STA442 Homework 2, Mixed effects models

student number: 1003942326

Yulin WANG

15/10/2019

Question 1: Math

Method and Results

We analyzed the MathAchieve dataset from the MEMSS package. Firstly, we treated the variable School as a random intercept effect while other three variables as fixed effects and fitted a linear mixed model to model the mathematics achievement scores(MathAch) as a function of School, Minority, Sex, and SES(socio-economic status) as follows:

$$Y_{ij} = \beta_0 + \beta_1 M_{ij} + \beta_2 S_{ij} + \beta_3 SES_{ij} + U_i + Z_{ij}$$

where Y_{ij} is the mathematics achievement score of jth student from ith school; M_{ij} is 1 if minority, 0 otherwise; S_{ij} is 1 if male, 0 female; SES_{ij} is the socio-economic status of jth student from ith school; U_i is the the random effect of ith school, and Z_{ij} is the residual term for jth student from ith school.

Table 1: Estimation of LMM treating School as random intercept

	MLE	Std.Error	DF	t-value	p-value
(Intercept)	12.8847	0.1935	7022	66.5930	0
MinorityYes	-2.9615	0.2058	7022	-14.3932	0
SexMale	1.2298	0.1627	7022	7.5583	0
SES	2.0894	0.1057	7022	19.7664	0
σ	1.9167	NA	NA	NA	NA
τ	5.9924	NA	NA	NA	NA

From the above lme table, we found that all the fixed effects were statistically significant. Holding all the other variables stay fixed, the minority students got lower scores, Males got greater scores than female students, and students with better socio-economic status tended to get higher scores respectively.

Additionally, we can notice that the between-group(School) variance is $\sigma^2 = (1.9167)^2 = 3.673739$ and the within-group(School) variance is $\tau^2 = (5.9924)^2 = 35.90886$ approximately. Then we can calculate the ICC(Intraclass Correlation Coefficient), which is $\frac{\sigma^2}{\sigma^2+\tau^2} = \frac{3.673739}{3.673739+35.90886} = 0.09281197$. This means that about 9.28% of the total variation due to School(i.e. can be explained by School). In other words, the mathematics achievement scores vary among students within a school, but not too much among schools. That is, the observations within schools are no more similar than observations from different schools.

Conclusion

In conclusion, there are not substantial differences between schools, and differences within schools are much greater than differences between students from different schools.

Limitation and furthur research

If we want to further analyze the differences between schools, we could try to fit another linear mixed model treating school as both random intercept and random slope effects by using 'ML' method instead of the default method 'REML'. After that we need to fit a model treating school as only random intercept again by using 'ML' method, and then do the likelihood ratio test between these two models. Here is the test result.

```
##          Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## Math_ML1     1  6 46398.84 46440.11 -23193.42
## Math_ML2     2  8 46397.71 46452.75 -23190.85 1 vs 2 5.127809   0.077
```

Since the p-value for the likelihood ratio test between these two models is about 0.077 which is not statistically significant, and our goal is to fit the simplest possible model, so there is no need to fit a linear mixed model with a random slope effect for School.

Question 2: Drugs

Introduction

We analyzed the Treatment Episode Data Set – Discharges (TEDS-D) by using an R version of the dataset available at https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/35074. The TEDS-D is a national census data system of annual discharges from substance abuse treatment facilities. We had two main analysis purposes. Firstly, we wanted to analyze that chance of a young person completing their drug treatment whether depends on the substance the individual is addicted to or not, with 'hard' drugs (Heroin, Opiates, Methamphetamine, Cocaine) being more difficult to treat than alcohol or marijuana. Additionally, we analyzed that whether some American states have particularly effective treatment programs whereas other states have programs which are highly problematic with very low completion rates.

Methods

Firstly, we cleaned the data(i.e. deleted the missing(NA)) and changed the response "completed" into numeric expression 1/0. In order not to modify the original dataset, we created a new variable "drugType" to rearrange the variable "SUB1" into two groups: "hard" and "soft", which represented two addiction levels of substances.

In order to approach Bayesian inference, we used INLA and treated drugType(hard/soft), GENDER(Male/Female), raceEthnicity, homeless(True/False), and AGE(four different age groups: "12-14", "15-17", "18-20", "21-24") as fixed effects while STFIPS(states) and TOWN as two random effects to fit a generalized linear mixed logistic model to model the log-odds of individual completing the treatment as a function of drugType, GENDER, raceEthnicity, homeless, AGE, STFIPS and TOWN.

We did two null hypotheses(H_0). Firstly, the chance of a young person completing their drug treatment does not depend on the substance the individual is addicted to. Secondly, there is no relationship between the chance of a young person completing their drug treatment and the state locations of thier treatment facilities.

Table 2: Posterior means, standard deviations and quantiles for model parameters.

	mean	sd	0.025quant	0.975quant
(Intercept)				
(Intercept)	0.652	1.1029	0.538	0.790
drugType				
soft	1.408	1.0099	1.381	1.436
GENDER				
FEMALE	0.916	1.0086	0.901	0.932
raceEthnicity				
Hispanic	0.814	1.0119	0.796	0.833
BLACK OR AFRICAN AMERICAN	0.626	1.0121	0.612	0.641
AMERICAN INDIAN (OTHER TH	0.737	1.0360	0.688	0.790
OTHER SINGLE RACE	0.845	1.0329	0.793	0.901
TWO OR MORE RACES	0.827	1.0385	0.768	0.891
ASIAN	1.133	1.0451	1.039	1.236
NATIVE HAWAIIAN OR OTHER	0.843	1.0630	0.748	0.951
ASIAN OR PACIFIC ISLANDER	1.440	1.0901	1.216	1.705
ALASKA NATIVE (ALEUT, ESK	0.864	1.1665	0.638	1.169
homeless				
TRUE	1.009	1.0163	0.977	1.041
AGE18-20				
AGE18-20	0.891	1.0101	0.874	0.909
AGE15-17				
AGE15-17	0.806	1.0112	0.788	0.823
AGE12-14				
AGE12-14	0.840	1.0204	0.807	0.874
SD				
STFIPS	0.581	0.0549	0.480	0.696
TOWN	0.539	0.0293	0.484	0.599

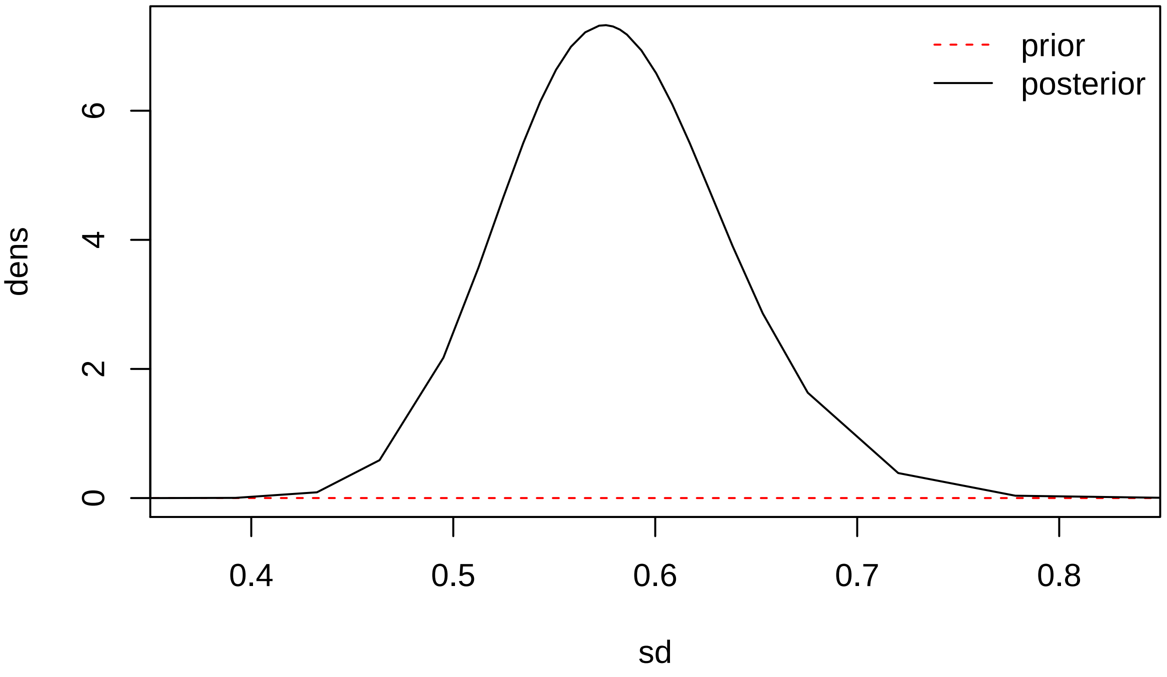


Table 3: Posterior means and quantiles for random effect(STFIPS)

ID	mean	0.025q	0.975q	ID	mean	0.025q	0.975q
ALABAMA	0.2	-0.3	0.7	MONTANA	-0.1	-0.9	0.6
ALASKA	0.0	-0.8	0.8	NEBRASKA	0.8	0.5	1.2
ARIZONA	0.0	-1.1	1.1	NEVADA	-0.1	-0.7	0.5
ARKANSAS	-0.1	-0.7	0.4	NEW HAMPSHIRE	0.2	-0.2	0.7
CALIFORNIA	-0.3	-0.6	0.0	NEW JERSEY	0.5	0.2	0.7
COLORADO	0.6	0.1	1.0	NEW MEXICO	-1.1	-1.8	-0.4
CONNECTICUT	0.1	-0.4	0.6	NEW YORK	-0.3	-0.6	-0.1
DELAWARE	1.0	0.7	1.3	NORTH CAROLINA	-0.9	-1.2	-0.6
WASHINGTON DC	-0.3	-0.6	0.1	NORTH DAKOTA	-0.3	-0.9	0.4
FLORIDA	1.0	0.7	1.3	OHIO	-0.2	-0.5	0.1
GEORGIA	-0.2	-0.8	0.4	OKLAHOMA	0.5	0.0	1.0
HAWAII	0.2	-0.6	1.0	OREGON	0.1	-0.2	0.5
IDAHO	-0.2	-1.0	0.6	PENNSYLVANIA	0.0	-1.1	1.1
ILLINOIS	-0.5	-0.8	-0.3	RHODE ISLAND	-0.2	-0.6	0.2
INDIANA	-0.1	-0.9	0.7	SOUTH CAROLINA	0.3	0.0	0.6
IOWA	0.4	0.1	0.7	SOUTH DAKOTA	0.5	-0.3	1.3
KANSAS	-0.2	-0.5	0.1	TENNESSEE	0.2	-0.2	0.7
KENTUCKY	-0.2	-0.5	0.2	TEXAS	0.6	0.3	0.9
LOUISIANA	-0.6	-1.0	-0.2	UTAH	0.1	-0.5	0.7
MAINE	0.1	-0.7	0.9	VERMONT	-0.2	-1.0	0.6
MARYLAND	0.5	0.2	0.8	VIRGINIA	-2.8	-3.2	-2.5
MASSACHUSETTS	0.8	0.4	1.2	WASHINGTON	-0.1	-0.4	0.3
MICHIGAN	-0.4	-0.7	0.0	WEST VIRGINIA	0.0	-1.1	1.1
MINNESOTA	0.4	0.0	0.9	WISCONSIN	0.0	-1.1	1.1
MISSISSIPPI	0.0	-1.1	1.1	WYOMING	0.0	-1.1	1.1
MISSOURI	-0.4	-0.7	-0.1	PUERTO RICO	0.6	-0.1	1.2

Results

Firstly, from the table 2, we could notice that all the parameters are statistically significant since all the credible intervals do not contain 0.

