Stat 744 Assignment 5

Yusang Hu 400177333

Oct 10, 2019

Introduction

"bdf" is a data set describes the language scores of 8-graders in the Netherlands. There are 28 variables in this data set, I'll pick some of them to do a research. I'll choose 'repeatgr' as response variable, 'IQ.perf','Sex','Minority','ses','mixedgra' and 'schoolNR' as predictos, where 'sex','Minority','mixedgra','IQ.perf','ses' are fixed effects and 'schoolNR' is random effect. Following are descriptions of these variables: repeatgr: Factor with 3 levels '0', '1', '2'. An ordered factor indicating if one or more grades have been repeated. IQ.perf: Number. A numeric vector of IQ scores. sex: Factor with 2 levels '0', '1'. Sex of the student. Minority: Factor with 2 levels 'N', 'Y'. A factor indicating if the student is a member of a minority group. ses: Number. A numeric vector of socioeconomic status indicators. mixedgra: Factor with 2 levels '0', '1'. A factor indicating if the class is a mixed-grade class. schoolNR: Factor with 131 levels. A factor denoting the school. From above we can observe that there are two types of fixed effects: factor and number. My research of interest is finding out the predictor that has more effects on students' likelihood to repeat a grade in each parameter type (i.e. factor, number). I'll fit the model by using both generalized linear mixed model and generalized linear model, and compare estimators between them by using ggplot.

Generalized linear mixed model

First I'll fit a mixed effects model.

```
library(ggplot2)
theme_set(theme_bw()+theme(panel.spacing=grid::unit(0,"lines")))
library(directlabels)
library(lme4)

## Loading required package: Matrix

library(dotwhisker)
library(mlmRev)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(purrr)
library(broom)
m1=glmer(repeatgr~IQ.perf+sex+Minority+ses+mixedgra+(1|schoolNR),data=b
df,family=binomial)
```

Scale continuous parameters

Fit a new model with scaled continuous parameters (i.e. ses, IQ.perf).

```
bdf=bdf%>%
  mutate(ses_sc=drop(scale(ses)),IQperf_sc=drop(scale(IQ.perf)))
m1_sc=update(m1,repeatgr~IQperf_sc+sex+Minority+ses_sc+mixedgra+(1|scho
olNR),data=bdf,family=binomial)
```

Generalized linear model

```
m1_fixed=glm(repeatgr~IQperf_sc+sex+Minority+ses_sc+mixedgra,data=bdf,f
amily=binomial)
```

Make a new dataframe

Now, I'm going to create a new dataframe by using tidy. This dataframe contains many regression results of GLMM and GLM, such as estimate, p value. Since I'm interested in finding out the predictor that has more effects on students' likelihood to repeat a grade in each parameter type, I'll add an extra column called 'para_type', which indicates type of each parameter.

```
m1_tidy <- map(list(GLMM=m1_sc,GLM=m1_fixed),tidy,
               conf.int=TRUE) %>%
  bind_rows(.id="model") %>%
  mutate(term=factor(term,levels=unique(term))) %>%
  filter(term!="(Intercept)")

## Warning in bind_rows_(x, .id): binding factor and character vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

para_type=c('Number','Factor','Factor','Number','Factor','','Number','F
actor','Factor','Number','Factor')
m1_cbtype=cbind(m1_tidy,para_type)
```

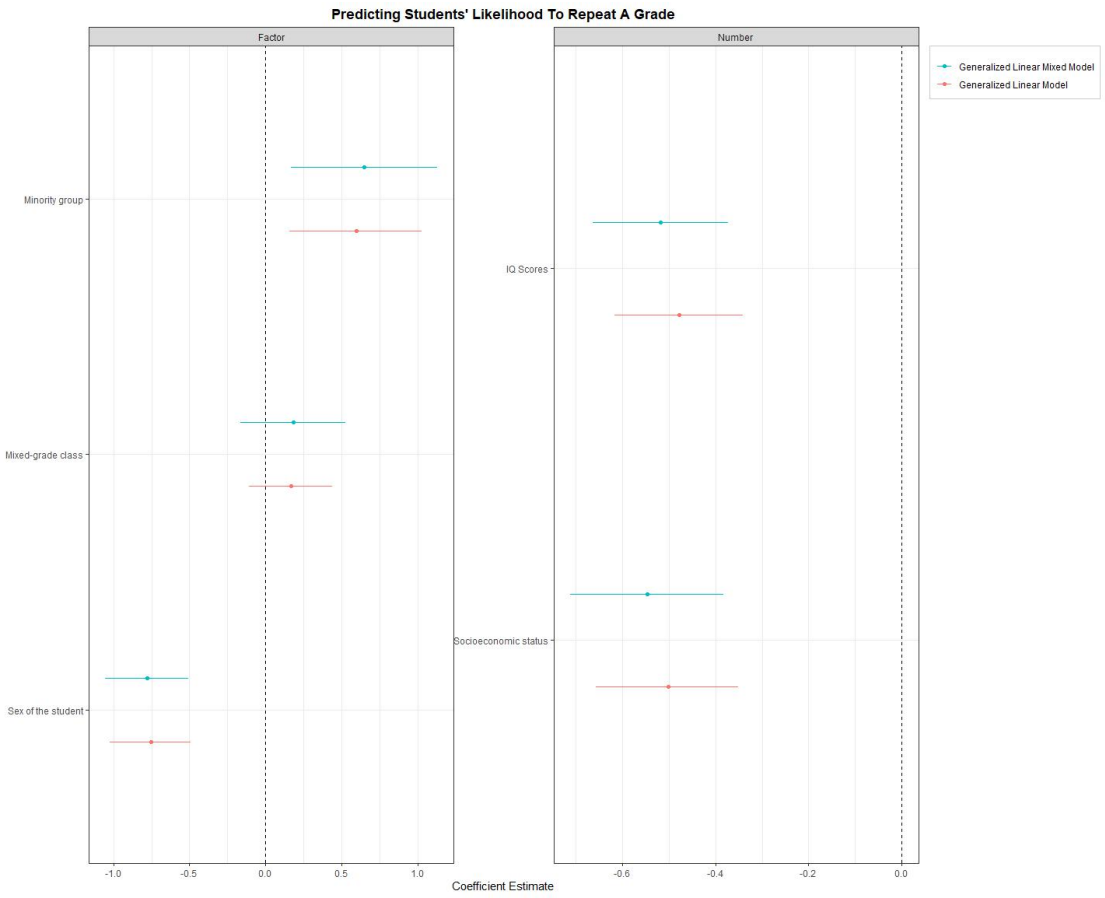
Plot

Note: Comments are attached after each code.

```r
renam=c(sex1="Sex of the student",MinorityY="Minority group",ses_sc="So
cioeconomic status",mixedgra1="Mixed-grade class",IQperf_sc="IQ Scores")
 ## preparation for renaming y-axis scales
pd <- ggstance::position_dodgev(height=0.5)
g1 <- ggplot(m1_cbtype[-6,],aes(x=estimate,y=term,colour=model))+ ## m1
_cbtype[-6,] -> drop 6th row of m1_cbtype, since we are not interested i
n the intercept (i.e. random effect 'schoolNR')
  geom_point(position=pd) + labs(y="")+
  ggstance::geom_linerangeh(aes(xmin=conf.low,xmax=conf.high),
                            position=pd)+
  scale_colour_brewer(palette="Dark2") + geom_vline(xintercept=0,lty=2)
+ ## draw a dashed line at estimate=0, this will help readers to underst
and which coefficient is not significant (i.e. including 0)
  facet_wrap(~para_type,scales = "free")+ ## separate plots by parameter
 type, set scales="free" so that first facet only have scales with type
'factor' and second facet only have scales with type 'number'
  xlab("Coefficient Estimate")+ ## change 'estimate' to 'Coefficient Est
imate', in order to add more detailed on the title of x-axis
  ggtitle("Predicting Students' Likelihood To Repeat A Grade")+ ## add t
itle for the plot
  theme(plot.title=element_text(face="bold",hjust=0.5), ## bold and cen
ter the title, to make it more aesthetic
          legend.justification = c(1,1), ## line 76-78: change the layout
 of legend, to make it more conspicuous
          legend.background = element_rect(colour="grey80"),
          legend.title=element_blank())+
  guides(colour=guide_legend(reverse=TRUE))+ ## inverse the legend, to m
ake the plot easier to read
  scale_color_discrete(labels=c("Generalized Linear Model","Generalized
 Linear Mixed Model"))+## rename legends, abbreviation sometimes make re
aders confused; we have already indicated that two legends are 'model',
so we can get rid of the legend title 'model' since it's a repetitive in
formation
  scale_y_discrete(labels=renam) ## rename y-axis scales, describe each
scale in detail

## Scale for 'colour' is already present. Adding another scale for
## 'colour', which will replace the existing scale.

m1_reorder <- mutate(m1_cbtype[-6,],term=reorder(term,estimate)) ## reo
rder y-axis by the magnitude of estimate
g1 %+% m1_reorder
```



Conclusions

From the plot, we can observe that: 1. it is clear that 'Mixed-grade class' (i.e. 'mixedgra1') is not significant since its confidence interval includes zero, so there is insufficient evidence to conclude that this predictor has effect on students' likelihood to repeat a grade. 2. When the type of parameter is 'factor', 'Sex of the student' (i.e. 'sex1') has more (negative) effects on students' likelihood to repeat a grade and 'Minority group' (i.e. 'MinorityY') has more (positive) effects on students' likelihood to repeat a grade; when the type of parameter is 'number', 'Socioeconomic status' (i.e. ses) has more (negative) effects on students' likelihood to repeat a grade. 3. we cannot determine which model is better without other informations, but we can conclude that all significant fixed effects in GLMM have larger effects (i.e. larger coefficients) than in GLM.