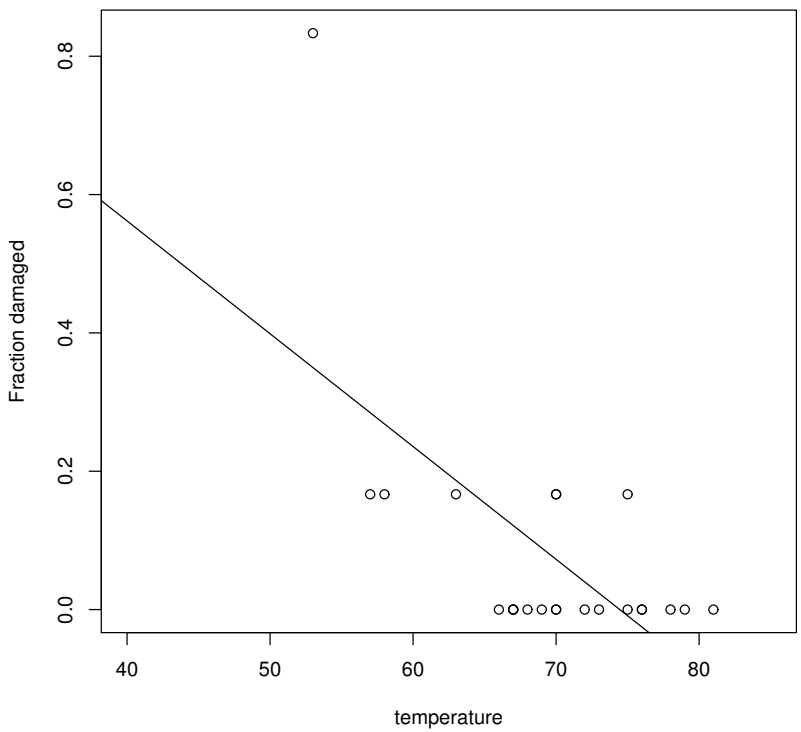
Also often problems with variance heterogeneity.

This lecture: focus on regression models for binary and binomial 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:

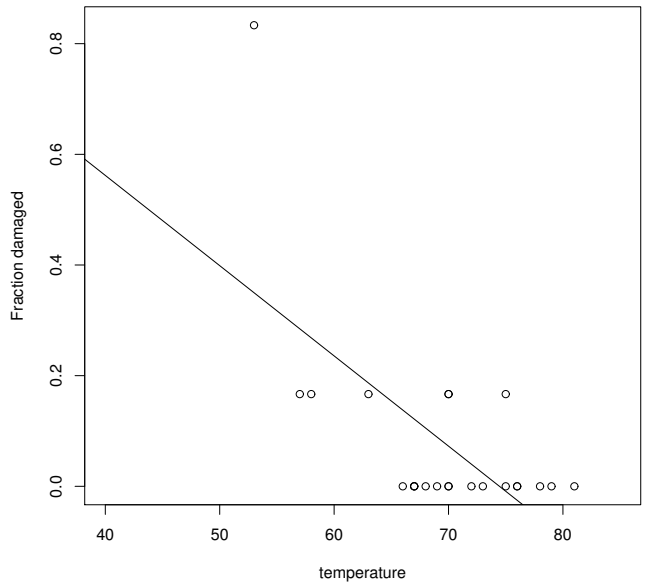


Problems with least squares fit:

- 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 count of defect rings, $Y_i = I_{i1} + I_{i2} + \cdots + I_{i6}$ 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

