Multilevel Models for Binary Responses

The Bangladesh Demographic and Health Survey 2004 Dataset

This tutorial will be analysing data from the Bangladesh Demographic and Health Survey (BDHS), a nationally representative cross-sectional survey of women of reproductive age (13-49 years).

The response variable (antemed) is a binary indicator of whether a woman received antenatal care from a medically-trained provider (a doctor, nurse or midwife) at least once before her most recent live birth.

In this practical, multilevel models are used to allow for and to explore between-community variance in antenatal care. The data have a two-level hierarchical structure with 5366 women at level 1, nested within 361 communities at level 2. In rural areas a community corresponds to a village, while an urban community is a neighbourhood based on census definitions.

A range of predictorvariables will be considered. At level 1, variables such as a woman’s age at the time of the birth and education. Level 2 variables include an indicator of whether the region of residence is classified as urban or rural. Further community-level measures can be derived by aggregating woman-level variables, for example the proportion of respondents in the community who are in the top quintile of a wealth index.

The file contains the following variables:

Variable name	Description and codes
comm	Community identifier
womid	Woman identifier
antemed	Received antenatal care at least once from a medically-trained provider, e.g. doctor, nurse or midwife (1=yes, 0=no)
bord	Birth order of child (ranges from 1 to 13)
mage	Mother's age at the child's birth (in years)
urban	Type of region of residence at survey (1=urban, 0=rural)
meduc	Mother's level of education at survey (1=none, 2=primary, 3=secondary or higher)
islam	Mother's religion (1=Islam, 0=other)
wealth	Household wealth index in quintiles (1=poorest to 5=richest)

Open the Stata dataset antenatal.dta

In [1]:
```
library (foreign)
bang <- read.dta ("antenatal.dta")
bang$comm <- as.factor (bang$comm)
```

The third command above ensures that the variable "comm", which identifies the community in which each woman lives, is treated as a factor variable. This will have no impact on the estimation of models, but will ensure commands used in previous sessions for presenting model findings

Two-level Null Model

The "table" command can be sued to check that the dependent variable ("antemed") is indeed coded 0,1, and to see the number of cases in each of those categories. In this case, the two categories are relatively evenly split meaning that logistic regression can be considered an appropriate form of model.

In [2]:
```
table (bang$antemed)
```

```
   0    1
2613 2753
```

As in previous seesions begin by fitting a null or empty two-level model; that is a model with only an intercept and community effects. The fitting of multi-level logit models is achieved through the "glmer" command in the "lme4" package. This is the same package used in previous sessions and is loaded as shown below.

In [3]:
```
library (lme4)
```

```
Loading required package: Matrix

Warning message:
“package ‘Matrix’ was built under R version 4.2.3”
```

The "glmer" command is used for multilevel mixed-effects generalised linear models. The broad syntax of this command is the same as the "lmer" command used in previous tutorials. The "glmer" command requires additional arguments to be provided, notably "family" which states the distribution that the distrbution the dependent variable follows.

This session begins by fitting a random effects logit regression of y (with no explanatory variables), also known as variance components model, for the probability of receiving antenatal care at least once with community random effects. The formula of the model is:

The R syntax below fits the two-level logit model (with women clustered in communities). Note that as the dependent variable is binary, the family is set to "binomial". The link argument can be used to specifiy the link function eg. link=logit requests a logit regression model is estimated (but this is the default option so can be left out).

In [4]:
```
nullmodel <- glmer (antemed~1+(1|comm), data=bang, family=binomial)
summary (nullmodel)
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: antemed ~ 1 + (1 | comm)
   Data: bang

     AIC      BIC   logLik deviance df.resid
  6639.5   6652.7  -3317.8   6635.5     5364

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.7779 -0.7458  0.3423  0.7118  2.6784

Random effects:
 Groups Name        Variance Std.Dev.
 comm   (Intercept) 1.464    1.21
Number of obs: 5366, groups:  comm, 361

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.14809    0.07178   2.063   0.0391 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The intercept consists of two components: a fixed effect , shared by all communities, and a random effect u0j, specific to community j. The random effect is assumed to follow a normal distribution with covariance matrix which in this simple model contains just one element, the between-community variance .

From the estimates of the model above, we can say that the log-odds of receiving antenatal care from a medically-trained provider in an ‘average’ community (one with u0j = 0) is estimated as β0 = 0.148. We can calculate the odds by exponentiating the estimated coefficient for β0 for an ‘average’ community (with uj = 0) such as exp(0.148) = 1.16, and the corresponding probability is 1.16/(1+1.16) = 0.53.

The intercept for community j is 0.148 + uj , where the variance of uj is estimated as 1.464.

As with linear models, log likelihood values can be used to compare model fit and establish if the multilevel approach is required (i.e. the multilevel model fits the data better than a single level model). The syntax below runs a single level logit model, and reports the log likelihood values for each model.

The difference in log likelihood values is circa 400, so comparing to the chi-squared distribution with 1 degree of freedom (note df=1 for the single level model and 2 for the two-level model) suggests the multilevel model is most appropriate.

In [5]:
```
singlelogit <-glm (antemed~1, data=bang, family=binomial)
summary(singlelogit)
logLik(singlelogit)
logLik(nullmodel)
```

