Logistic regression and generalized linear models

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

October 28, 2021

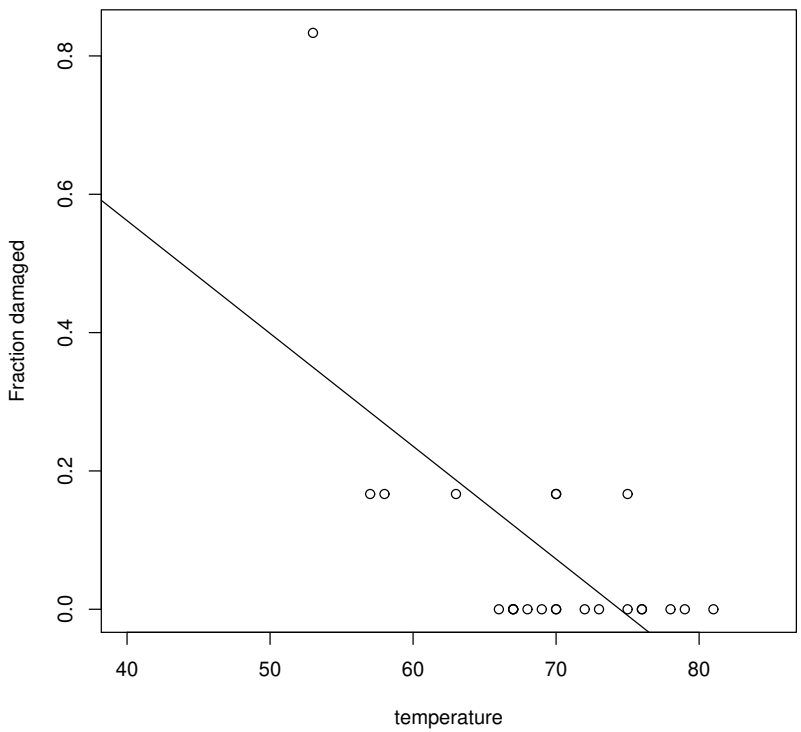
Topics of the day

- Logistic regression
- Overdispersion
- Logistic regression with random effects

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:

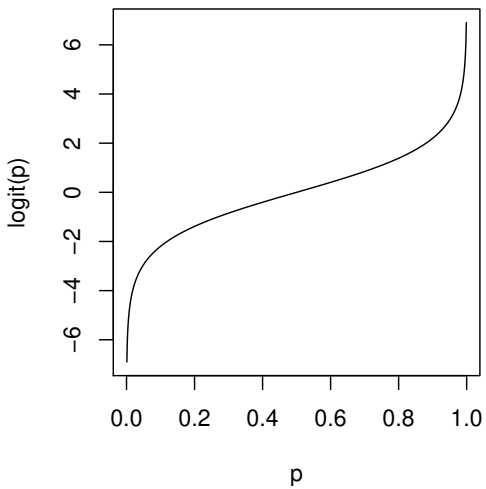
$$\eta = \text{logit}(p) = \log(\frac{p}{1-p})$$

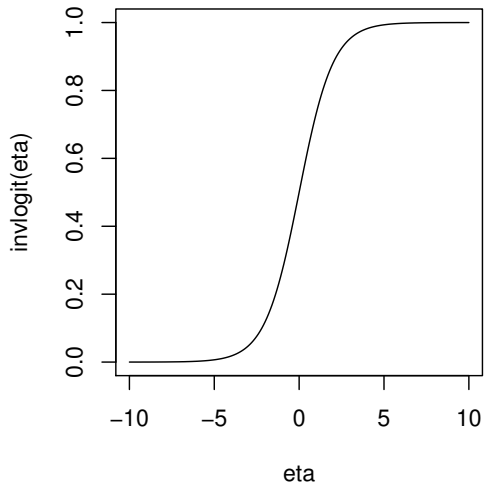
Note: logit injective function from $]0,1[$ to $\mathbb{R}$. Hence we may apply linear model to η and transform back:

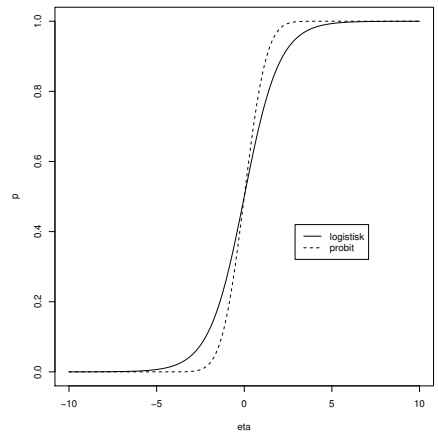
$$\eta = \alpha + \beta t \Leftrightarrow p = \frac{\exp(\alpha + \beta t)}{\exp(\alpha + \beta t) + 1}$$

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.

