Does school affect Math Achievement?

Xinqi Shen

1 Introduction

The MathAchieve dataset is from the MEMSS package, and this dataset provides us with information about math achievement scores, socio-economic status, school of students and whether the student is a member of a minority group. In this report, we will discuss if there are substantial differences between schools, or are differences within schools nearly as big as differences between students from different schools.

2 Method

We fit a linear mixed model to analyze the math achievement scores, treating school as a random effect and other factors as fixed effects. Since students in the same school are not independent and we want to analyze the difference between schools and within schools. Given the following model:

$$MathAchieve_{ij}|School_i \sim N(\mu_{ij}, \tau^2)$$
$$\mu_{ij} = X_{ij}\beta + School_i$$
$$School_i \sim N(0, \sigma^2)$$

where $MathAchieve_{ij}$ is a math achievement score for jth student in ith school. $X_{ij}\beta$ contains intercept, gender, socio-economic status and whether a student is a member of a minority racial group. $School_i$ is the random effect term.

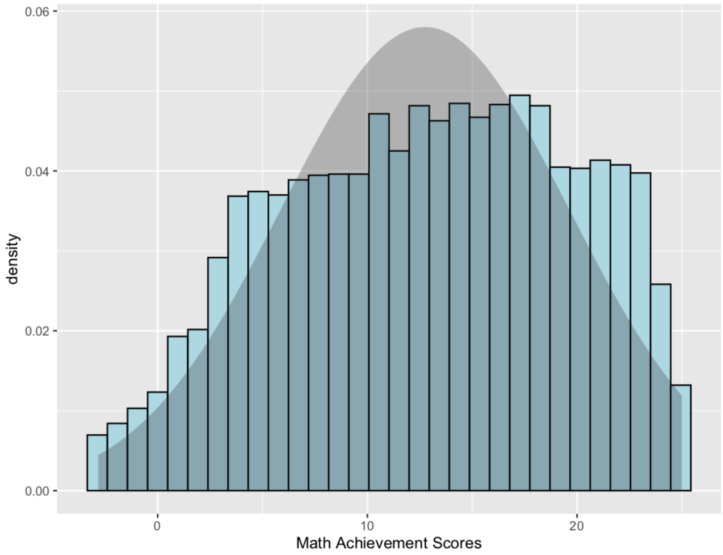


Figure 1: The histogram of normal empirical distribution

From the histogram shown above, the math achievement score approximately follows a normal distribution, which satisfies our model assumption.

3 Result

	MLE	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
σ	1.92	NA	NA	NA	NA
τ	5.99	NA	NA	NA	NA

Table 1: Estimated parameters

Based on the summary table shown above, fixed effects are all significant given their p values less than 0.05, which indicates the fixed effect supports the correctness of our model. In order to analyze the difference between schools and within schools, we can look at the variances between schools σ^2 and the variance within schools τ^2. We find that the variance between schools is about 10.2% of variance within schools. In other words, the proportion of variance due to school(random effect) is about 9.3%$(\frac{\sigma^2}{\sigma^2+\tau^2})$. Thus, we can conclude that there are no significant difference in math achievement scores between schools.

4 Conclusion

Based on the analysis of the result of the dataset from MEMSS package, the differences in math achievement score between schools are much smaller than the differences within schools. In other words, there are no substantial differences between schools.

5 Appendix

```
library("nlme")
library(xtable)
library(ggplot2)
sd=sd(MathAchieve$MathAch)
mean=mean(MathAchieve$MathAch)
ggplot(MathAchieve, aes(x=MathAch)) +
            geom_histogram(aes(y=..density..),color="black",fill="lightblue")
            + stat_function(fun=dnorm,args=list(mean=mean,sd=sd),n=1000,
            geom="ribbon",alpha=0.3,mapping = aes(ymin=0,ymax=..y..))
            + xlab("Math Achievement Scores")
data("MathAchieve", package = "MEMSS")
model1 <- lme(MathAch ~ Minority + Sex + SES, random = ~1 | School,
              data=MathAchieve)
knitr::kable(Pmisc::lmeTable(model1), digits = 2, escape = FALSE,
             format = "latex")
```

American drug treatment completion

Xinqi Shen

1 Introduction

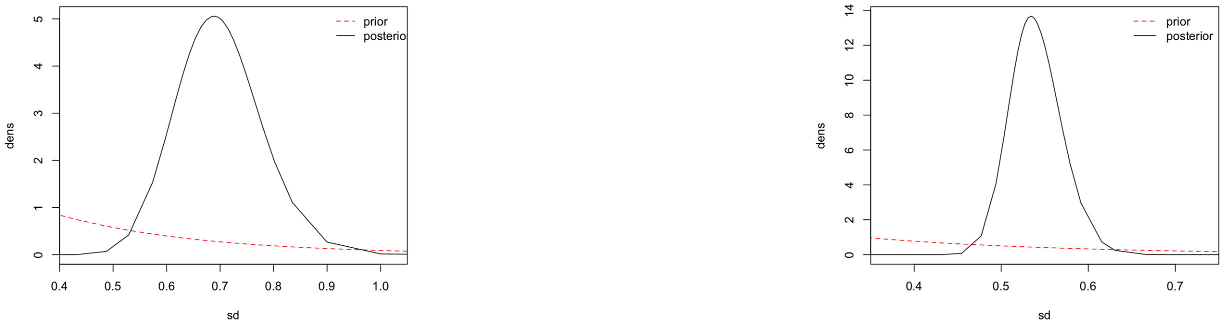
The Treatment Episode Data Set – Discharges (TEDS-D) is a national census data system of annual discharges from substance abuse treatment facilities.TEDS-D provides annual data on the number and characteristics of persons discharged from public and private substance abuse treatment programs that receive public funding. We will focus on addressing two hypotheses. First of all, whether the chance of a young person completing their drug treatment depends on the substance. Secondly, whether some American states have particularly effective treatment programs.

2 Method

We will fit a logistic generalized linear mixed model since the response variable 'completed' is either TRUE or FALSE. And we will treat state and town as random effects because people living in the same place more likely have similar effects of drug treatment. Also, we include gender, age, race, whether a person is homeless and the substance which is the individual's primary addiction as fixed effects. Given the following model:

$$Completed_{ijk}|States_i, Town_{ij} \sim Binomial(N, \mu_{ijk})$$
$$log(\frac{\mu_{ijk}}{1-\mu_{ijk}}) = X_{ijk}\beta + State_i + Town_{ij}$$
$$State_i \sim N(0, \sigma_1^2)$$
$$Town_{ij} \sim N(0, \sigma_2^2)$$

where $Completed_{ijk}$ is if the ith person who lives in the jth town of ith state completed their treatment. $X_{ijk}\beta$ contains an intercept, and all the fixed effects which include gender, age, race, substance and whether a person is homeless. $State_i$ and $Town_{ij}$ are both random effects.



(a) Posterior and Prior for state

(b) Posterior and Prior for town

Figure 1: Posterior and Prior distribution

We use penalized complexity prior for our σ_1 and σ_2 with $P(\sigma_1 > 0.8) = 0.05$ and $P(\sigma_2 > 0.7) = 0.05$. Based on the posterior and prior distribution plots shown above, the choice of prior looks appropriate.

3 Result

variable	0.5quant	0.025quant	0.975quant
INTERCEPT			
(INTERCEPT)	0.716	0.573	0.895
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	0.916	0.994
COCAINE/CRACK	0.855	0.814	0.898
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	0.728	0.679	0.781
OTHER SINGLE RACE	0.866	0.812	0.923
TWO OR MORE RACES	0.855	0.794	0.921
ASIAN	1.132	1.038	1.235
NATIVE HAWAIIAN OR OTHER	0.844	0.748	0.953
ASIAN OR PACIFIC ISLANDER	1.454	1.227	1.723
ALASKA NATIVE	0.845	0.624	1.145
HOMELESS			
TRUE	1.005	0.973	1.037
AGE			
AGE18-20	0.935	0.916	0.953
AGE15-17	0.926	0.905	0.947
AGE12-14	0.972	0.934	1.012
SD			
STFIPS	0.693	0.554	0.865
TOWN	0.537	0.484	0.600

Table 1: Posterior means and quantiles for model parameters.

The table 1 provides the posterior means and quantiles including random effects and fixed effects for our model parameters. By looking at the 95% posterior credible interval of odds ratio whether includes 1, we find 'homeless' is an insignificant fixed effect, thus we can ignore this variable.

3.1 Substance

First of all, we can figure out whether the chance of a young person completing their drug treatment depends on the substance the individual is addicted to, with 'hard' drugs (Heroin, Opiates, Methamphetamine, Cocaine) being more difficult to treat than alcohol or marijuana. By focusing on the outputs in SUB1 part, given marijuana as the reference group, we find that only the odds ratio of alcohol is grater than 1 and other substances(Heroin, Opiates, Methamphetamine, Cocaine) are all smaller than 1. In other words, only people who are addicted to alcohol are more likely to finish their drug treatment than people who are addicted to marijuana. Thus, we can conclude that people who are addicted to the 'hard' drugs are more difficult to complete their drug treatments than alcohol or marijuana.

3.2 State

ID	mean	0.025quant	0.975quant
ALABAMA	0.2	-0.3	0.8
ALASKA	0.0	-0.9	0.9
ARIZONA	0.0	-1.4	1.4
ARKANSAS	-0.1	-0.7	0.5
CALIFORNIA	-0.3	-0.6	0.0
COLORADO	0.5	0.1	1.0
CONNECTICUT	0.1	-0.4	0.7
DELAWARE	1.0	0.7	1.3
FLORIDA	1.0	0.7	1.4
WASHINGTON DC	-0.3	-0.6	0.1
VIRGINIA	-2.9	-3.3	-2.6

Table 2: Random Effect of Some American States.