```
Call:
glm(formula = antemed ~ 1, family = binomial, data = bang)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.200  -1.200   1.155   1.155   1.155

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.05219    0.02731   1.911    0.056 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7435.2  on 5365  degrees of freedom
Residual deviance: 7435.2  on 5365  degrees of freedom
AIC: 7437.2

Number of Fisher Scoring iterations: 3

'log Lik.' -3717.601 (df=1)

'log Lik.' -3317.762 (df=2)
```

Looking at Residuals

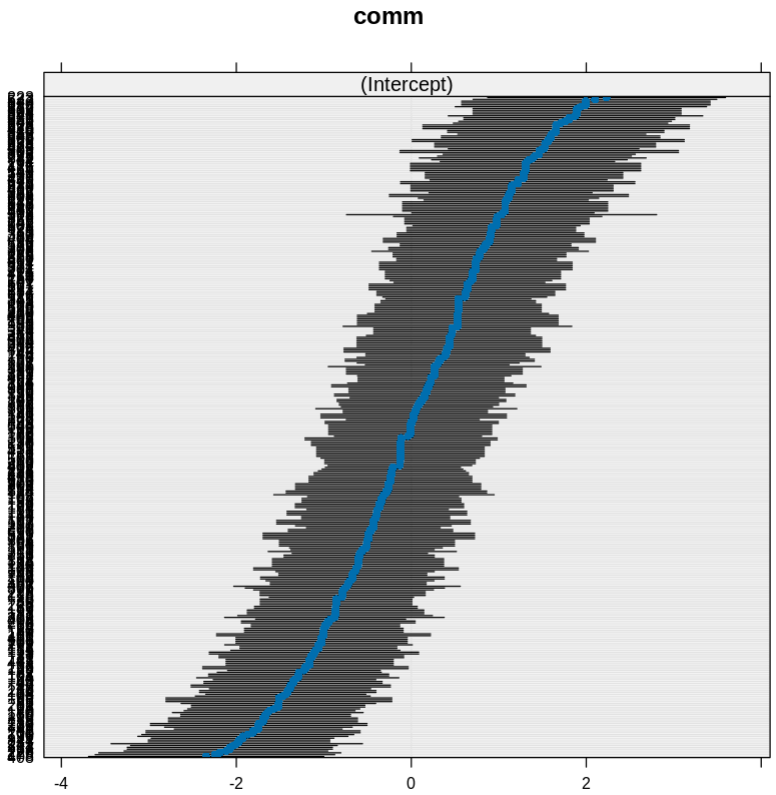
Note that are no level-1 residuals in logit models as expected value = pi

As with previous linear models, it is possible to the level 2 residuals in order to get a sense of the amount of variation between communities. The syntax below produces the relevant caterpillar plot showing variation in intercepts between the different communities based on "nullmodel". This syntax is the same as in previous sessions.

In [6]:
```
library (lattice)
dotplot(ranef(nullmodel))
```

```
Warning message:
“package ‘lattice’ was built under R version 4.2.3”

$comm
```



The plot shows the estimated residuals for all 361 communities in the sample. For a number of communities, the 95% confidence interval does not overlap the vertical line at zero, indicating that uptake of antenatal care in these communities is significantly above average (to the right of the zero line) or below average (to the left of the zero line). The relatively large standard errors are due to the small number of individuals (sample size) in each of the communities, leading to large standard errors for the estimated community residual u0j.

Adding an Explanatory Variable - Random Intercept Model

Next, expand the model to include maternal age as an explanatory variable in the model.

Expanding the null multilevel logit regression ("nullmodel" above) to be a random intercept model is teh same as for previous linear models; the explanatory variable(s) are added to the model fomula, seperated by "+" as shown below, i.e the syntax below would add a woman's age as a level 1 explanatory variable.

In [7]:
```
ri1 <- glmer (antemed~ 1+mage + (1|comm), data=bang, family=binomial)
summary (ri1)
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: antemed ~ 1 + mage + (1 | comm)
   Data: bang

