Mixed models for binary data

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

October 12, 2021

Variance for binomial distribution

For binomial variables, variance is determined by mean.

Y binomial $b(n, p)$:

$$\mathbb{E}Y = np \quad \mathbb{V}\mathrm{ar}\,Y = np(1-p)$$

Binary case, $n = 1$:

$$\mathbb{E}Y = p \quad \mathbb{V}\mathrm{ar}\,Y = p(1-p)$$

Overdispersion

Binomial default model in case of binary data.

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

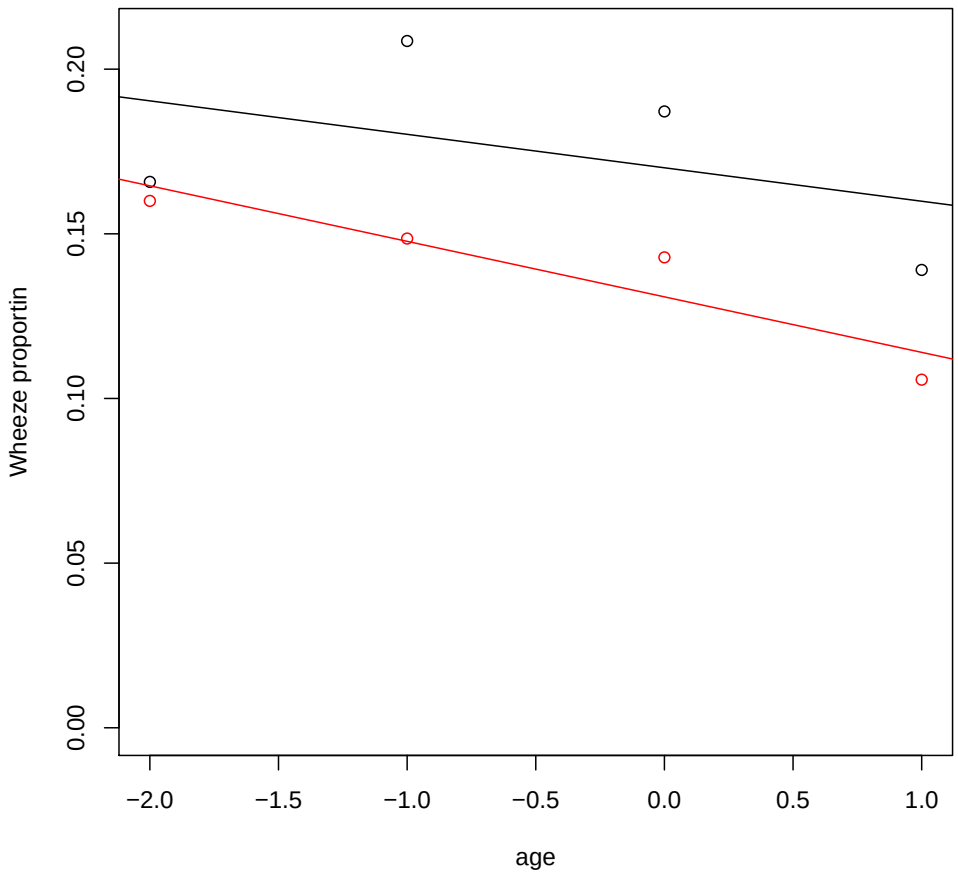
This is called overdispersion and can be due to correlation in the data, latent factors, biological heterogeneity, genetics,....

Latent factors can be modeled explicity using random effects - i.e. mixed models for binary and count data.

Wheezing data

The wheezing (Ohio) data has variables resp (binary indicator of wheezing status), id, age (of child), smoke (binary, mother smoker or not). Each child has 4 observations.

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



Closer look at data

Let Y_{ij} denote wheezing status of ith child at jth age.

Looking at the data I got suspicious.

Consider sum of observations $Y_{i\cdot} = Y_{i1} + \ldots Y_{i4}$ for each child. Possible values 0,1,2,3,4.

Distribution of $Y_{i\cdot}$'s and probabilities for binomial $b(4, 0.15)$:

k	0	1	2	3	4
Proportion equal to k	0.66	0.18	0.08	0.04	0.03
$b(4, 0.15)$	0.52	0.37	0.10	0.01	0.00

There appear to be too many $Y_{i\cdot}$ with value 0 or 4 !

