STA442 A2

Tuoyue Huang Student number:1003906712

Saturday 15 October 2019

"Report of Math Dataset"

Summary

Through analysis of data set "MathAchieve" by fitting a mixed effect model with school as the random effect, we could arrive the conclusion that the differences of math score is mostly explained by within school variation instead of between schools variation.

Introduction

In the report, we are interested in comparisons of math score differences between schools or within schools and differences between students from different schools? The data set use here is "MathAchieve" from package "MEMSS".

Methods

Our purpose of study are differences between and within groups, which indicates we should carry out a random effects model and treat School as a random effect. In addition, we have data which has multiple observations from the same school, so it is reasonable to treat School as a random effect. We can also verify our guess through Figure 1, the red point is the median of math score which is randomly distributed from 5 to 20. The model assuption here we have:

$$Y_{ij}|U_i \overset{ind}{\sim} N(\mu_{ij}, \tau^2) \quad \mu_{ij} = X_{ij}\beta + U_i \quad U_i \overset{ind}{\sim} N(0, \sigma^2)$$

where Y_{ij} is the math score for $Student_j$ in $School_i$ and $X_{ij}\beta$ is the fixed effect of covariates, like sex. U_i here is the individual random effect from School, which is saying each school's deviation from the population average math score. Since we have the normality assumption of School, we check the normal QQ plot which is Figure 2. All the points located around the diagnol line except the two bottom ponits, which is overall pretty good. In order to see the differences, we need to calculate the proportions of between and within subject variance, which we can get from the summary table of the model fitted. We use restricted maximunm likelihood to get unbiased variance of parameter.

Results

Figure 1: Boxplot of Math Score Among Schools

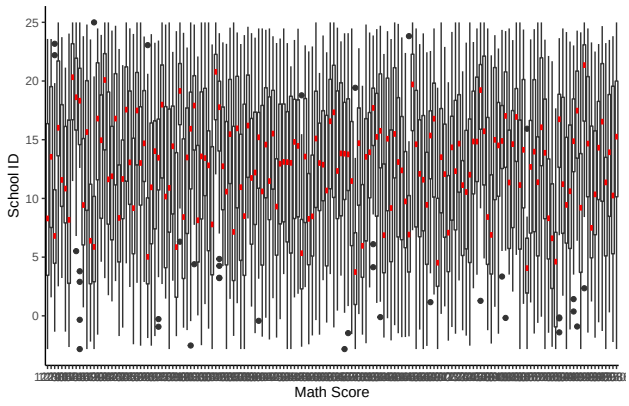


Figure 2: Normal QQ Plot, Predicted Random Intercepts

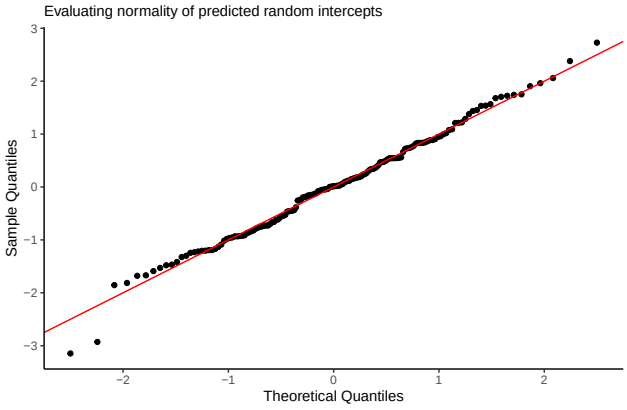


Table 1 is our summary table of the model. Minority, Sex and SES are all fixed effect variables, whcih we can interpret as usuall. For example, if the student is from minority racial group, then his math score is likely to be about 3 marks lower. The corresponding p-values are all 0, resulting we have significant evidence to conclude that. The last two rows gives us between and within school standard deviation,whcih we could use to calculate the following: Between School Variance $= (1.92^2/(1.92^2 + 5.99^2) * 100 = 9\%$; Within School Variance $= (5.99^2/(1.92^2 + 5.99^2) * 100) = 91\%$ This is saying different schools have similar math scores since between school variance is only 9%, while 91% of variability in math score is attributed to within schools which is students. Therefore, we could claim that there are not substantial differences between schools and 91% part of the difference is explained by the within school variation.

