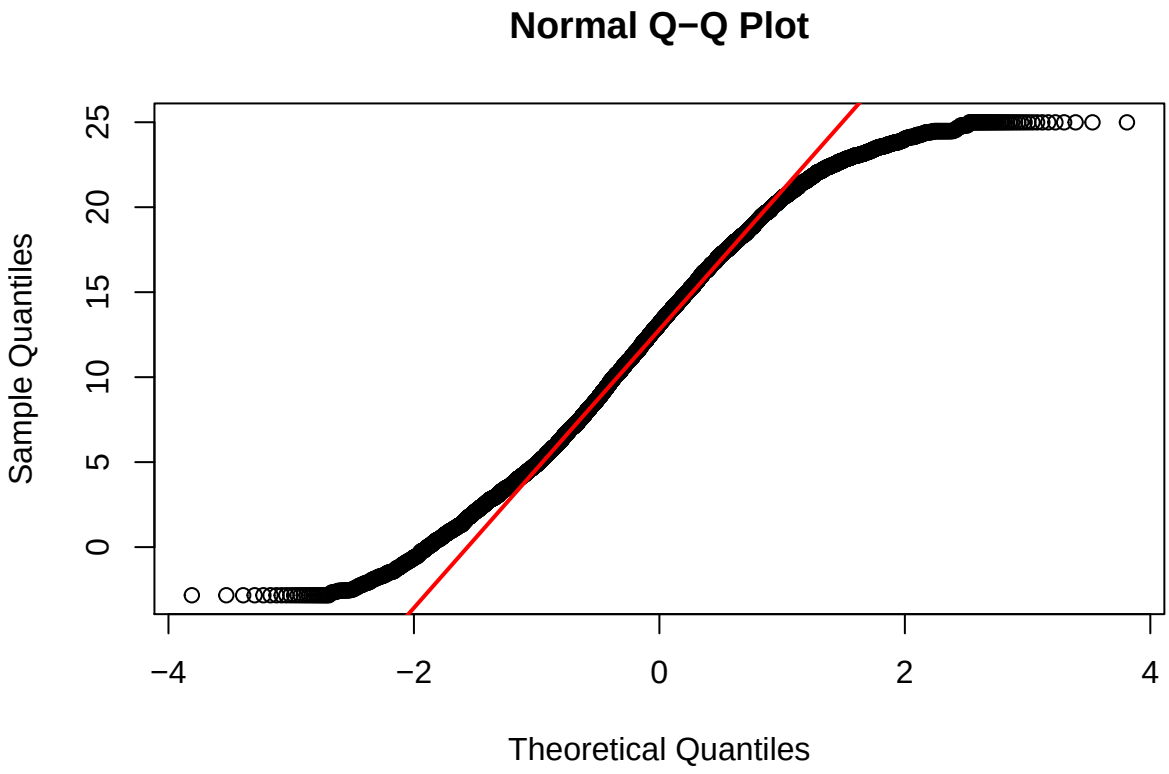
STA442 Assignment2

Xinyue Jiang

2019/10/16

Q1

We analyzed the MathAchieve dataset from MEMSS to figure out what elements influenced the math achieves of students, especially whether or not there were substantial differences between schools. We found that MathAchieve almost followed normal distribution by checking the normal QQ plot. Also,Due to the non independence in the data (the math achivements from the same school), we chose the Linear Mixed model: $Y_{ij} = X_{ij}\beta + U_i + \epsilon_{ij}$,where
X_{ij} had the indicator variables for SEX, Minority and SES.
$X_{ij}\beta$ was the fixed effect which represented how Sex, Minority (levels yes and no), and SES (socio-economic status) influenced the math achievement score of each individual student.
U_i was an school-level random effect
The summary table shows that Sex, Minority and SES are significant ($p_{Sex} = 0$, $p_{minority} = 0$, and $p_{SES} = 0$),which means they have effect on the math achievement scores of students. School was considered as the random effect. To understand the difference in the math achievement scores between different schools, we used intraclass correlation coefficient(ICC) to calculate how much variance can be explained by the random effect. The summary table showed the standard deviation of the random effect is 1.92. By using ICC ($\rho = \frac{\sigma_u^2}{\sigma_u^2+\sigma_\epsilon^2}$), we got that the variance explained by between schools was 9.3% of the total variance, whereas the variance explained by within schools was nearly 91% of the total variance. In conclusion, differences within schools are greater than differences between students from different schools.



	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
SES	2.089	0.106	7022	19.766	0
SexMale	1.230	0.163	7022	7.558	0
σ	1.917	NA	NA	NA	NA
τ	5.992	NA	NA	NA	NA

Q2

Introduction

The Treatment Episode Data Set – Discharges (TEDS-D) provides data on the number and characteristics of persons discharged from public and private substance abuse treatment programs. We analyzed TEDS-D to explore the elements that affected the discharge rate. We focused on two hypothesises.<1> The 1st is that the chance of a young person completing their drug treatment depends on the substance the individual is addict to, with 'hard' drugs (Heroin, Opiates, Methamphetamine, Cocaine) being more difficult to treat than alcohol or marijuana. <2>The 2nd 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.

Methods

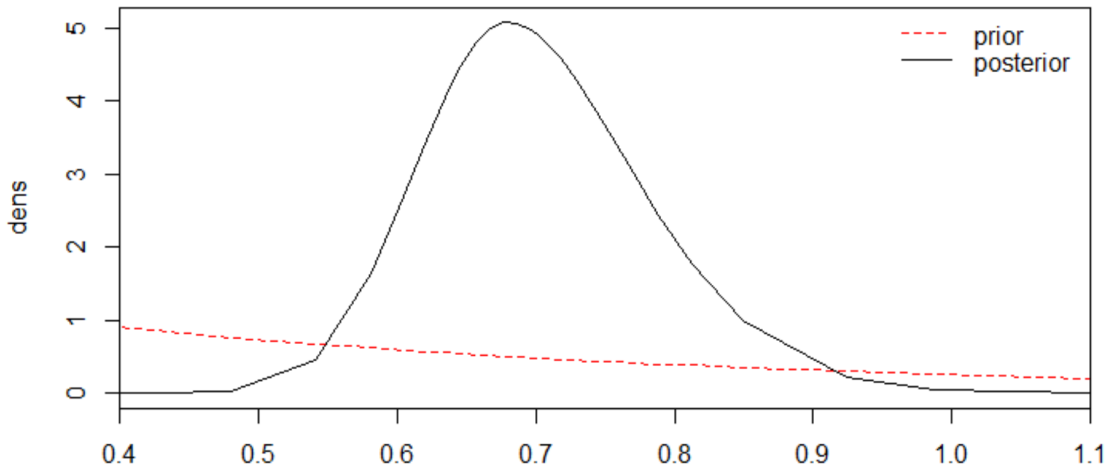
We used Bayesian Inference and logistic linear mixed model to explore the two hypothesises because treatment completed or not followed bernoulli distribution (y~ Bernoulli(λ)).Also, the data had groupings(STFIPS and TOWNS),such that we had dependent responses.

Y_{ijk}~$Bernoulli(\lambda_{ijk})$
$logit(\lambda_{ijk}) = X_{ijk}\beta + U_i + V_j$
U_i an state-level random effect
V_j an town-level random effect
X_{ijk} indicator variables for Age, SUB1, GENDER, raceEthnicity and homeless
Y_{ijk} whether or not the individual k, in state i and town j, in question completed their treatment.

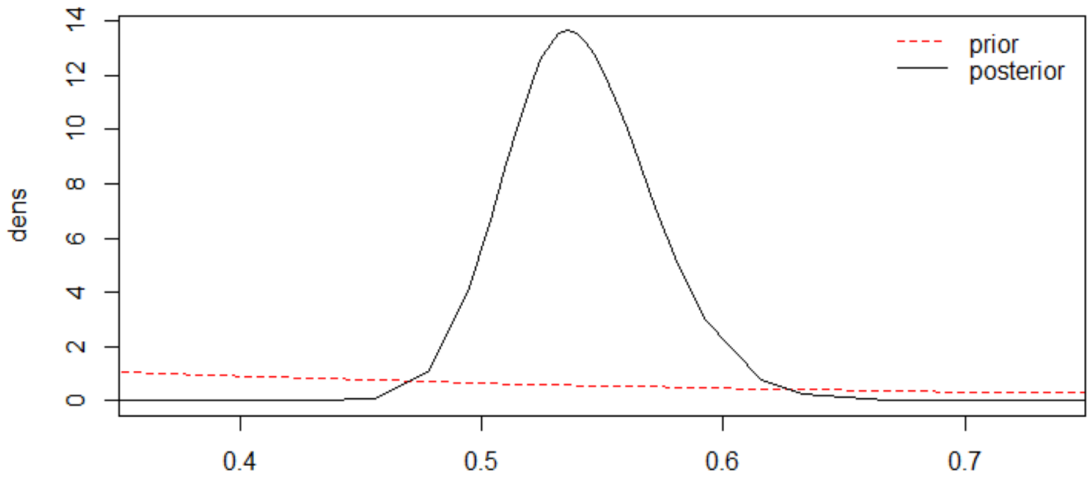
we used penalized complexity prior as our prior distribution for σ_{U_i}, and we chose the parameters (0.9, 0.15) because we thought there was 15% probability that the between STFIPS standard deviation > 0.9 ($P(\sigma_{U_i} > 0.9) = 0.15$).We also used penalized complexity prior as our prior distribution for σ_{V_j}, then adjusted it by using parameters (0.7, 0.1) because we thought there was 10% probability that the between TOWNs standard deviation > 0.7 ($P(\sigma_{V_j>0.7}) = 0.1$). We used the data Y, to adjusted the priors of σ_{U_i} and σ_{V_j}, got the posteriors for σ_{U_i} and σ_{V_j}, which are $[\sigma_{Ui}|Y]$ and $[\sigma_{Vj}|Y]$

To test Hyphothesis 1, we checked the summary table for the fixed effect and looked at the 0.5 quantile of $[\beta_{Heroin}|Y]$, $[\beta_{Opiates}|Y]$, $[\beta_{Methamphetamine}|Y]$, $[\beta_{Cocaine}|Y]$, then compared them to the 0.5 quantile of $[\beta_{alcohol}|Y]$ and $[\beta_{marijuana}|Y]$ to see how drugs' effects on the odds of completing drug treatments differed.

To test Hyphothesis 2, we checked the summary table for the state-level random effect, and looked at the posterior means of different states in America to see how different states affected the odds of completing drug treatments.



s.d. for the state-level random effect



s.d. of town-level random effect

Result

In the following table, the 0.5 quantile of $[\beta_{alcohol}|Y]$ and $[\beta_{marijuana}|Y]$ are 1 and 1.609 respectivey. The 0.5 quantile of $[\beta_{Heroin}|Y]$, $[\beta_{Opiates}|Y]$, $[\beta_{Methamphetamine}|Y]$, and $[\beta_{Cocaine}|Y]$ are 0.872,0.901,0.955, 0.855 repectively, which are smaller than those of alcohol and marijuana. Further more, We can see the 95% credible interval for $[\beta_{alcohol}|Y]$ is [1.574, 1.645], even the lower bond of the credible interval is much larger than the 0.975 quantile of the 'hard' drugs. In concusion, alchhol and marijuna have higher effect on the odds of success of drug treatment, which means alchol and marijuna addiction are easier to treat than hard drugs.