Looking at variances

Assuming Y_{ij} is $b(p_{ij}, 1)$ we try logistic regression

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

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 - if 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:

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

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`

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 '

Residual deviance: 1819.9  on 2145  degrees of freedom
```

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

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 (plus/minus two standard deviations) for U_i is $[-4.686, 4.686]$.

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

Smoke effect: 0.41 (not significant) and age (centered) ranges between -2 to 1 with coefficient -0.18.

Interpretation in terms of marginal variance ?

For linear mixed model we can directly interpret variances of random effects in terms of proportions of variance and intra-class correlation for the response variable.

This is not possible for logistic mixed models.

For logistic mixed regression model, the variance is

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

where the expectation and variance is with respect to U_i in

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

There is no simple formula for the above variance.

Interpretation in terms of odds

The odds are

$$O_i = \frac{p_i}{1-p_i} = \exp(\alpha + \beta z_i + U_i)$$

and the odds ratio between individuals i and j is

$$\frac{O_i}{O_j} = \exp(\beta(z_i - z_j) + U_i - U_j)$$

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

Larsen et al. (2000) suggested to consider the median summary

$$\text{MOR} = \exp[\beta(z_i - z_j) + \text{MED}(|U_i - U_j|)] = \exp[\beta(z_i - z_j)] \exp[\sqrt{2\tau^2}0.6744)$$

between individuals i and j.

Here factor $\exp[\sqrt{2\tau^2}0.6744)]$ is median odds ratio between the individual (i or j) with highest random effect and the individual with lowest random effect (note we consider absolute value $|U_i - U_j|$).

95% intervals for probabilities or odds

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

Hence odds O_i in interval

$$[\exp(\alpha + \beta z_i - 1.96\tau); \exp(\alpha + \beta z_i + 1.96\tau)]$$

with probability 95%.

For probability p_i the interval is

$$\left[\frac{\exp(\alpha + \beta z_i - 1.96\tau)}{1 + \exp(\alpha + \beta z_i - 1.96\tau)}; \frac{\exp(\alpha + \beta z_i + 1.96\tau)}{1 + \exp(\alpha + \beta z_i + 1.96\tau)}\right]$$

Wheezing data

With $\tau = 2.343$ we get MOR=9.34.

That is, keeping all fixed factors equal ($z_i = z_j$), 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 probability of 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.

Computation

Due to non-linear relation between mean of observations and random effects, computation of likelihood is not straightforward.

Huge statistical literature on how to compute good approximations of the likelihood.

`glmer` uses numerical integration (adaptive Gaussian quadrature) and the accuracy is controlled using the argument `nAGQ` (default is `nAGQ=1`).

SPSS use so-called penalized quasi-likelihood based on (very crude) approximation of likelihood.

For the wheeze data set `R` and SPSS estimates differ but we get qualitatively similar results regarding significance of fixed effects.

Wheeze results with different values of nAGQ

5 quadrature points:

```
> fiter5=glmer(resp~age+smoke+(1|id),family=binomial,
                                          data=ohio,nAGQ=5)
Groups Name        Variance Std.Dev.
 id     (Intercept) 4.198    2.049
Number of obs: 2148, groups:  id, 537

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.02398    0.20353 -14.857  < 2e-16 ***
age         -0.17319    0.06718  -2.578  0.00994 **
smoke        0.39448    0.26305   1.500  0.13371
```

10 quadrature points:

```
> fiter10=glmer(resp~age+smoke+(1|id),family=binomial
                                        ,data=ohio,nAGQ=10)
Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 4.614    2.148
Number of obs: 2148, groups:  id, 537

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.08959    0.21557 -14.332  < 2e-16 ***
age         -0.17533    0.06762  -2.593  0.00952 **
smoke        0.39799    0.27167   1.465  0.14293
```

Some sensivity regarding variance estimate. Fixed effects results quite stable.

Results with 20 quadrature points very similar to those with 10 quadrature points.

Summary

- logistic regression very useful for binary data where linear normal models not appropriate.
- 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.

Exercises

1. 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`.
   - 1.1 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 ?
   - 1.2 What can you conclude about the STOCK effects ?
   - 1.3 Is there a BRANCHNR effect ? Does this make sense ?