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

If n_i not too small $D \approx \chi^2(m - p)$ where m number of observations and p 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}.

Pearson's X^2:

$$X^2 = \sum_{i=1}^{m} \frac{(y_i - n_i \hat{p}_i)^2}{n_i \hat{p}_i (1 - \hat{p}_i)}$$

is asymptotically equivalent alternative to D.

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
...
```

$n_i = 6$ so residual deviance approximately $\chi^2(21)$

Residual deviance not large compared with numbers of degrees of freedom.

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.

Overdispersion

In some applications we see larger variability in the data than predicted by variance formulas for binomial.

This is also sometimes revealed by large residual deviance or X^2 relative to degrees of freedom.

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

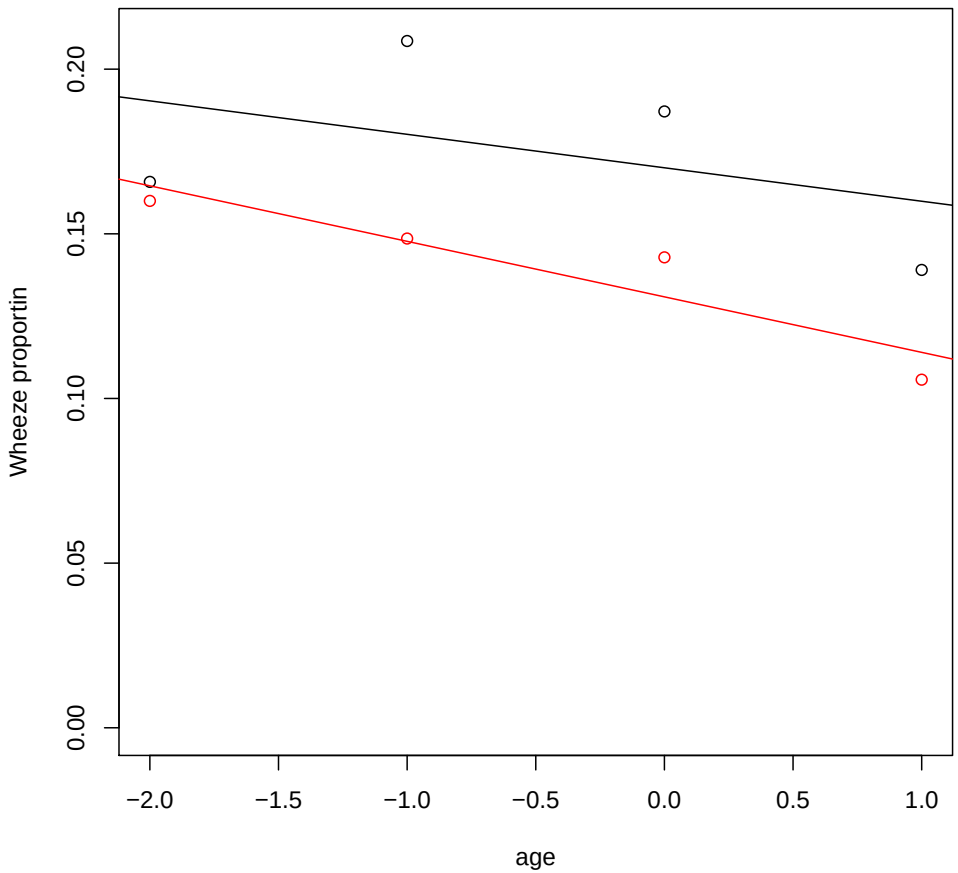
Overdispersion may be caused e.g. by genetic variation between subjects, variation between batches in laboratory experiments, or variation in environment in agricultural trials.

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

Wheezing data

The wheezing (Ohio) data has variables resp (binary indicator of wheezing status), id, age (of child), smoke (binary, mother smoker or not).

Aggregated data: (black=smoke, red=no smoke)



Let Y_{ij} denote wheezing status of ith child at jth age. Assuming Y_{ij} is $b(1, p_{ij})$ we try logistic regression

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{smoke}_{ij}$$

Assuming independence between observations from the same child, and letting $Y_{i\cdot}$ be the sum of observations from ith child,

$$\begin{aligned}
&\mathbb{V}\text{ar}\, Y_{i\cdot} \\
=&\mathbb{V}\text{ar}(Y_{i1} + Y_{i2} + Y_{i3} + Y_{i4}) = \mathbb{V}\text{ar}\, Y_{i1} + \mathbb{V}\text{ar}\, Y_{i2} + \mathbb{V}\text{ar}\, Y_{i3} + \mathbb{V}\text{ar}\, Y_{i4} \\
=&p_{i1}(1 - p_{i1}) + p_{i2}(1 - p_{i2}) + p_{i3}(1 - p_{i3}) + p_{i4}(1 - p_{i4})
\end{aligned}$$

Note: same variance of $Y_{i\cdot}$ for all children with same value of smoke.

We can calculate above theoretical variance from fitted model and compare with empirical variances.

Smoke=0: theoretical: 0.58 empirical: 1.22.

Smoke=1: theoretical: 0.48 empirical: 0.975

Issue: observations from same child are correlated[1] we know first observation is non-wheeze then very likely three remaining observations non-wheeze too.

Correlation can be due to genetics or the environment (more or less polluted) for the child.

Explicit model these effects using random effect:

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{smoke}_{ij} + U_i$$

where U_i are $N(0, \tau^2)$ and independent among children.

Such a model can be fitted by the *R*-procedure `glmer` with syntax very close related to `lmer` and `glm`

[1] $\mathbb{V}\text{ar}(\sum_{i=1}^m Y_i) = \sum_{i=1}^m \mathbb{V}\text{ar}(Y_i) + 2\sum_{i<j} \mathbb{C}\text{ov}(Y_i, Y_j)$

Logistic regression