Table 1: Mixed Effect Model of Math Score

	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

"Report of Treatment Episode Data Set Discharges(TEDSD)"

Introduction

Abuse consumption of alcohol and drugs is certainly something we should need to take seriously and get a treatment with. But how is the effectiveness of this type of treatment. Is it depend on the kinds of addiction substances or it is affected by the quality of treatment programs of different states? To answer these two questions, we carry out an analysis on Treatment Episode Data Set Discharges(TEDSD). The TEDS-D data set 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.

Methods

In the analysis, we use the generalized linear mixed model with bayesian inference. The reason is that our response variable,'Completed', is a bernoulli variable, thus we should use logistic regression. In addition, we need to set covariates 'STFIPS', which is states, and 'TOWN' as random effects, since multiple measurements are come from the same state and the same town. We suppose patients from different towns and states may have different global probability of success of treatment. From using bayesian inference, we could get a posterior distribution of our interested parameters instead of point estimation. Posterior distributions is just the estimated distribution of parameters given the prior information we provided. Thus, we have our model with the following assuptions:

$$Y_{ijk} \overset{ind}{\sim} Bernoulli(p_k) \quad logit(p_k) = \mu + X_{ijk}\beta + U_i + V_j \quad U_i \overset{i.i.d.}{\sim} N(0, \sigma_1^2) \quad V_j \overset{i.i.d.}{\sim} N(0, \sigma_2^2)$$

Priors: $\mu \sim N(0, 10^2) \quad \beta \sim N(0, 3^2) \quad \sigma_1 \sim Exponential(\sigma_1^2) \quad \sigma_2 \sim Exponential(\sigma_2^2)$

where Y_{ijk} is response variable of $patient_k$ living in $Town_j$ of $state_i$,and X_{ijk} is the fixed effects of covariates, like gender and race. Here U_i and V_j are Latent variables for random effects due to states and towns.

We first should exponentiate the summary table due to logistic model, then we get the odds of success in drug addiciton treatment. In the summary table, we will have posterior means and marginal 95% credible intervals for parameters. We need to know odds and probability transformation is monotonic, then a higher odds means a relatively higher probability of success in treatment. Therefore, we could get our conclusions by comparing the 0.5 quantile, which is aproximately equals to the mean due to the normality assumption, of different groups in Table 2.

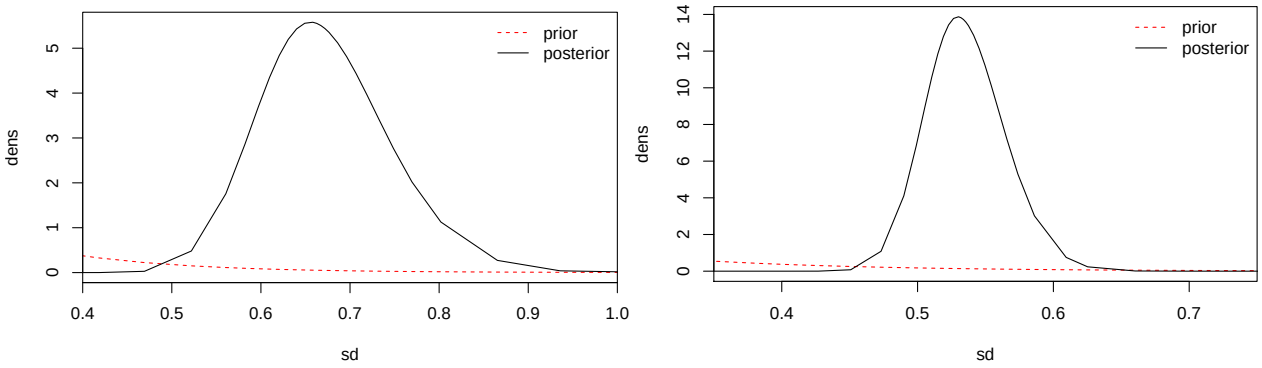
Table 2: Posterior means and quantiles for model parameters.

