STA442 Homework 2

Kevin Dang - 1003205079

Due Wednesday October 16, 2019

Question 1: Math

In this study, our primary goal is to investigate whether there are significant differences in mathematics achievement scores between different schools, or whether the differences within schools are just as large as between students attending different schools. To answer this question, we first fit a linear mixed model with mathematics achievement scores (`MathAch`) as the response, `Minority`, `Sex`, `SES`, `MEANSES` as fixed effects and `School` as a random effect. `Minority` represents whether the student is a minority or not, `Sex` indicates male or female, `SES` is a measure of socio-economic status, `MEANSES` is the mean `SES` for each school and `School` is an identifier for each school. The model can be written as follows with b_i as the random effect for school:

$$MathAch = \beta_0 + \beta_1 MinorityYes + \beta_2 SexMale + \beta_3 SES + \beta_4 MEANSES + b_i$$

The model assumptions need to be checked before we can start interpreting the coefficients from Table 1. The Normal QQ Plot in Figure 1 shows that the random effects are normally distributed. Figure 2 represents the 95% prediction intervals for each school and we can see that the intervals are quite spread out and most are not centered around zero, so it is appropriate to include `School` as a random effect in the model.

The first impressions to note are that students who are minorities perform worse than those who are not minorities and males perform better than females on average. Those with higher socio-economic status perform better than those of lower socio-economic status and this also applies to the mean level of socio-economic status. The previous statements apply when all other covariates are held constant; for instance if both a male and female student are minorities and are of the same level of socio-economic status then the model predicts that the male student will achieve a higher score. These results are given by the extremely small p-values corresponding to those coefficients in Table 1 which were all rounded to 0.

The next part involves the discussion of the random effect variable, `School`. Two individuals who are members of a minority racial group and have the same socio-economic status while attending the same school will have a difference of mathematics achievement scores with standard deviation $\sqrt{2}\tau = \sqrt{2} \times 5.992$. The school level effect is $\sigma = 1.563$. Converting them to variances we have $\tau^2 = 35.904$ and $\sigma^2 = 2.443$. The proportions of variances are $\frac{\tau^2}{\sigma^2+\tau^2} = 0.936$ and $\frac{\sigma^2}{\sigma^2+\tau^2} = 0.064$ for within-school variance and between-school variance respectively.

Since the between-school variance represents 6.4% of the total variance, and many 95% prediction intervals for each school do not include 0 we can conclude that there are significant differences in mathematics achievement scores between different schools.

Table 1: Linear Mixed Model for Math Achievement Scores

	MLE	Std.Error	DF	t-value	p-value
(Intercept)	12.830	0.172	7022	74.733	0
MinorityYes	-2.731	0.203	7022	-13.422	0
SexMale	1.218	0.161	7022	7.571	0
SES	1.926	0.108	7022	17.761	0
MEANSES	2.882	0.368	158	7.840	0
σ	1.563	NA	NA	NA	NA
τ	5.992	NA	NA	NA	NA