```
> fit=glm(resp~age+smoke,family=binomial,data=ohio)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.88373    0.08384 -22.467   <2e-16 ***
age         -0.11341    0.05408  -2.097   0.0360 *
smoke        0.27214    0.12347   2.204   0.0275 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

    Null deviance: 1829.1  on 2147  degrees of freedom
Residual deviance: 1819.9  on 2145  degrees of freedom
```

According to above results, age and smoke both significant at the 5% level.

χ^2 distribution of deviance residual not trustworthy here since $n_i = 1$.

We can increase n_i by aggregrating over 8 categories for age $\times$ smoke but then variability between children hidden.

Mixed model analysis

```
> fiter=glmer(resp~age+smoke+(1|id),family=binomial,data=oh
> summary(fiter)
Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 5.491    2.343
Number of obs: 2148, groups:  id, 537

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.37396    0.27496 -12.271   <2e-16 ***
age         -0.17677    0.06797  -2.601   0.0093 **
smoke        0.41478    0.28705   1.445   0.1485
```

Now only age is significant on the 5% level.

Note large variance 5.491 for the U_i.

Interpretation of variance of random effects

Variance 5.491 corresponds to standard deviation 2.343. This means 95% probability interval for U_i is $[-4.686, 4.686]$.

Large part of the variation explained by the U_i relative to the fixed effects.

Interpretation in terms of marginal variance ?

For logistic regression with random effects, the variance of an observation Y_{ij} is[2]

$$\mathbb{V}\mathrm{ar}\, Y_{ij} = \mathbb{E}p_{ij}(1 - p_{ij}) + \mathbb{V}\mathrm{ar}\, p_{ij} \tag{1}$$

where the expectation and variance is with respect to U_i in

$$p_{ij} = \frac{\exp(\alpha + \beta^{\mathsf{T}} z_{ij} + U_i)}{1 + \exp(\alpha + \beta^{\mathsf{T}} z_{ij} + U_i)}$$

There is no simple formula for this variance.

Here $p_{ij}(1 - p_{ij})$ is the conditional variance of Y_{ij} given U_i - but this can not be evaluated since U_i is unobserved.

[2] $\mathbb{V}\mathrm{ar}\, Y = \mathbb{E}\mathbb{V}\mathrm{ar}[Y|X] + \mathbb{V}\mathrm{ar}\mathbb{E}[Y|X]$

Interpretation in terms of odds

The odds are

$$O_{ij} = \frac{p_{ij}}{1 - p_{ij}} = \exp(\alpha + \beta^{\mathsf{T}} z_{ij} + U_i)$$

and the odds ratio between observations ij and kl is

$$\frac{O_{ij}}{O_{kl}} = \exp(\beta^{\mathsf{T}}(z_{ij} - z_{kl}) + U_i - U_k)$$

where $U_i - U_k \sim N(0, 2\tau^2)$.

Larsen et al. (2000) suggested to consider the median odds ratio between the individual with highest random effect and the individual with lowest random effect.

In other words: MOR is

$$\exp[\beta^{\mathsf{T}}(z_{ij} - z_{kl}) + \text{MED}(|U_i - U_k|)] = \exp[\beta^{\mathsf{T}}(z_{ij} - z_{kl})] \exp[\sqrt{2}\tau 0.6744)]$$

(note $|U_i - U_k|$ is $U_i - U_k$ if $U_i > U_k$ and vice versa).

95% intervals for probabilities or odds

U_i is between -1.96τ and 1.96τ with 95% probability.