For the variable drugType, there is a positive relationship between the log-odds of completing the treatment and the 'soft' drugs. The log-odds of a young person who is addicted to 'soft' drugs is about 1.408 greater than that of a young person addicted to 'hard' drugs. That is, the treatment completion rate for 'hard' drugs addicts is lower than that for 'soft' drugs addicts.

From the prior and posterior plot for State-level standard deviation, we can find that the sd of STFIPS(state) is highly approximately between 0.48 and 0.70, which means that there are differences between different states. And from table 3, we find that parameters for different states have quite different posterior means and quantiles. Additionally, after ignoring those not statistically significant parameters, some states have positive relationship with the log-odds while other have negative relationship instead.

Conclusions

There is a statistically significant relationship bwtween the substance that young person is addicted to and the chance of completing their drug treatment. Completion rate of a young person who is addicted to 'hard' drugs is lower than that of a young person who is addicted to alcohol or marijuana, where the 'hard' drugs include Heroin, Opiates, Methamphetamine, and Cocaine.

Additionally, the treatment given locations are closely correlated to the chance of success treatment. Some American states with particularly effective treatment programs have higher treatment completion rates than other states that have highly problematic programs.

Limitation and furthur research

During our analysis, we did not take the interactions between different predictors into consideration. For instance, people with different ages may have different resistance to drugs living in different states and towns. Thus, we could conduct further research with adding some interaction terms.

Appendix

Question 1: Math

```
library('Pmisc')
library('nlme')
library('Matrix')
library('sp')
library('parallel')
library('INLA')
data("MathAchieve", package = "MEMSS")
Matdata <- na.omit(MathAchieve)    #delete missing data


# fit the model treating school as a random intercept effect
Math_lme1 <- lme(MathAch ~ Minority + Sex + SES, random = ~1 | School, data = Matdata)
#summary(Math_lme1)
knitr::kable(Pmisc::lmeTable(Math_lme1), digits = 4, caption = "Estimation of LMM treating School as ran


#do the likelihood ratio test by using method='ML'
Math_ML1 <- lme(MathAch ~ Minority + Sex + SES, random = ~1 | School, data = Matdata, method='ML')
Math_ML2 <- lme(MathAch ~ Minority + Sex + SES, random = ~1+ SES | School, data = Matdata, method='ML')
# compare random intercept and random slope
anova(Math_ML1, Math_ML2)
```

