STA442 Homework2-Mixed Effect Models

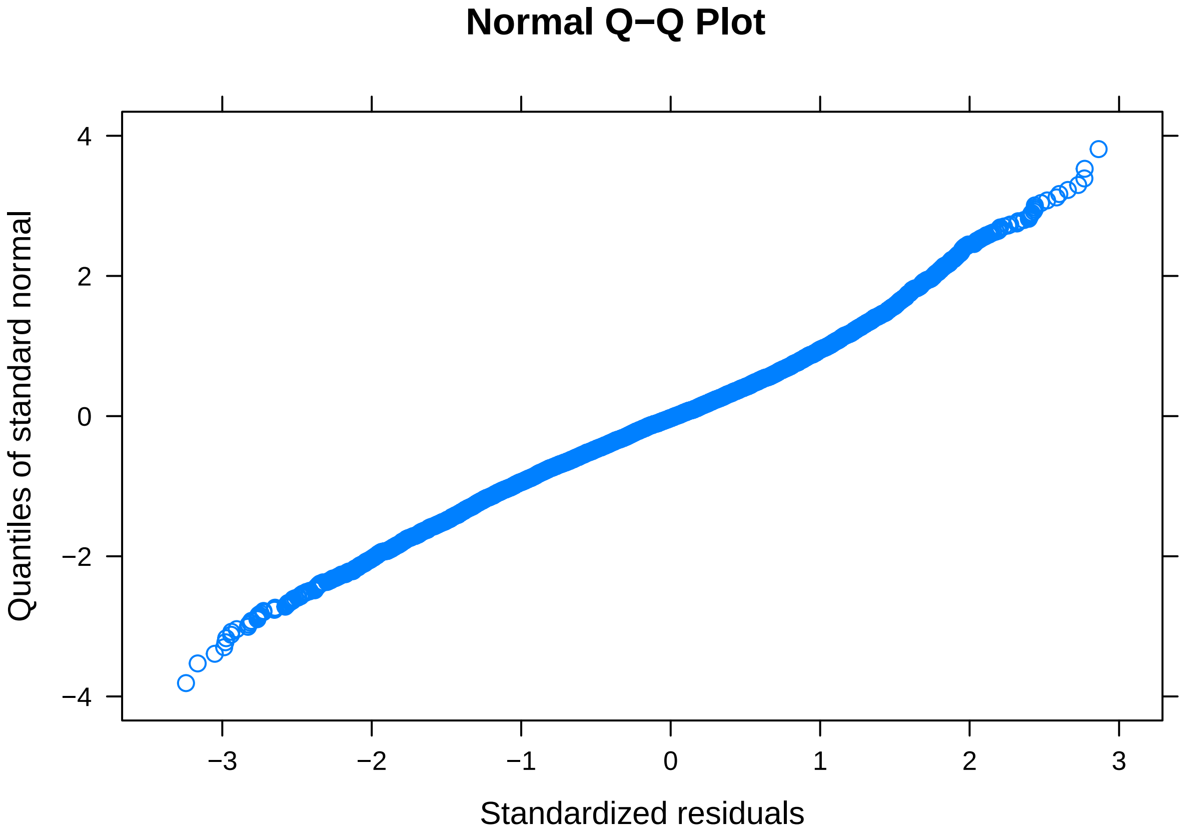
Yingzhu Zhang

16/10/2019

1. Math

Report

In this analysis on MathAchieve dataset, we investigated whether the mathematics achievement scores of students have differences within schools nearly as big as the differences between students from different schools. We treated School as a random effect, since we were interested in the differences in mathematics achievement scores between specific schools. If we were just interested in controlling it to see the differences within one school only, then fixed effect will be more appropriate. The fixed effects in this model are Minority (whether a student is a minority racial group), Sex (female or male), and SES (socio-economic status). Additionaly, the normality of the data is confirmed by the Q-Q Plot below as we can see there is no systmetic deviation.



We fitted the mixed effect model and carried out the analysis treating School as a random effect. The results for the fixed effects were summarized in Table 1. To find whether the difference between schools are greater than within school, we need to check the random effect. In the summary of the model, we found that the standard deviation between school is 1.92, whereas the standard deviation within is 5.99. After calculating the variances by squaring the standard deviations, we can say that only 9.28% of the variance was explained by the random factor.