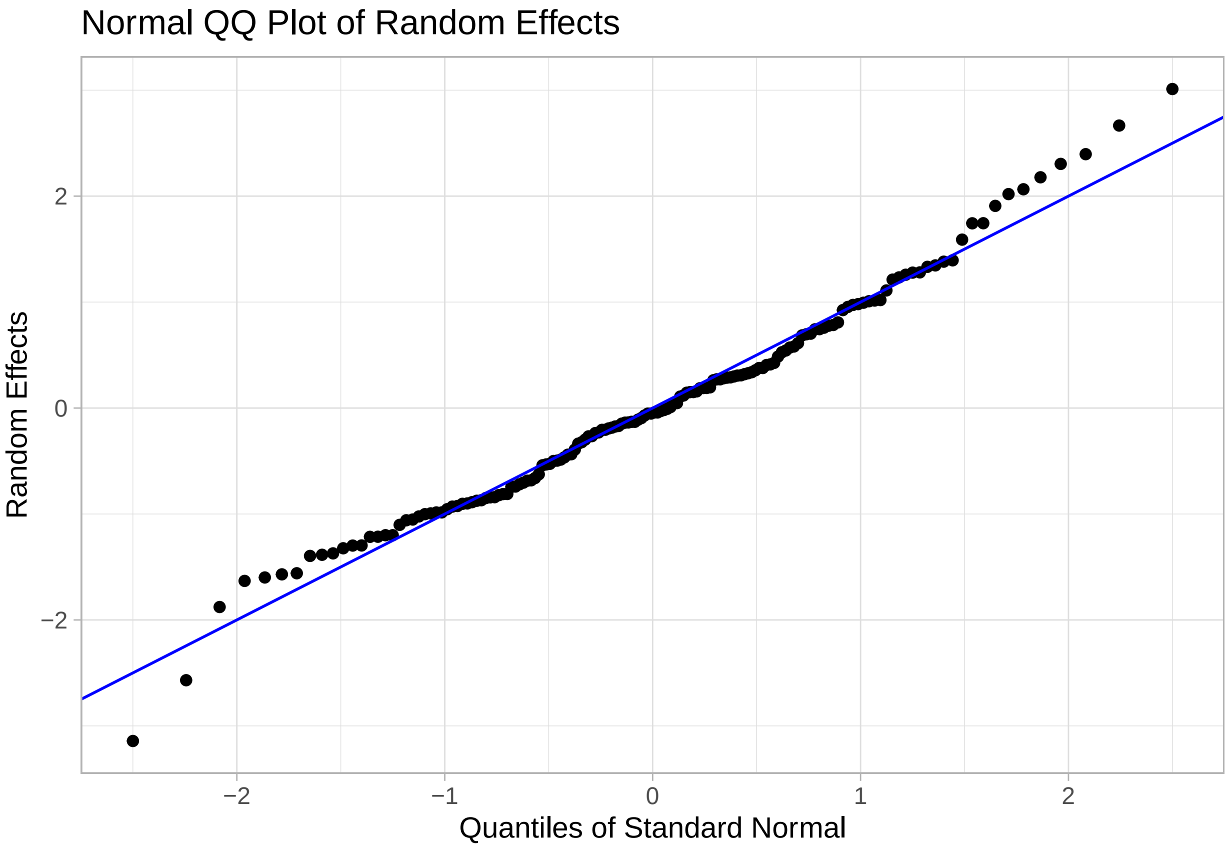

Figure 1: Normal QQ Plot

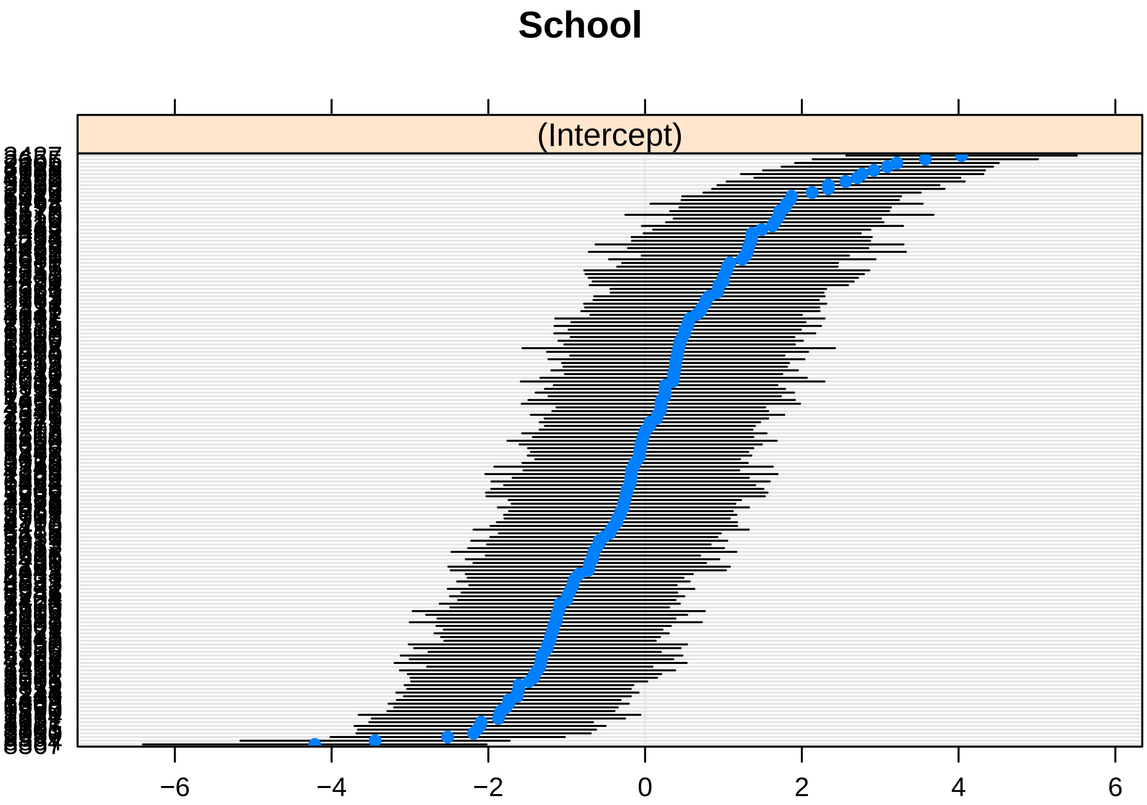

Figure 2: 95% prediction intervals for each school

Appendix A: Math Code

```r
library(tidyverse)
library(nlme)
library(lme4)
library(knitr)

data("MathAchieve", package = "MEMSS")

# Linear Mixed Model with School as a random effect
mathlme <- lme(MathAch ~ Minority + Sex + SES + MEANSES, random = ~1 | School, data = MathAchieve)
kable(Pmisc::lmeTable(mathlme), digits = 3, escape = FALSE,
             caption = "\\label{tab:tab1}Linear Mixed Model for Math Achievement Scores") # Table 1

# Use lmer for random effects plots
mathlmer <- lmer(MathAch ~ Minority + Sex + SES + MEANSES + (1|School), data = MathAchieve)

# Normal QQ Plot
data.frame(randef = ranef(mathlmer)$School[ ,1]) %>%
  mutate_at("randef",funs( (. - mean(.)) / sd(.))) %>%
  arrange(randef) %>%
  mutate(q = qnorm(seq(1:nrow(ranef(mathlmer)$School))/(1 + nrow(ranef(mathlmer)$School)))) %>%
  ggplot(aes(x = q,y = randef)) +
  theme_light() +
  geom_point() +
  geom_abline(slope = 1,intercept = 0,colour = "blue") +
  labs(title = "Normal QQ Plot of Random Intercepts",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles")

# Plot random effects
randef <- ranef(mathlmer, condVar=TRUE)
lattice::dotplot(randef)
```

Question 2: Drugs

Introduction

The Treatment Episode Data Set contains information from treatment facilities that include substance abuse, discharges and patient characteristics. The first hypothesis that we want to investigate is the effect of addiction to hard drugs (i.e. heroin, opiates, methamphetamine, cocaine) versus the effect of addiction to alcohol or marijuana on the chances of a young person completing their treatment program. The second hypothesis is that treatment programs in some American states are more effective than in other states which have low completion rates.

Methods

The analysis involves Bayesian inference. The model that is used is a Binomial Mixed Model with a logit link and a PC prior, which takes the most basic form as follows:

