STA442_HW2

Zeyang Zhang

16/10/2019

Math Report

Introduction

We analyse the dataset MathAchieve from MEMSS package. "School" is an ordered factor identifying the school that the student attends. "Minority" is a factor with levels **No** and **Yes** indicating if the student is a member of a minority racial group. "Sex" is a factor with levels **Male** and **Female**. "SES" is a numeric vector of socio-economic status. "MathAch" is a numeric vector of mathematics achievement scores. "MEANSES" is a numeric vector of the mean SES for the school. We hope to see if there are substantial differences between schools by comparing the differences within schools and the differences between students from different schools.

Methods

We set "MathAch" as our response variable, fit a mixed model that treats factor "Minority", factor "Sex" and the variable "SES" as fixed effects and treats School as a random effect. The model can be written as:

$$
\begin{aligned}
&Y_{ij}|U_i \overset{ind}{\sim} N(\mu_{ij}, \tau^2) \\
&\mu_{ij} = \beta_o + \beta_1 I_{Minority} + \beta_2 I_{Sex} + \beta_3 X_{SES} + U_i \\
&U_i \overset{ind}{\sim} N(0, \sigma^2)
\end{aligned}
$$

Y_{ij} is the response variable, representing the mathematics achievement scores of the jth student in ith School.

$I_{Minority}$ is 1 if the student is a member of a minority racial group.

I_{Sex} is 1 if the student is male.

X_{SES} is a number representing the student's socio-economic status.

U_i is the random effect for the ith School.

τ^2 is the randomness associated with each observation.

Results

Table 1: Estimation of fixed effects and random effects in the mixed model of math achievement dataset

	MLE	Std.Error	DF	t-value	p-value
(Intercept)	12.885	0.193	7022	66.593	0
MinorityYes	-2.961	0.206	7022	-14.393	0
SexMale	1.230	0.163	7022	7.558	0
SES	2.089	0.106	7022	19.766	0
σ	1.917	NA	NA	NA	NA
τ	5.992	NA	NA	NA	NA

From Table 1, we can see coefficients for Minority, Sex and SES are statistically significant. For i th School and j th student in that school, the model can be written as:

$$Y_{ij}|U_i \overset{ind}{\sim} N(\mu_{ij}, 5.992^2)$$
$$\mu_{ij} = 12.885 - 2.961 \times I_{Minority} + 1.230 \times I_{Male} + 2.089 \times X_{SES} + U_i$$
$$U_i \overset{ind}{\sim}, N(0, 1.917^2)$$

From Table 1, we found that Standard Deviation between schools ($\sigma = 1.917$) is smaller than the Standard Deviation within each school (τ =5.992). Therefore, we can say that there are no substantial differences between mathematics achievement scores of students in different schools.

Conclusions

In conclusion, in the fixed model we fit, the response variable is significantly associated with factor Minority, Sex and variable SES. However, there are no substantial differences between mathematics achievement scores of students in different schools since the variance between each school is smaller than the variance within each school.

Drugs Report

Introduction

This is a data set from the Treatment Episode Data Set – Discharges (TEDS-D), which provides annual data on the number and characteristics of persons discharged from public and private substance abuse treatment programs that receive public funding.

In this report, we want to investigate 2 hypothesis:

- 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.
- Some American states have particularly effective treatment programs whereas other states have programs which are highly problematic with very low completion rates.

Methods

First, to solve hypothesis 1, we encode Heroin, Opiates, Methamphetamine, Cocaine as harddrugs == TRUE and encode Marijuana and alcohol as harddrugs = FALSE.

We use **complete** as our response binary variable and consider the binomial model with logit link function. Factors includes **harddrugs** (whether the substance is hard drug), **GENDER**, **AGE**, **raceEthnicity** and **homeless**. And naturally we consider **STFIPS**, **TOWN** (the US state and town in which the treatment was given) as random effect.

We use package "inla" to fit a Bayesian Generalized Linear Mixed Models, which can be written as:

$$
\begin{aligned}
&Y_{ijk} \overset{ind}{\sim} Bernoulli(\lambda_{ij}) \\
&log(\frac{\lambda_{ij}}{1-\lambda_{ij}}) = \beta_0 + \boldsymbol{BX}_{ijk} + U_i + U_j \\
&U_i \overset{ind}{\sim} N(0, \sigma_1^2) \\
&U_j \overset{ind}{\sim} N(0, \sigma_2^2)
\end{aligned}
$$

Y_{ijk} is 1 if the **k**th individual from **i**th state **j**th town complete the treatment and 0 otherwise.

λ_{ij} corresponds to the probability that student from **i**th state **j**th town complete the treatment.