Table 1: Estimation of fixed effects in linear mixed model of math achievement dataset

	Value	Std.Error	DF	t-value	p-value
(Intercept)	12.88	0.19	7022	66.59	0
MinorityYes	-2.96	0.21	7022	-14.39	0
SexMale	1.23	0.16	7022	7.56	0
SES	2.09	0.11	7022	19.77	0

Hence, we conclude that the differences between schools are not greater than what can be explained by within-school variation. In another word, the math achievement score does not variate much for the students from different schools but does deviate more within one school.

2. Drugs

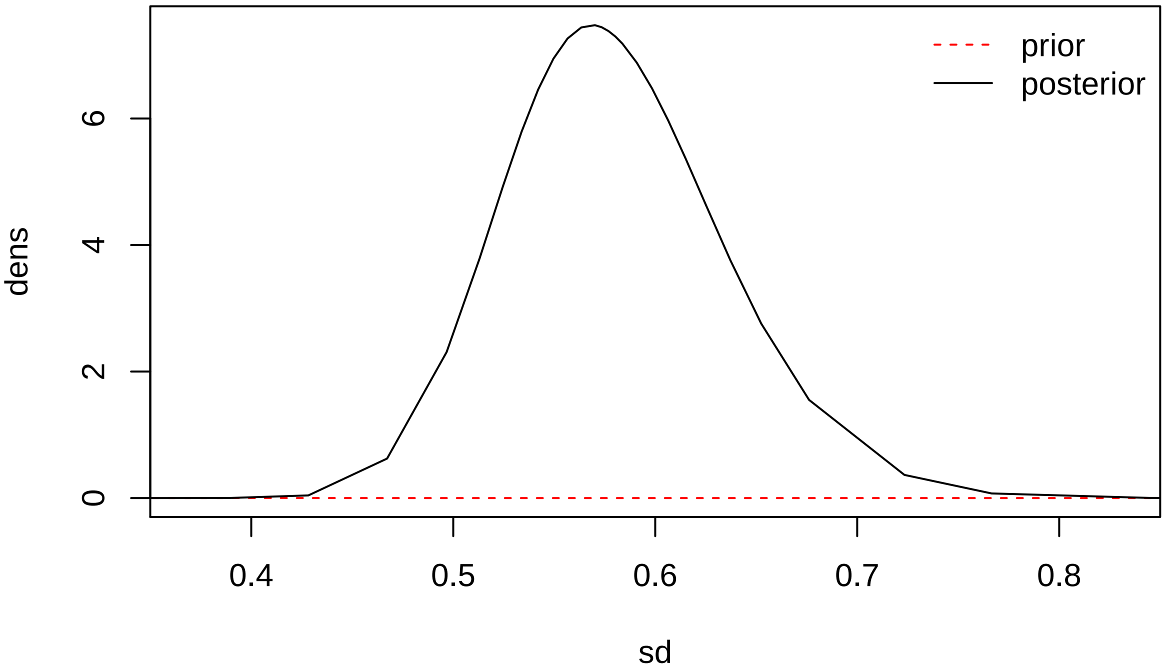
Introduction

We analyzed the Treatment Episode Data Set – Discharges, the dataset provides annual data on the number and characteristics of persons discharged from public and private substance abuse treatment programs that receive public funding. We wanted to test the hypothesis that chance of a young person completing their drug treatment with 'hard' drugs (Heroin, Opiates, Methamphetamine, Cocaine) being more difficult to treat than alcohol or marijuana. Another hypothesis is that some American states have particularly effective treatment programs whereas other states have programs which are highly problematic with very low completion rates. In this analysis using R-studio, we fitted the Generalized Linear Mixed Model and treated SUB1 (types of drugs) and TOWN (American states) as random effects. Finally, by checking the completion rates, we can further investigate the above hypothesis.

Methods

The following plot shows the distribution of the posteriors and penalized complexity prior.

- The posterior probability is this case is the odds of a young man completes his drug treatment.
- The penalized complexity prior probability is some uncertain events, in this case, the types of drug that the individual is addicted to, or the location (states) of the treatment program.



