HW3

Charlotte Li

5/8/2020

```
setwd("~/Documents/nyu/1stGradSpring/Generalized Linear Model/dataset")
#setwd("~/Desktop/Generalized Linear Model/dataset")
math<-read.csv("math.csv")
attach(math)
head(math)
```

```
##   student school minority    sex    ses mathlev size sector pracad disclim
## 1       1   1224       No Female -1.528       0  842 Public   0.35   1.597
## 2       2   1224       No Female -0.588       1  842 Public   0.35   1.597
## 3       3   1224       No   Male -0.528       1  842 Public   0.35   1.597
## 4       4   1224       No   Male -0.668       0  842 Public   0.35   1.597
## 5       5   1224       No   Male -0.158       1  842 Public   0.35   1.597
## 6       6   1224       No   Male  0.022       0  842 Public   0.35   1.597
##   meanses
## 1  -0.428
## 2  -0.428
## 3  -0.428
## 4  -0.428
## 5  -0.428
## 6  -0.428
```

Q1

Total number of schools and average students in each school

```
#number of schools
length(unique(school))
```

```
## [1] 160
```

```
##avg students in each school
nrow(math)/length(unique(school))
```

```
## [1] 44.90625
```

Q2. Calculate the math proficiency rate (i.e. the percentage of math proficient students) in each school, and answer the following TRUE/FALSE question. Half of the schools have math proficiency rate lower than 41.59%.

```
prof<-tapply(math$mathlev,math$school,sum)
stu.per.sch<-table(math$school)
table(prof/stu.per.sch<0.4159)/length(table(math$school))
```

```
##
## FALSE  TRUE
##   0.5   0.5
```

```
## median
median(prof/stu.per.sch)
```

```
## [1] 0.4158805
```

True

Q3 Student level predictor

Minoirty, Sex, ses, pracad

Q4 School level predictor

size, disclim, meanses, sector,pracad

Q5 Recode

```
library(plyr)
math$minority2 <- revalue(math$minority, c("Yes"="1", "No"="0"))
math$female <- revalue(math$sex, c("Female"="1", "Male"="0"))
math$public2 <- revalue(math$sector, c("Public"="1", "Catholic"="0"))
```

Q6 Run a logistic regression including only student level predictors

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
Q6<-glm(mathlev~minority2+female+ses,math,family=binomial(logit))
summary(Q6)
```

```
##
## Call:
## glm(formula = mathlev ~ minority2 + female + ses, family = binomial(logit),
##     data = math)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6752  -0.9983  -0.6310   1.1164   2.2928
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.00689    0.03962   0.174    0.862
## minority21    -0.77528    0.06249 -12.407   < 2e-16 ***
## female1     -0.41595    0.05103  -8.152 3.59e-16 ***
## ses          0.73691    0.03610  20.413  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9727.3  on 7184  degrees of freedom
## Residual deviance: 8853.6  on 7181  degrees of freedom
## AIC: 8861.6
##
## Number of Fisher Scoring iterations: 3
```

Q7 Suggest at least one way to improve the logistic regression you run in question 6 that will help us better understand the effects of various factors on students' math proficiency. Briefly explain why.

Modeling the cluster effects via random effect coefficients A regression model for clustered data that include both the fixed effect and random is called mixed effect model. Multilevel models, random effect models, random coefficients models, hierarchical models.

Q8 Run a random effect logistic segression with an additional school level random effect

```
Q8<-glmer(mathlev~minority2+female+ses+(1|school),math,family=binomial(logit))
summary(Q8)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: mathlev ~ minority2 + female + ses + (1 | school)
##    Data: math
##
##      AIC      BIC   logLik deviance df.resid
##   8683.3   8717.7  -4336.7   8673.3     7180
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9595 -0.7684 -0.4105  0.8871  3.5781
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  school (Intercept) 0.2834   0.5323
## Number of obs: 7185, groups:  school, 160
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.02711    0.06178  -0.439    0.661
## minority21  -0.84426    0.07743 -10.904  < 2e-16 ***
## female1     -0.39655    0.05915  -6.705 2.02e-11 ***
## ses          0.64531    0.04052  15.925  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) mnrt21 femal1
## minority21 -0.309
## female1    -0.496  0.027
## ses        -0.093  0.128  0.024
```

Q9 Variance of random effect

0.2834

Q10 Explain variance

The variance of the random effect explains the variability between schools.The random effect is 0.2834 suggesting it explains 0.2834 of the variance in log odds of students' math proficiency level.

Q11 ICC

```
0.2834 /(0.2834 +(3.1415926)^2/3)
```

```
## [1] 0.07931115
```

Q12

```
exp(Q6$coefficients)
```

```
## (Intercept)  minority21     female1         ses
##   1.0069139   0.4605742   0.6597162   2.0894641
```

```
exp(Q8@beta)
```

```
## [1] 0.9732550 0.4298769 0.6726389 1.9065827
```

After exponentiate the coefficients from Q6 and Q8, the exp(betas) did not differ too much. If there is a big random effect, then the difference would be big between the two approaches. Yet, the difference is rather small. ICC is only 0.08, meaning only 8% of the variation is in school level random effect, and that is pretty small.

#Q13 The random effect logistic regression model in question 8 has smaller AIC value than the logistic regression model in question 6. True

Q14 Expand Q8 model by including school level predictors

```
Q14<-glmer(mathlev~minority2+female+ses+size+public2+pracad+disclim+meanses+(1|school),math,family=bino
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.0206416 (tol = 0.002, component 1)
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unider
##  - Rescale variables?;Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
```

```r
summary(Q14)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: mathlev ~ minority2 + female + ses + size + public2 + pracad +
##     disclim + meanses + (1 | school)
##    Data: math
##
##      AIC      BIC   logLik deviance df.resid
##   8602.4   8671.2  -4291.2   8582.4     7175
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.7102 -0.7653 -0.4082  0.8802  4.3010
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  school (Intercept) 0.1161   0.3407
## Number of obs: 7185, groups:  school, 160
##
## Fixed effects:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.758e-01  1.979e-01  -3.415 0.000637 ***
## minority21  -8.487e-01  7.561e-02 -11.225  < 2e-16 ***
## female1     -4.061e-01  5.771e-02  -7.038 1.96e-12 ***
## ses          5.701e-01  4.140e-02  13.771  < 2e-16 ***
## size         2.198e-04  7.089e-05   3.100 0.001936 **
## public21    -1.883e-01  1.268e-01  -1.485 0.137674
## pracad       9.796e-01  2.666e-01   3.674 0.000239 ***
## disclim     -1.052e-01  6.077e-02  -1.731 0.083472 .
## meanses      2.389e-01  1.364e-01   1.751 0.079986 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##            (Intr) mnrt21 femal1 ses    size   pblc21 pracad disclm
## minority21 -0.006
## female1    -0.186  0.025
## ses        -0.026  0.078  0.023
## size       -0.366 -0.102  0.014  0.001
## public21   -0.537  0.093 -0.019 -0.003 -0.231
## pracad     -0.861 -0.116  0.054  0.005  0.091  0.377
## disclim     0.061 -0.028  0.106 -0.005 -0.060 -0.491  0.204
## meanses     0.437  0.223  0.002 -0.246 -0.104 -0.084 -0.564  0.004
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
## convergence code: 0
## Model failed to converge with max|grad| = 0.0206416 (tol = 0.002, component 1)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
## Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
```

Significant variables: minoirty, female, ses, size,pracad

15 The variance of the random effect in Q14 is smaller than in Q8

true