Question 2: Drugs

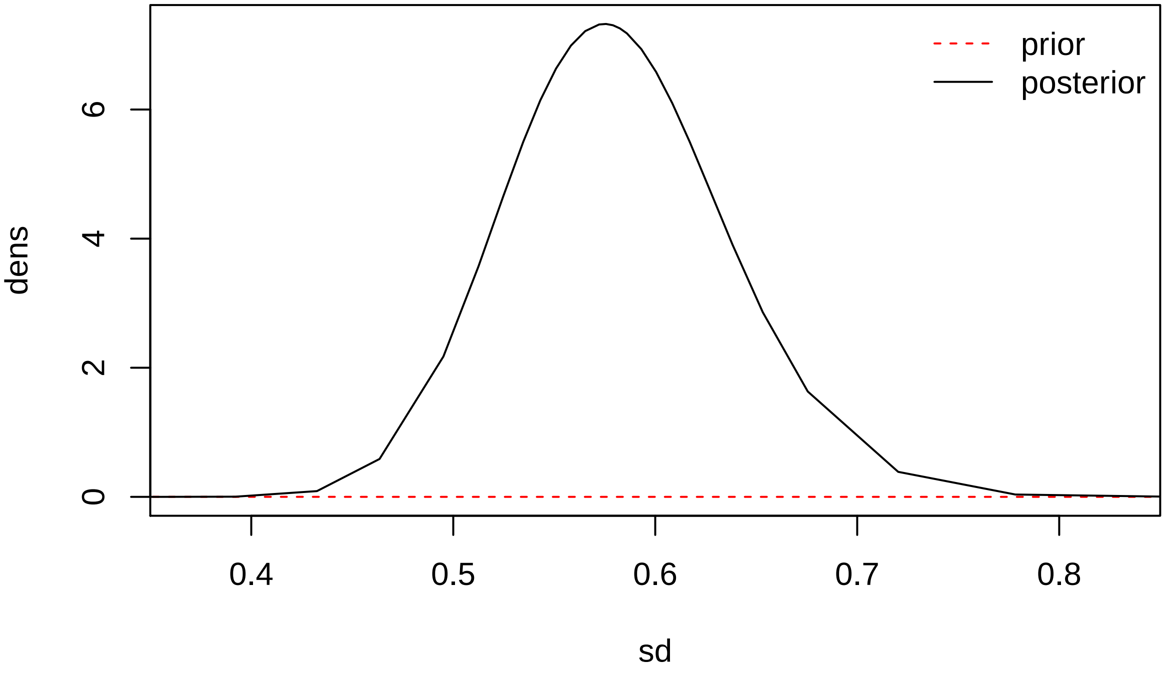
```
#download.file("http://pbrown.ca/teaching/appliedstats/data/drugs.rds", "drugs.rds")
xSub = readRDS("drugs.rds")
forInla = na.omit(xSub) #delete the missing data
forInla$y = as.numeric(forInla$completed) #change response into numeric expression 1/0

#create a new variable called drugType that groups the SUB1 into 'hard' and 'soft'
forInla$drugType = rep(NA,dim(forInla)[1])
forInla$drugType[forInla$SUB1 != '(4) MARIJUANA/HASHISH' & forInla$SUB1 != '(2) ALCOHOL'] <- 'hard'
forInla$drugType[forInla$SUB1 == '(4) MARIJUANA/HASHISH' | forInla$SUB1 == '(2) ALCOHOL'] <- 'soft'
#convert drugType to factor
forInla$drugType <- factor(forInla$drugType)

library("INLA")
#fit the model
ires = inla(y ~ drugType + GENDER + raceEthnicity + homeless + AGE +
              f(STFIPS, hyper=list(prec=list( prior='pc.prec', param=c(0.1, 0.05)))) + f(TOWN), data=for


#generate the prior and posterior plot for State-level standard deviation
```

```
sdState = Pmisc::priorPostSd(ires)
do.call(matplot, sdState$STFIPS$matplot)
do.call(legend, sdState$legend)
```



```
#generate a table of posterior means, standard deviations and quantiles for model parameters.
toPrint = as.data.frame(rbind(exp(ires$summary.fixed[, c(1, 2, 3, 5)]), sdState$summary[, c(1, 2, 3, 5)]
sss = "^(raceEthnicity|drugType|GENDER|homeless|age group|SD)(.[[:digit:]]+.[[:space:]]+| for )?"
toPrint = cbind(variable = gsub(paste0(sss, ".*"), "\\1", rownames(toPrint)),
                category = substr(gsub(sss,"", rownames(toPrint)), 1, 25), toPrint)

Pmisc::mdTable(toPrint, digits = 3, mdToTex = TRUE,
               guessGroup = TRUE, caption = "Posterior means, standard deviations and quantiles for mode


#generate a table of posterior means and quantiles for random effect STFIPS
ires$summary.random$STFIPS$ID = gsub("[[:punct:]]|[[:digit:]]", "", ires$summary.random$STFIPS$ID)

ires$summary.random$STFIPS$ID = gsub("DISTRICT OF COLUMBIA", "WASHINGTON DC", ires$summary.random$STFIP


toprint = cbind(ires$summary.random$STFIPS[1:26, c(1, 2, 4, 6)],
                ires$summary.random$STFIPS[-(1:26), c(1, 2, 4, 6)])

colnames(toprint) = gsub("uant", "", colnames(toprint))
knitr::kable(toprint, digits = 1, format = "latex", caption = "Posterior means and quantiles for random
```