	0.5quant	0.025quant	0.975quant
(Intercept)			
(Intercept)	0.681	0.550	0.843
SUB1			
ALCOHOL	1.642	1.608	1.677
HEROIN	0.898	0.875	0.921
OTHER OPIATES AND SYNTHET	0.924	0.897	0.952
METHAMPHETAMINE	0.982	0.944	1.022
COCAINE/CRACK	0.876	0.834	0.920
GENDER			
FEMALE	0.895	0.880	0.910
raceEthnicity			
Hispanic	0.829	0.810	0.849
BLACK OR AFRICAN AMERICAN	0.685	0.669	0.702
AMERICAN INDIAN (OTHER TH	0.730	0.680	0.782
OTHER SINGLE RACE	0.864	0.810	0.921
TWO OR MORE RACES	0.851	0.790	0.917
ASIAN	1.133	1.038	1.236
NATIVE HAWAIIAN OR OTHER	0.847	0.750	0.955
ASIAN OR PACIFIC ISLANDER	1.451	1.224	1.719
ALASKA NATIVE (ALEUT, ESK	0.845	0.623	1.144
homeless			
TRUE	1.015	0.983	1.048
SD			
STFIPS	0.667	0.542	0.831
TOWN	0.534	0.482	0.597

Results

To answer our first hypothesis, we look at the first part of Table 2. Our base line group, the intercept, is the odds of success in treatment for white male patients with marijuana addiction and is not homeless. Since we are only interested in the effect of addictive drug, we should focus on section of 'SUB1'. Our reference group is marijuana with odds $e^0 = 1$. Then we can compare the posterior mean for each drug with value 1. We can see the group alcohol has odds of 1.64, which is saying the odds increases by 64% compare with the marijuana. Thus, people with alcohol addition have a higher probability of success in the addictive drug treatment. The other following groups are all having odds ratio less than one, infering that 'hard' drugs, like heroin, opiates, are most likely to have a lower probability of successful treatment than marijuana. Among 'hard' drugs, cocaine is the most difficult one to treat with a 13% less odds ratio than the reference group. By checking the 95% credible interval, each of them has reasonable small range, and we can conclude we have significant evidence to say that alcohol and marijuana is more easier to treat than other 'hard' drugs.

ID	mean	0.025q	0.975q	ID	mean	0.025q	0.975q
ALABAMA	0.2	-0.3	0.8	MONTANA	-0.2	-1.0	0.6
ALASKA	0.0	-0.8	0.8	NEBRASKA	0.8	0.4	1.2
ARIZONA	0.0	-1.3	1.3	NEVADA	-0.1	-0.8	0.5
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	-1.9	-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	-1.1	-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	1.0	OREGON	0.1	-0.3	0.5
IDAHO	-0.2	-1.0	0.6	PENNSYLVANIA	0.0	-1.3	1.3
ILLINOIS	-0.5	-0.8	-0.2	RHODE ISLAND	-0.2	-0.6	0.2
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	1.3
KANSAS	-0.2	-0.6	0.1	TENNESSEE	0.3	-0.2	0.7
KENTUCKY	-0.1	-0.5	0.2	TEXAS	0.6	0.3	0.9
LOUISIANA	-0.5	-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	-2.5
MASSACHUSETTS	0.8	0.4	1.2	WASHINGTON	-0.1	-0.5	0.3
MICHIGAN	-0.4	-0.7	0.0	WEST VIRGINIA	0.0	-1.3	1.3
MINNESOTA	0.4	0.0	0.9	WISCONSIN	0.0	-1.3	1.3
MISSISSIPPI	0.0	-1.3	1.3	WYOMING	0.0	-1.3	1.3
MISSOURI	-0.4	-0.7	-0.1	PUERTO RICO	0.6	-0.1	1.3



For the second hypothesis, we look at the second half of Table 2. These are all the marginal posterior means and 95% credible intervals for the random effects U_i of 52 American states. States with positive means are relatively have more effective local treatment pragrams and vice versa. The reason is that we add U_i on the right side of our model, a positive U_i will eventually lead to a higher value of p. Majority of the states have posterior mean around 0, indicating their local treatment programs have very little or no influence to the probability of successful treatments. Let us look at some significant states, for example states Delaware(1.0), Massachusetts(0.8), Nebraska(0.8), Texas(0.6) and some others get mean above or eaqual to 0.5 have relatively higher means, meaning thier local treatment may be more effective compare to the rest states. However on the other sides, in states Virginia(-2.9),New Mexico(-1.1), North Carolina(-0.8), Louisiana(-0.5) are getting relatively lower and negative means, thus their programs tends to have lower probability of success due to problematic design. By checking the 95% credible intervals, those significant ones do not include 0. Therefore, there are enough evidence to say that completion rates are different between states. Delaware is the best and Virginia is the worst state in drug treatment.