In this analysis, we treated SUB1 and TOWN as a random effect to see the odds of a young man complete the treatment, and the completion rate in different American states. Mathematically, the GLMM model is described as:

$$Y_i \sim Bernoulli(\lambda_i)$$
$$logit(\lambda_i) = X_i\beta + U_i + V_i$$
$$U_i \sim i.i.d.\mathrm{N}(0, \sigma_U^2)$$
$$V_{ij} \sim i.i.d.\mathrm{N}(0, \sigma_V^2)$$

where:

- Y_i = the chance of a young man i completes the drug treatment, hence the completion rate
- $X_i\beta$ = the fixed effect where X_i is the indicator variables
- U_i = SUB1(types of drugs) as an individual-level random effect
- V_i = TOWN (American states) as an individual-level random effect

Additionally, the null hypothesis (H_0) for the first hypothesis test is that the completion rates for all kinds of drugs are the same, the alternative hypothesis (H_a) is that one of the completion rate deviates among other drugs. Correspondingly, H_0 for the second test is that the completion rates for all states are the same, H_a is that one of the completion rate deviates among other states.

To interpret the result, we know that the posteriors for the fixed effects tend to be approximately normal so there is little difference between mean, median and mode. The posterior means indicate the odd of a young man complete treatment programs, hence the completion rate. By comparing the mean of each variables, we can determine whether the hypothesis is correct.

Results

As shown in Table 2 above, the completion rate for alcohol addiction is 1.609 at the 0.5 quantile. Since marijuana is the reference group, the completion rate for it is 1.0. Comparing to other 'hard' drugs, for example, the chance of an individual addicted to heroin complete the treatment is only 0.872 and for cocaine is 0.855, which are significantly lower than alcohol and marijuana.

Table 2: Posterior means and quantiles for model parameters.

	0.5quant	0.025quant	0.975quant
(Intercept)			
(Intercept)	0.716	0.593	0.865
SUB1			
ALCOHOL	1.609	1.574	1.645
HEROIN	0.872	0.849	0.896
OTHER OPIATES AND SYNTHET	0.901	0.874	0.929
METHAMPHETAMINE	0.955	0.917	0.994
COCAINE/CRACK	0.855	0.814	0.899
GENDER			
FEMALE	0.893	0.878	0.909
raceEthnicity			
Hispanic	0.832	0.812	0.851
BLACK OR AFRICAN AMERICAN	0.682	0.666	0.699
AMERICAN INDIAN (OTHER TH	0.728	0.679	0.781
OTHER SINGLE RACE	0.865	0.812	0.923
TWO OR MORE RACES	0.855	0.793	0.921
ASIAN	1.132	1.037	1.235
NATIVE HAWAIIAN OR OTHER	0.845	0.749	0.953
ASIAN OR PACIFIC ISLANDER	1.454	1.227	1.723
ALASKA NATIVE (ALEUT, ESK	0.845	0.624	1.145
homeless			
TRUE	1.005	0.973	1.037
AGE18-20			
AGE18-20	0.935	0.916	0.953
AGE15-17			
AGE15-17	0.926	0.905	0.947
AGE12-14			
AGE12-14	0.973	0.934	1.013
SD			
STFIPS	0.577	0.483	0.698
TOWN	0.518	0.469	0.575