$$
\begin{aligned}
Y_{ij} &\sim Bin(N_i, \mu_i) \\
logit(\mu_i) &= \beta_0 + \beta X_j + U_i \\
U_i &\sim iid \;\; N(0, \sigma^2) \\
\\
\text{Priors:} & \\
\beta &\sim N(0, 10^2 I) \\
\sigma &\sim PC(0.1, 0.05)
\end{aligned}
$$

$$U_i \overset{iid}{\sim} N(0, \sigma^2)$$

Where Y_{ij} is an individual i treated for addiction to substance j. N_i is the number of observations while μ_i is the probability of an individual completing the treatment. X_j is substance type which includes heroin, opiates, methamphetamine, cocaine, alcohol and marijuana. U_i represents the random effect of each US state.

There are some important confounders that need to be included in the model such as `GENDER`, `AGE` and `raceEthnicity`, so we can expand the model below:

$$log(\frac{\mu_i}{1-\mu_i}) = \beta_0 + \beta_1 I(Substance) + \beta_2 Gender + \beta_3 I(Age) + \beta_4 I(raceEthnicity) + U_i$$

where β_1 represents 5 parameters since there are 6 substances in total (with marijuana as the baseline), β_3 represents 3 parameters since there are 4 age groups (with `21-24` as the baseline) and β_4 represents 9 parameters in total since there are 10 ethnic groups in the data (with `WHITE` as the baseline). The written model was simplified with indicator variables for clarity and convenience.

β is given a Gaussian prior which has a variance of 100, which is smaller than the default 1000. This is because the coefficients are expected to be small as we are using a binomial family. The PC prior is used for the scale parameter σ which is the standard deviation of the random effects. The parameters for this prior were suggested by Patrick Brown.

Results

Due to lack of computational resources, the model was not able to run on the full dataset, so a random sample of 100000 observations was taken from the dataset to be used in the model output in Table 2. The model parameter estimates have all been exponentiated for easier interpretation (e.g. β becomes exp(β)).