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

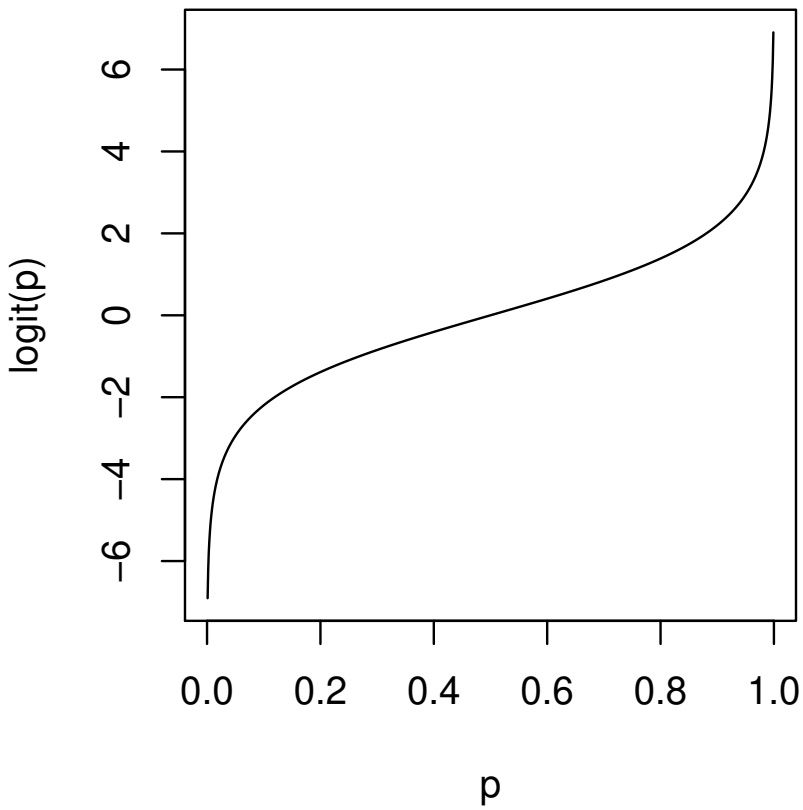
is the *odds* of an event happening with probality p.

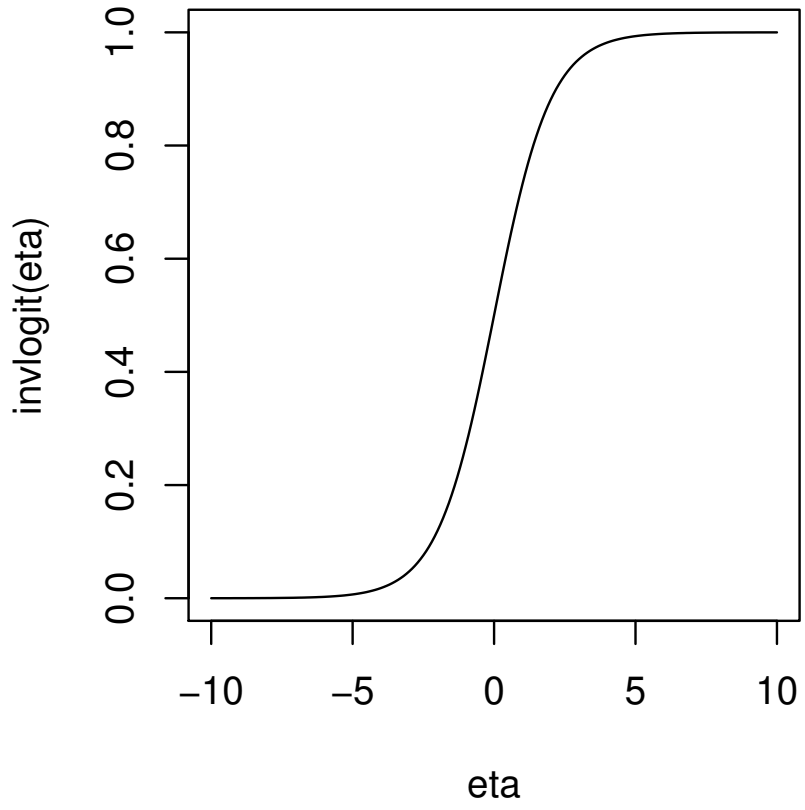
Note: logit injective function from $]0,1[$ to $\mathbb{R}$. Hence we may apply linear model to η and transform back:

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

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

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

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.

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
...
```

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`.

Generalized linear models

Logistic regression special case 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

For a binomial variabl $Y \sim bin(n, p)$, $\mathbb{E}Y = np$ and $\mathbb{V}\mathrm{ar}\,Y = np(1-p)$.

For Poisson (details omitted) $\mathbb{E}Y = \mathbb{V}\mathrm{ar}\,Y$.

Overdispersion happens when actual variance of Y is bigger than predicted by the model. For example $\mathbb{V}\mathrm{ar}\,Y > np(1-p)$.

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.

Overdispersion can also be due to correlation between trials I_{ij} forming a count variable $Y_i = I_{i1} + \ldots + I_{in_i}$.

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

Exercises

1. Suppose the probability that the race horse Flash wins is 10%. What are the odds that Flash wins ?
2. Suppose 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).