The two plots above(I tried to add a title to these two plots, but failed) show the priors we choose for state

and town and the posterior distributions r has estimated. Since I have no idea about the probability of successful drug addictive treatment would be, I just use penalized complexity priors for both random variable. I randomly guess there's a 40% chance that the between subject variability is > 0.05.

Conclusion

In concequence, through analysis of TEDSD data set, we learned drug treatment completion rates of young people is affected by addicted substance and in what states are these people receving the treatment. We can claim that alcohol has the highest probability of completing, following by marijuana, and Heroin, Opiates, Methamphetamine, Cocaine have a lower probability, where Cocaine is the lowest and is 13% less in odds ratio compare to marijuana. We also conclude, programs in Delaware, Massachusetts and Nebraska have relatively higher chance to complete, while programs may be problematic in Virginia, New Mexico and North Carolina since completion rates are lower than the rest states.

Appendix(code used)

```
#####FIRST
data("MathAchieve", package = "MEMSS")
math <- as_data_frame(MathAchieve)

plot1 <- ggplot(data = math, aes(math$School,math$MathAch))+
  geom_boxplot() +
  theme_classic()
boxplotdat <-ggplot_build(plot1)$data[[1]]
plot1 +
  geom_segment(data = boxplotdat ,
               aes(x=xmin, xend=xmax, y=middle, yend=middle),
               colour="red", size=2) +
   labs(title= "Figure 1: Boxplot of Math Score Among Schools",x="Math Score",y="School ID") +
  theme(plot.title = element_text(hjust = 0.5,face = "bold"))

#Same model for plot QQ plot
mod1 = lme4::lmer(MathAch ~ SES + Minority + Sex + (1|School), REML = TRUE,
                  data = MathAchieve)
#Same model for summary table
mod2 = nlme::lme(MathAch ~ Minority + Sex + SES, random = ~1|School,
                 data=MathAchieve)


data_frame(b = ranef(mod1)$School[ ,1]) %>%
  mutate_at("b",funs( (. - mean(.)) / sd(.))) %>%
  arrange(b) %>%
  mutate(q = qnorm(seq(1:nrow(ranef(mod1)$School))/(1 + nrow(ranef(mod1)$School)))) %>%
  ggplot(aes(x = q,y = b)) +
  theme_classic() +
  geom_point() +
  geom_abline(slope = 1,intercept = 0,colour = "red") +
  labs(title = "Figure 2: Normal QQ Plot, Predicted Random Intercepts",
       subtitle = "Evaluating normality of predicted random intercepts",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles") +
  theme(plot.title = element_text(hjust = 0.5,face = "bold"))
```

```r
knitr::kable(Pmisc::lmeTable(mod2), digits = 2, escape = FALSE,
             caption = "Mixed Effect Model of Math Score")

#####SECOND
download.file("http://pbrown.ca/teaching/appliedstats/data/drugs.rds", "drugs.rds")
xSub = readRDS("drugs.rds")
#table(xSub$SUB1,xSub$completed)
forInla = na.omit(xSub)
forInla$y = as.numeric(forInla$completed)

ires = inla(y ~ SUB1 + GENDER + raceEthnicity + homeless
            + f(STFIPS, model = "iid",
                hyper=list(prec=list( prior='pc.prec', param=c(0.4, 0.05))))
            + f(TOWN, model = "iid",
                hyper=list(prec=list( prior='pc.prec', param=c(0.4, 0.05)))),
            data=forInla, family='binomial',
            control.fixed = list(
              mean = 0, mean.intercept = 0,
              prec = 3^(-2), prec.intercept = 100^(-2)),
            control.family = list(
              link = 'logit'),
            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)

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

colnames(toprint) = gsub("uant", "", colnames(toprint))
knitr::kable(toprint, digits = 1, format = "latex")
#kable_styling(latex_options = "hold_position")
```