Hence odds O_{ij} in interval

$$[\exp(\alpha + \beta^{\mathsf{T}} z_{ij} - 1.96\tau); \exp(\alpha + \beta^{\mathsf{T}} z_{ij} + 1.96\tau)]$$

with probability 95%.

For probability p_{ij} the interval is

$$\left[\frac{\exp(\alpha + \beta^{\mathsf{T}} z_{ij} - 1.96\tau)}{1 + \exp(\alpha + \beta^{\mathsf{T}} z_{ij} - 1.96\tau)}; \frac{\exp(\alpha + \beta^{\mathsf{T}} z_{ij} + 1.96\tau)}{1 + \exp(\alpha + \beta^{\mathsf{T}} z_{ij} + 1.96\tau)}\right]$$

Wheezing data

E.g. with $\tau = 2.343$ we get MOR=9.34.

That is, keeping all fixed factors equal, for two randomly picked children, the median odds ratio between the child with highest random effect and the child with lowest random effect is 9.34.

For child of centered age 0 and with smoking mother the 95% interval for wheezing is

$$\left[\frac{\exp(-3.37+0.41-1.96*2.34)}{1+\exp(-3.37+0.41-1.96*2.34)}; \frac{\exp(-3.37+0.41+1.96*2.34)}{1+\exp(-3.37+0.41+1.96*2.34)}\right] = [0.00; 0.84]$$

Mean probability (by Monte Carlo) is 0.16.

Emphasizes the large individual specific effects.

Summary

- logistic regression very useful for binary data
- in some applications there is evidence of overdispersion (extra variance)
- easy to add random effects to model sources of overdispersion and thereby correctly model correlation between observations e.g. for same subject.
- thereby we get more trustworthy standard deviations for fixed effects estimates.
- disadvantage: not easy to interpret random effects variances in terms of variances and correlations of the response variable Y_i.
- likelihood function very complicated

Next time: computation of the likelihood (how does `glmer` work ?)

Exercises

1. (Threshold model) 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.
2. 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). Compare logistic regression analyses for the original and aggregated data.
3. The variance of a (standard) logistic distribution is $\pi^2/3$. Argue why this implies that a logistic regression with parameter β roughly corresponds to a probit regression with parameter $\beta\frac{\sqrt{3}}{\pi} \approx \beta/1.8$

4. An experiment was designed to assess the effect of different stocks on the robustness of cherry flowers to frost. For 20 cherry trees of 5 different stock varieties, three branches were sampled and on each branch the status of 5 buds (dead=1 or alive=0) were recorded. The data are available as `cherries_red.txt`.
   - 4.1 Fit a logistic model with systematic STOCK and BRANCHNR effects and with random BRANCHID and TREEID effects. Is there scope for simplification of the random part of the model ?
   - 4.2 What can you conclude about the STOCK effects ?
   - 4.3 Is there a BRANCHNR effect ? Does this make sense ?
5. Verify the variance expression (1).
6. Suppose that X_1 and X_2 are independent $N(0, \tau^2)$. Show that the median of $|X_1 - X_2|$ is $\sqrt{2}\tau 0.6744$.

   Hint: consider first median of $(X_1 - X_2)^2$ which has a well-known distribution.