     AIC      BIC   logLik deviance df.resid
  6603.4   6623.2  -3298.7   6597.4     5363

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.9757 -0.7431  0.3357  0.7190  3.2358

Random effects:
 Groups Name        Variance Std.Dev.
 comm   (Intercept) 1.462    1.209
Number of obs: 5366, groups:  comm, 361

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.909350   0.142570   6.378 1.79e-10 ***
mage        -0.032357   0.005235  -6.181 6.37e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
     (Intr)
mage -0.864
```

This model has run without hitch, but sometimes there can be issues where Noteable fails to converge on set of estimates for parameters. If this occurs it may be linked to having too many variables in your model (overparamterisation) or having strongly correlated independent variables. The software may be telling you that a number of solutions in terms of coefficient estimates are yielding a very similar model fit.

However, sometimes it may be that specifying a different way to fit a model can overcome this issue. For this reason we show an adaption of the model fitting procedure and recode that may be useful if you hit this problem in future.

In the case of generalised mixed effect models, the "nAGQ" argument is often useful for achieving convergance. This argument states how many points are used for evaluating the adaptive Gauss-Hermite approximation to the log-likelihood. The default value of 1 corresponding to the Laplace approximation. Values greater than 1 produce greater accuracy in the evaluation of the log-likelihood at the expense of speed. A value of zero uses a faster but less exact form of parameter estimation for GLMMs.

Hence the command below requests more points be used, providing a more "accurate result", and allowing converagnce to be achieved

In [8]:
```
ri2 <- glmer (antemed~ 1+mage + (1|comm), data=bang, family=binomial, nAGQ=10)
summary (ri2)
```

```
Generalized linear mixed model fit by maximum likelihood (Adaptive
  Gauss-Hermite Quadrature, nAGQ = 10) [glmerMod]
 Family: binomial  ( logit )
Formula: antemed ~ 1 + mage + (1 | comm)
   Data: bang

     AIC      BIC   logLik deviance df.resid
  6594.6   6614.3  -3294.3   6588.6     5363

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.9947 -0.7416  0.3336  0.7184  3.2502

Random effects:
 Groups Name        Variance Std.Dev.
 comm   (Intercept) 1.5      1.225
Number of obs: 5366, groups:  comm, 361

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.909347   0.144074   6.312 2.76e-10 ***
mage        -0.032335   0.005284  -6.119 9.41e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
     (Intr)
mage -0.863
```

The fitted line for a given community will differ from the average line in its intercept, by an amount for community j. A plot of the predicted

Expanding the Model

It is possible to think of several ways in which the basic random intercept model above can be expanded:

```
1) Adding a quadratic function of age to teh model to check if the relationship with age is curvi-linear
2) Adding additional explanatory variables (at both the indivdual and community levels)
3) Testing to see if random slopes associated with different level 1 explanatory variables add to the explana
tion.
```

You may consider any of the ideas above that you think are of interest and try modifying the above syntax to run appropriate regression models.

N.B the changes to syntax required to introduce random slopes is the same as in lab session 3, that is to say that the lower level variables that are to be treated as random should be included in the random part of the model formula,

i.e. rs1 <- glmer (antemed~ 1+mage2 + (1+mage2|comm), data=bang, family=binomial)

Variance Partition Coefficient

Level 1 variance in a multilevel logit model has a fixed value of 3.29 (pi). VPC measures can therefore be calculated as in lab session 2, but using 3.29 as the level 1 variance.

Going back to model "nullmodel", level 2 variance is given as 1.462.

The variance partition coefficient (VPC) is calculated as 1.462/(1.462+3.29) = 0.308.

The approximation suggests that 30.8% of the variation in the use of antenatal medical care is attributable to the community level. However, this is statistics is contested and you should note that there are alternative ways of approximating the VPC in a multilevel logistic regression. In particular you should be very cautious on comparing the VPC across models. As we add level 1 explanatory variable, and assuming such variables have some explanatory power then it might be argued that the amount of unexplained variance will be less than in earlier models with fewer parameters; yet it is treated as 3.29 in both all multilevel models.

A more approprate approach is to simply evaluate if the level 2 variance has changed across models, as with the comparrison made earlier where it was suggested that introducing a woman's age to the model had done little to explain variation in outcome between communities.

In [ ]: