Mixed models for binary and count data

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

October 9, 2017

Variance for binomial and Poisson

For binomial and Poisson 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, $n = 1$:

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

Y Poisson med middelværdi λ:

$$\mathbb{E}Y = \lambda \quad \mathbb{V}\mathrm{ar}\, Y = \lambda$$

Overdispersion

Binomial and Poisson default models in case of binary and count data.

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

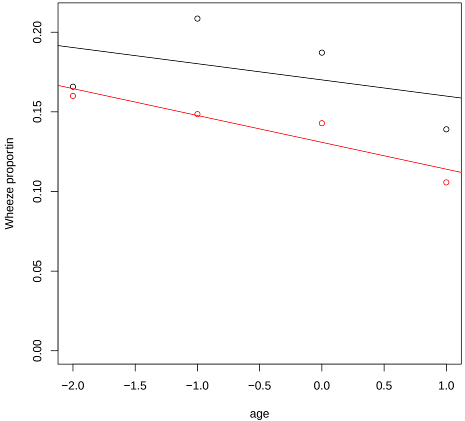
This is called overdispersion and can be due to correlation in the data, latent factors, biological heterogeneity, genetics,....

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).

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



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

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

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 ’
```

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=o
> 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.

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.

Poisson regression with random effects

For count data we can also add random effects to the model.

Recall in this case we model the logarithm of the mean λ_i for ith observation Y_i:

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

Again use `glmer` but now with family `poisson`.

Interpretation of variance components

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 and Poisson mixed models.

E.g. for logistic regression, 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 this variance.

Here $p_i(1-p_i)$ is the conditional variance of Y_i given U_i - but this can not be evaluated since U_i is unobserved.

For the Poisson case with

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

we have (complicated) formulae for the mean

$$\mathbb{E}Y_i = \exp(\alpha + \beta z_i + \tau^2/2)$$

and variance:

$$\mathbb{V}\mathrm{ar}\,Y_i = \mathbb{E}Y_i \left[\exp(\alpha + \beta z_i + 3\tau^2/2) - \exp(\alpha + \tau^2/2) + 1\right]$$

Note: τ^2 not a simple proportion of total variance.

Formula indeed shows that $\tau^2 > 0$ gives overdispersion:

$$\frac{\mathbb{V}\mathrm{ar}\,Y_i}{\mathbb{E}Y_i} = \exp(\alpha + \beta z_i + 3\tau^2/2) - \exp(\alpha + \tau^2/2) + 1 > 1$$

if $\tau^2 > 0$.

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 and Poisson regression very useful for binary and count 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. Consider again the epilepsy data. Introduce subject specific random intercepts. What is the fitted variance for the random intercepts ? Compare the results regarding fixed effects with those of the previous analysis.