Table 2: Binomial Mixed Model for Drug Treatment Completion

	mean	sd	0.025quant	0.5quant	0.975quant	mode	kld
(Intercept)	0.70	1.10	0.58	0.70	0.83	0.70	1
SUB1(2) ALCOHOL	1.61	1.02	1.55	1.61	1.67	1.61	1
SUB1(5) HEROIN	0.89	1.02	0.85	0.89	0.93	0.89	1
SUB1(7) OTHER OPIATES AND SYNTHETICS	0.90	1.03	0.86	0.90	0.95	0.90	1
SUB1(10) METHAMPHETAMINE	0.96	1.04	0.89	0.96	1.03	0.96	1
SUB1(3) COCAINE/CRACK	0.90	1.04	0.83	0.90	0.98	0.90	1
GENDER(2) FEMALE	0.88	1.01	0.86	0.88	0.91	0.88	1
AGE18-20	0.93	1.02	0.90	0.93	0.96	0.93	1
AGE15-17	0.94	1.02	0.90	0.94	0.97	0.94	1
AGE12-14	0.97	1.04	0.91	0.97	1.04	0.97	1
raceEthnicityHispanic	0.86	1.02	0.83	0.86	0.89	0.86	1
raceEthnicityBlack	0.69	1.02	0.66	0.69	0.71	0.69	1
raceEthnicityAmerican Indian	0.75	1.06	0.67	0.75	0.85	0.75	1
raceEthnicityOther	0.81	1.06	0.73	0.81	0.90	0.81	1
raceEthnicityTwo or more	0.80	1.07	0.71	0.80	0.91	0.80	1
raceEthnicityASIAN	1.09	1.08	0.94	1.09	1.27	1.09	1
raceEthnicityHawaiian/Other Pacific	0.79	1.11	0.64	0.79	0.97	0.79	1
raceEthnicityAsian/Pacific Islander	1.63	1.16	1.23	1.63	2.17	1.63	1
raceEthnicityAlaska Native	0.96	1.29	0.59	0.96	1.58	0.96	1

From the output we can see that holding all other variables constant, the odds of an alcohol user completing their treatment is $100 \times (1.61 - 1) = 61\%$ greater than a marijuana user. The odds (relative to marijuana) for heroin is $100 \times (0.89 - 1) = -11\%$, opiates is -10%, methamphetamine is -4% and cocain is -10%. Three out of the four hard substances groups have odds under 1 for their upper bound of the 95% interval, so users of these hard drugs have a lower probability of completing their treatment compared to alcohol and marijuana users. The one exception is methamphetamine users, where the upper bound of the 95% interval is 1.03, so this group cannot be ruled out in having a similar likelihood of completing their treatment as marijuana users if they have very similar characteristics.

Next we look at 95% credible intervals for various American. states in Table 3. Quite a few of the credible intervals do not include 0, so it appears as though that there is a difference in completion of treatment programs across different states. For example, Florida has a 95% credible interval of (1.1, 1.5) whereas Virginia has an interval of (-3.0, -2.3) so there is a significant difference between these two states.

Table 3: Random Intercepts with 95% credible intervals

ID	mean	0.025q	0.975q	ID	mean	0.025q	0.975q
(1) ALABAMA	0.2	-0.3	0.7	(30) MONTANA	-0.2	-0.5	0.2
(2) ALASKA	0.1	-0.3	0.4	(31) NEBRASKA	0.0	-0.3	0.3
(4) ARIZONA	0.0	-1.2	1.2	(32) NEVADA	-0.1	-0.3	0.1
(5) ARKANSAS	0.0	-0.5	0.4	(33) NEW HAMPSHIRE	0.2	-0.1	0.5
(6) CALIFORNIA	-0.2	-0.3	0.0	(34) NEW JERSEY	0.6	0.4	0.7
(8) COLORADO	0.8	0.7	1.0	(35) NEW MEXICO	-2.0	-2.5	-1.5
(9) CONNECTICUT	0.3	0.2	0.5	(36) NEW YORK	-0.1	-0.3	0.1
(10) DELAWARE	0.5	0.3	0.8	(37) NORTH CAROLINA	-0.7	-0.9	-0.5
(11) DISTRICT OF COLUMBIA	-0.5	-0.8	-0.2	(38) NORTH DAKOTA	0.0	-0.7	0.7
(12) FLORIDA	1.3	1.1	1.5	(39) OHIO	-0.1	-0.3	0.1
(13) GEORGIA	-0.4	-0.8	0.0	(40) OKLAHOMA	0.4	0.1	0.6
(15) HAWAII	0.4	0.2	0.7	(41) OREGON	0.1	0.0	0.3
(16) IDAHO	-0.5	-0.8	-0.1	(42) PENNSYLVANIA	0.0	-1.2	1.2