ID	mean	0.025q	0.975q	ID	mean	0.025q	0.975q
ALABAMA	0.2	-0.3	0.7	MONTANA	-0.2	-0.9	0.6
ALASKA	0.0	-0.8	0.8	NEBRASKA	0.8	0.4	1.1
ARIZONA	0.0	-1.1	1.1	NEVADA	-0.1	-0.7	0.5
ARKANSAS	-0.1	-0.6	0.4	NEW HAMPSHIRE	0.2	-0.3	0.6
CALIFORNIA	-0.3	-0.5	0.0	NEW JERSEY	0.5	0.2	0.8
COLORADO	0.5	0.1	0.9	NEW MEXICO	-1.1	-1.8	-0.5
CONNECTICUT	0.1	-0.4	0.6	NEW YORK	-0.3	-0.6	0.0
DELAWARE	1.0	0.7	1.3	NORTH CAROLINA	-0.8	-1.1	-0.6
WASHINGTON DC	-0.3	-0.6	0.1	NORTH DAKOTA	-0.3	-0.9	0.3
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	0.4
IDAHO	-0.2	-0.9	0.6	PENNSYLVANIA	0.0	-1.1	1.1
ILLINOIS	-0.5	-0.8	-0.2	RHODE ISLAND	-0.2	-0.6	0.2
INDIANA	-0.1	-0.8	0.7	SOUTH CAROLINA	0.4	0.1	0.7
IOWA	0.4	0.1	0.7	SOUTH DAKOTA	0.4	-0.3	1.2
KANSAS	-0.2	-0.5	0.1	TENNESSEE	0.3	-0.2	0.7
KENTUCKY	-0.2	-0.5	0.2	TEXAS	0.6	0.3	0.9
LOUISIANA	-0.5	-0.9	-0.2	UTAH	0.1	-0.5	0.6
MAINE	0.1	-0.6	0.9	VERMONT	-0.2	-1.0	0.6
MARYLAND	0.5	0.2	0.8	VIRGINIA	-2.9	-3.2	-2.5
MASSACHUSETTS	0.8	0.4	1.2	WASHINGTON	-0.1	-0.4	0.2
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	0.5	-0.1	1.2

In Table 3 above, the positive mean represents a positive completion rate, indicating an effective treatment program in the state, whereas the negative represents a problematic program. In the total of 52 states, 58% (30 states) of them have a positive completion rate. Furthermore, Florida has one of the highest completion rates, which is 1.0. On the contrary, the completion rates in Virginia is only -2.9.

Conclusion

By conducting the GLMM analysis, we can conclude that the chance for a young person who has alcohol addiction completing the treatment is twice as high as those who have ‘hard’ drugs addiction. The reasoning behind this result is probably treatment for drug addiction is more timeconsuming than alcohol addiction. People who are addicted to drugs have more severe effects and tend to need more attention comparing to those who addicted to alcohol. Hence, the process of healing a drug addiction is more challenging.

The effectiveness of treatment varies depending on which state it’s located, and there are still 42% of the states are running ineffective programs. States like Florida, Delaware, and Massachusetts have the highest completion rates, probably due to their accessible and diversified drugs rehab centers. Whereas Virginia, New Mexico, and North California have the lowest rates, showing problematic and ineffective programs.

Appendix

```r
# Load the dataset
data("MathAchieve", package = "MEMSS"); attach(MathAchieve)
library(nlme)
# Model 1
model1 <- lme(MathAch ~ Minority + Sex + SES, random=~1|School, data=MathAchieve)
qqnorm(model1, main = "Normal Q-Q Plot")
tab<-summary(model1)$tTable[,]
tab = round(tab,2)
knitr::kable(tab, caption = "Estimation of fixed effects in
             linear mixed model of math achievement dataset")
# sigma U = 1.916676, sigma E = 5.992412
# Proportion of variance explained by School
1.916676^2 / (1.916676^2 + 5.992412^2)

# drugs data
library(Matrix)
library(sp)
library(parallel)
library(raster)
library(Pmisc)
library(data.table)
library(Hmisc)

# load the dataset
download.file("http://pbrown.ca/teaching/appliedstats/data/drugs.rds",
              "drugs.rds")
xSub = readRDS("drugs.rds")

table(xSub$SUB1)
table(xSub$STFIPS)[1:5]
table(xSub$TOWN)[1:2]

forInla = na.omit(xSub)
forInla$y = as.numeric(forInla$completed)
library("INLA")
ires = inla(y ~ SUB1 + GENDER + raceEthnicity + homeless + AGE +
              f(STFIPS, hyper=list(prec=list(prior='pc.prec', param=c(0.1, 0.05)))) +
              f(TOWN, hyper=list(prec=list(prior='pc.prec', param=c(0.1, 0.05)))),
            data=forInla, family='binomial',
            control.inla = list(strategy='gaussian', int.strategy='eb'))

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

toPrint = as.data.frame(rbind(exp(ires$summary.fixed[, c(4, 3, 5)]),
                              sdState$summary[, c(4, 3, 5)]))
sss = "^(raceEthnicity|SUB1|GENDER|homeless|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 and quantiles for model parameters.")
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$STFIPS$ID)
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")
```