Secondly, we will address the hypothesis that whether some American states have particularly effective treatment programs whereas other states have programs which are highly problematic with very low completion rates. Based on 95% posterior credible interval of SD in Table 1, we find the standard deviation for state(STFIPS) do not contain 0, which indicates state as a random effect is significant. In other words, the completion rate of drug treatment differ from state to state. Table 2 demonstrates more detailed information about random effect of some American states(Full table is in Appendix). The 95% CI of some states include 0 such as Alabama and Alaska, which implies the drug treatment in these states does not obviously affect the completion rate with small mean value. Conversely, for some states do not include 0 in their 95% CI, we can compare their completion rate by mean value. For example, both Delaware and Colorado have particularly effective treatment programs given their large positive mean value 1.0 and 0.5, and it also shows Delaware performs better than Colorado. However, Virginia has highly problematic with very low completion rates, given its large negative mean value -2.9.

4 Conclusion

Based on the analysis of the result of the dataset from TEDS-D, we found some interesting results. Firstly, young person who is addicted to 'hard' drugs such as Heroin, Opiates, Methamphetamine, Cocaine are difficult to complete their treatment, compared with person who is addicted to alcohol or marijuana. Thus, young people should try not to get in touch with 'hard' drugs. Not surprisingly, some American states have particularly effective treatment programs such as Delaware and Florida, compared with some states are highly problematic with very low completion rates such as Virginia. Thus, young people who are addicted to drugs can seek help in these states.

5 Appendix

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

forInla = na.omit(xSub)
forInla$y = as.numeric(forInla$completed)
library("INLA")
library("xtable")

ires = inla(y ~ SUB1 + GENDER + raceEthnicity + homeless +
            f(STFIPS, hyper=list(prec=list(
              prior='pc.prec', param=c(0.8, 0.05)))) +
              f(TOWN,hyper=list(prec=list(
                prior='pc.prec', param=c(0.7, 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$legend)
do.call(matplot, sdState$TOWN$matplot)
do.call(legend, sdState$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)
print(xtable(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")
```

ID	mean	0.025q	0.975q	ID	mean	0.025q	0.975q
ALABAMA	0.2	-0.3	0.8	MONTANA	-0.2	-1.0	0.7
ALASKA	0.0	-0.9	0.9	NEBRASKA	0.8	0.4	1.2
ARIZONA	0.0	-1.4	1.4	NEVADA	-0.1	-0.8	0.6
ARKANSAS	-0.1	-0.7	0.5	NEW HAMPSHIRE	0.2	-0.3	0.7
CALIFORNIA	-0.3	-0.6	0.0	NEW JERSEY	0.5	0.2	0.8
COLORADO	0.5	0.1	1.0	NEW MEXICO	-1.2	-2.0	-0.5
CONNECTICUT	0.1	-0.4	0.7	NEW YORK	-0.3	-0.6	0.0
DELAWARE	1.0	0.7	1.3	NORTH CAROLINA	-0.8	-1.2	-0.5
WASHINGTON DC	-0.3	-0.6	0.1	NORTH DAKOTA	-0.3	-1.0	0.4
FLORIDA	1.0	0.7	1.4	OHIO	-0.2	-0.6	0.1
GEORGIA	-0.2	-0.8	0.4	OKLAHOMA	0.6	0.0	1.1
HAWAII	0.2	-0.6	1.1	OREGON	0.1	-0.3	0.5
IDAHO	-0.2	-1.1	0.7	PENNSYLVANIA	0.0	-1.4	1.4
ILLINOIS	-0.5	-0.8	-0.2	RHODE ISLAND	-0.2	-0.6	0.3
INDIANA	-0.1	-0.9	0.8	SOUTH CAROLINA	0.4	0.0	0.7
IOWA	0.4	0.1	0.7	SOUTH DAKOTA	0.5	-0.3	1.4
KANSAS	-0.2	-0.6	0.1	TENNESSEE	0.3	-0.2	0.7
KENTUCKY	-0.2	-0.5	0.2	TEXAS	0.6	0.3	0.9
LOUISIANA	-0.6	-1.0	-0.1	UTAH	0.1	-0.5	0.7
MAINE	0.1	-0.7	1.0	VERMONT	-0.2	-1.1	0.6
MARYLAND	0.5	0.2	0.8	VIRGINIA	-2.9	-3.3	-2.6
MASSACHUSETTS	0.8	0.4	1.3	WASHINGTON	-0.1	-0.5	0.3
MICHIGAN	-0.4	-0.7	0.0	WEST VIRGINIA	0.0	-1.4	1.4
MINNESOTA	0.4	0.0	0.9	WISCONSIN	0.0	-1.4	1.4
MISSISSIPPI	0.0	-1.4	1.4	WYOMING	0.0	-1.4	1.4
MISSOURI	-0.4	-0.7	-0.1	PUERTO RICO	0.6	-0.1	1.3

Table 3: Random Effect of American States