ID	mean	0.025q	0.975q	ID	mean	0.025q	0.975q
(17) ILLINOIS	-0.1	-0.3	0.1	(44) RHODE ISLAND	0.2	0.0	0.4
(18) INDIANA	-0.1	-0.7	0.5	(45) SOUTH CAROLINA	0.2	0.0	0.4
(19) IOWA	0.5	0.3	0.7	(46) SOUTH DAKOTA	1.0	0.7	1.2
(20) KANSAS	-0.4	-0.6	-0.1	(47) TENNESSEE	0.2	0.0	0.4
(21) KENTUCKY	-0.7	-1.0	-0.5	(48) TEXAS	0.6	0.5	0.8
(22) LOUISIANA	-0.6	-0.8	-0.4	(49) UTAH	0.3	0.1	0.5
(23) MAINE	0.2	-0.1	0.5	(50) VERMONT	-0.3	-0.6	0.0
(24) MARYLAND	0.4	0.2	0.6	(51) VIRGINIA	-2.7	-3.0	-2.3
(25) MASSACHUSETTS	0.7	0.6	0.9	(53) WASHINGTON	-0.1	-0.3	0.1
(26) MICHIGAN	-0.5	-0.7	-0.3	(54) WEST VIRGINIA	0.0	-1.2	1.2
(27) MINNESOTA	0.5	0.3	0.7	(55) WISCONSIN	0.0	-1.2	1.2
(28) MISSISSIPPI	0.0	-1.2	1.2	(56) WYOMING	0.0	-1.2	1.2
(29) MISSOURI	-0.3	-0.5	-0.1	(72) PUERTO RICO	0.7	0.1	1.3

Conclusions

Our analysis supports the two hypotheses described in the introduction. The likelihood of a young person completing their drug treatment program depends on the substance that they are addicted to, with alcohol or marijuana being easier to treat compared to hard drugs such as heroin, opiates, methamphetamine and cocaine. An interesting result is that methamphetamine users have nearly the same chance of completing their treatment as marijuana users, provided that their demographic characteristics are the same. The analysis also showed that some American states have effective treatment programs whereas other states have programs with very low completion rates.

Appendix B: Drugs Code

```r
library(INLA)
library(plyr)

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

df = na.omit(xSub)
df$y = as.numeric(df$completed)

# Rename raceEthniticity levels to save space
df$raceEthnicity <- revalue(df$raceEthnicity,
                            c("BLACK OR AFRICAN AMERICAN"="Black",
                              "AMERICAN INDIAN (OTHER THAN ALASKA NATIVE)"="American Indian",
                              "OTHER SINGLE RACE"="Other","TWO OR MORE RACES"="Two or more",
                              "NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER"="Hawaiian/Other Pacific",
                              "ASIAN OR PACIFIC ISLANDER"="Asian/Pacific Islander",
                              "ALASKA NATIVE (ALEUT, ESKIMO, INDIAN)"="Alaska Native"))

# Randomly sample to get subset of data, not enough computation power
set.seed(1998)
sample <- sample_n(df,100000)

# Binomial Mixed Model
ires = inla(y ~ SUB1 + GENDER + AGE + raceEthnicity +
              f(STFIPS, hyper=list(prec=list(prior='pc.prec',param=c(0.1, 0.05)))),
            data=sample, family='binomial',
            control.fixed = list(mean = 0, mean.intercept = 0, prec = 10^(-2),
                                 prec.intercept = 10^(-2)),
            control.family = list(link = "logit"),
            control.inla = list(strategy='gaussian', int.strategy='eb'))
kable(exp(ires$summary.fixed), digits=2,
      caption = "\\label{tab:tab2}Binomial Mixed Model for Drug Treatment Completion")

# 95% credible intervals for the Random Intercepts
states = cbind(ires$summary.random$STFIPS[1:26, c(1, 2, 4, 6)],
                ires$summary.random$STFIPS[-(1:26), c(1, 2, 4, 6)])
colnames(states) = gsub("uant", "", colnames(states))
kable(states, digits = 1,
      caption = "\\label{tab:tab3} Random Intercepts with 95% credible intervals")
```