β_0 is the intercept of the model, represent the log ratio of completeness from the individuals whose harddrugs is defaluted as "FALSE", race is defaluted as "WHITE", age is defaluted as "21-24", homeless is defaluted as "FALSE" and gender is defaluted as "MALE".

$\boldsymbol{B}$ is the vector of the posterior means.

$\boldsymbol{X}_{ijk}$ is a vector that includes harddrugs (whether the substance is hard drug), GENDER, AGE, raceEthnicity and homeless infomation of the **k**th individual from **i**th state **j**th town.

U_i is the random effect of **i**th state.

U_j is the random effect of **j**th town.

Prior:

- σ_1 follows an exponential distribution with $pr(\sigma_1 > 0.1) = 0.05$.
- σ_2 follows an exponential distribution with $pr(\sigma_2 > 0.1) = 0.05$.

From the Figure1 (State-level standard deviation, see appendix) we can see the prior is flat(weakly informative) and the posterior standard deviation density curve is very different from the prior.

Table 2: Posterior means and quantiles for model parameters.

	0.5quant	0.025quant	0.975quant
(Intercept)	0.918	0.760	1.108
harddrugsTRUE	0.710	0.697	0.724

Table 3: Random effect SD quantiles

	0.5quant	0.025quant	0.975quant
SD for STFIPS	0.575	0.481	0.696
SD for TOWN	0.518	0.469	0.575

Results

For hypothesis **1**: The null hypothesis is that there is no difference between the odd ratio of completion for a young man using hard drugs(namely,"HEROIN", "OTHER OPIATES AND SYNTHETICS", "METHAMPHETAMINE" and "COCAINE/CRACK") and the odd ratio for a young man using alcohol or marijuana. From the Table 2 (Full table see appendix), we can see the 95% credible interval for **harddrugs** does not contain 1 (after exponentialized), which means the effect of hard drugs is significant under 95% credible interval. Comparing with the alcohol or marijuana group, the odd ratio of completion for a young man using hard drugs decreases 27.6% ~ 30.3%. So, we can reject the null hypothesis above.

For hypothesis **2**: The null hypothesis is that there is no difference in completion rates between different American states, which means the SD for STFIPS should be not significantly different from 0. From the Table 3 we can see the SD for random effect "State" is larger than 0 in 95% credible interval. So we can reject the null hypothesis above. In addition, we have a Table "Posterior means and quantiles for each states" in the appendix which gives more infomation about which states have particularly effective treatment programs and which states have programs which are highly problematic with very low completion rates.

Conclusions

From the data set from the Treatment Episode Data Set – Discharges (TEDS-D) we can fit a Bayesian Generalized Linear Mixed Models. The first hypothesis is true because we found evidence that the chance of a young person completing the drug treatment depends on the substance the individual is addicted to. Generally, those with 'hard' drugs (Heroin, Opiates, Methamphetamine, Cocaine) are more difficult to treat than alcohol or marijuana. The second hypothesis is also true that some American states have particularly effective treatment programs whereas other states have programs which are highly problematic with very low completion rates.

Appendix

```
# Math Report
data("MathAchieve", package = "MEMSS")
library("nlme")
m <- nlme::lme(MathAch ~ Minority + Sex + SES, random = ~1 | School, data = MathAchieve)
knitr::kable(Pmisc::lmeTable(m), digits=3,
  caption = "Estimation of fixed effects and random effects in the mixed model of math achievement dataset
```

Table 4: Estimation of fixed effects and random effects in the mixed model of math achievement dataset

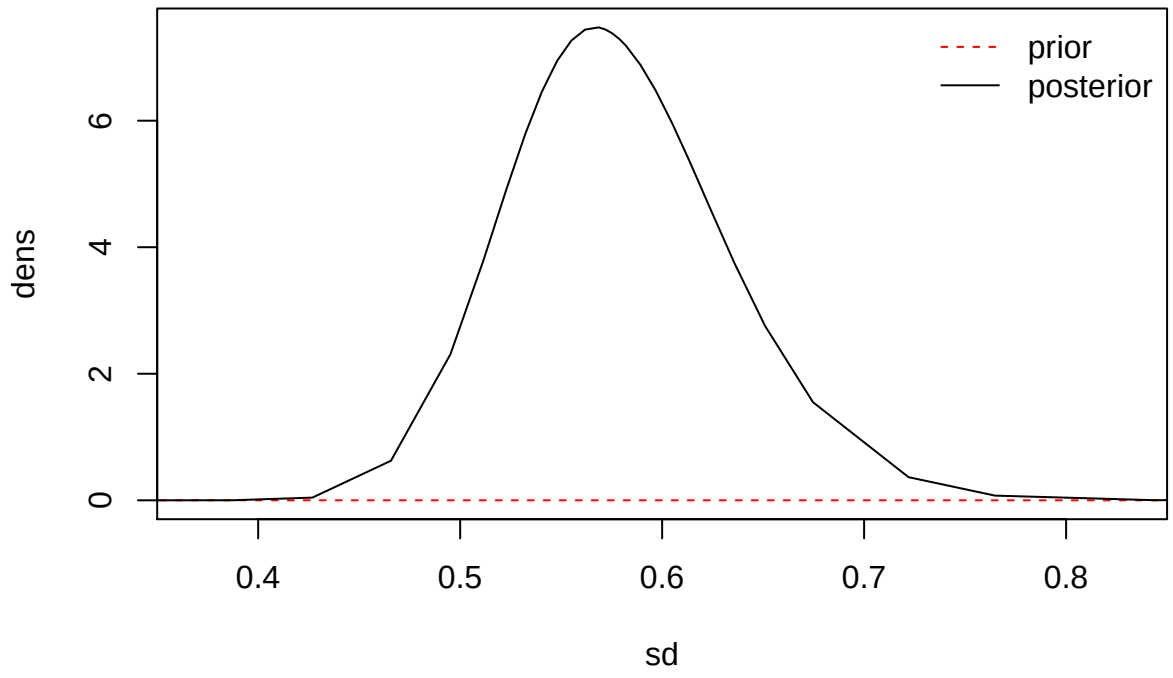
	MLE	Std.Error	DF	t-value	p-value
(Intercept)	12.885	0.193	7022	66.593	0
MinorityYes	-2.961	0.206	7022	-14.393	0
SexMale	1.230	0.163	7022	7.558	0
SES	2.089	0.106	7022	19.766	0
σ	1.917	NA	NA	NA	NA
τ	5.992	NA	NA	NA	NA

```r
# Drug Report
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)
forInla$harddrugs = ifelse(forInla$SUB1=="(4) MARIJUANA/HASHISH"|
                           forInla$SUB1=="(2) ALCOHOL", FALSE, TRUE)

library("INLA")
ires = inla(y ~ harddrugs + GENDER + raceEthnicity + AGE  + homeless +
              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.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)
title("Figure1: State-level standard deviation", cex.main= 0.9, line=1)
```

Figure1: State−level standard deviation




```
toPrint = as.data.frame(rbind(exp(ires$summary.fixed[,
  c(4, 3, 5)]), sdState$summary[, c(4, 3, 5)]))
sss = "^(harddrugs|GENDER|raceEthnicity|AGE|homeless)(.[[: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",
             caption = "Posterior means and quantiles for each states.")
```

Table 5: Posterior means and quantiles for model parameters.

	0.5quant	0.025quant	0.975quant
(Intercept)			
(Intercept)	0.918	0.760	1.108
harddrugs			
TRUE	0.710	0.697	0.724
GENDER			
FEMALE	0.916	0.901	0.932
raceEthnicity			
Hispanic	0.814	0.796	0.833
BLACK OR AFRICAN AMERICAN	0.626	0.612	0.641
AMERICAN INDIAN (OTHER TH	0.737	0.688	0.790
OTHER SINGLE RACE	0.845	0.793	0.901
TWO OR MORE RACES	0.827	0.768	0.891
ASIAN	1.133	1.039	1.236
NATIVE HAWAIIAN OR OTHER	0.843	0.748	0.951
ASIAN OR PACIFIC ISLANDER	1.440	1.216	1.705
ALASKA NATIVE (ALEUT, ESK	0.864	0.639	1.169
AGE			
18-20	0.891	0.874	0.909
15-17	0.806	0.788	0.823
12-14	0.840	0.807	0.874
homeless			
TRUE	1.009	0.977	1.041
SD for STFIPS			
SD for STFIPS	0.575	0.481	0.696
SD for TOWN			
SD for TOWN	0.518	0.469	0.575

Table 6: Posterior means and quantiles for each states.

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	-0.5	0.0	NEW JERSEY	0.5	0.2	0.7
COLORADO	0.6	0.2	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	0.3
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	-0.5	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	-0.8	0.7	SOUTH CAROLINA	0.3	0.0	0.6
IOWA	0.4	0.1	0.7	SOUTH DAKOTA	0.5	-0.3	1.2
KANSAS	-0.2	-0.5	0.1	TENNESSEE	0.2	-0.2	0.6
KENTUCKY	-0.2	-0.5	0.2	TEXAS	0.6	0.3	0.9
LOUISIANA	-0.6	-1.0	-0.2	UTAH	0.1	-0.4	0.6
MAINE	0.1	-0.6	0.9	VERMONT	-0.2	-1.0	0.6
MARYLAND	0.5	0.2	0.8	VIRGINIA	-2.8	-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	0.6	-0.1	1.2