	variable	category	0.5quant	0.025quant	0.975quant
(Intercept)	(Intercept)	(Intercept)	0.716	0.574	0.893
AGE18-20	AGE18-20	AGE18-20	0.935	0.916	0.953
AGE15-17	AGE15-17	AGE15-17	0.926	0.905	0.947
AGE12-14	AGE12-14	AGE12-14	0.972	0.934	1.012
SUB1(2) ALCOHOL	SUB1	ALCOHOL	1.609	1.574	1.645
SUB1(5) HEROIN	SUB1	HEROIN	0.872	0.849	0.896
SUB1(7) OTHER OPIATES AND SYNTHETICS	SUB1	OTHER OPIATES AND SYNTHET	0.901	0.874	0.929
SUB1(10) METHAMPHETAMINE	SUB1	METHAMPHETAMINE	0.955	0.916	0.994
SUB1(3) COCAINE/CRACK	SUB1	COCAINE/CRACK	0.855	0.814	0.898
GENDER(2) FEMALE	GENDER	FEMALE	0.893	0.878	0.909
raceEthnicityHispanic	raceEthnicity	Hispanic	0.832	0.812	0.851
raceEthnicityBLACK OR AFRICAN AMERICAN	raceEthnicity	BLACK OR AFRICAN AMERICAN	0.682	0.666	0.699
raceEthnicityAMERICAN INDIAN (OTHER THAN ALASKA NATIVE)	raceEthnicity	AMERICAN INDIAN (OTHER TH	0.728	0.679	0.781
raceEthnicityOTHER SINGLE RACE	raceEthnicity	OTHER SINGLE RACE	0.866	0.812	0.923
raceEthnicityTWO OR MORE RACES	raceEthnicity	TWO OR MORE RACES	0.855	0.794	0.921
raceEthnicityASIAN	raceEthnicity	ASIAN	1.132	1.038	1.235
raceEthnicityNATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER	raceEthnicity	NATIVE HAWAIIAN OR OTHER	0.845	0.748	0.953
raceEthnicityASIAN OR PACIFIC ISLANDER	raceEthnicity	ASIAN OR PACIFIC ISLANDER	1.454	1.227	1.723
raceEthnicityALASKA NATIVE (ALEUT, ESKIMO, INDIAN)	raceEthnicity	ALASKA NATIVE (ALEUT, ESK	0.845	0.624	1.145
homelessTRUE	homeless	TRUE	1.005	0.973	1.037
SD for STFIPS	SD	STFIPS	0.697	0.562	0.883
SD for TOWN	SD	TOWN	0.539	0.487	0.603

The following table shows how different states in America affect the odds of completing drug treatment. Some states such as Colorado, Oklahoma, Florida, with posterior means equals to 0.5, 0.6, 1.0 respectively, have positive effect on the odds of completing treatment. States such as California, Illinois, North California with posterior means equals to -0.3, -0.5, -0.8 respectively, have negative effect on the odds of completing treatment. According to the table, Delaware and Florida have the most effective treatment programs; whereas New Mexico and North California are highly problematic with very low completion rates.

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	0.8	NEBRASKA	0.8	0.4	1.2
ARIZONA	0.0	-1.3	1.3	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	-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	-1.2	-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	-1.0	0.6	PENNSYLVANIA	0.0	-1.3	1.3
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	1.3
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	-2.5
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

Conclusion

We analyzed the data provided by TEDS-D, which contained the information of annual discharges from substance abuse treatment facilities. We used the logistic linear mixed model to fit the data due to the dependence brought by the groupings of states and towns and the bernoulli distributed completing treatment. We also chose the Bayes Inference to estimate the value of each beta and the standard deviation of state-level and town-level radom effects. By fitting the GLMM model, we found hard drugs like Heroin, Opiates, Methamphetamine and Cocaine, were more difficult to treat than alchhol and marijuna. Also, the effectiveness of treatment programs between states were different.

Appendix

```
#Question1

install.packages("MEMSS")
install.packages("nlme")
install.packages("lme4")
library(nlme)
library(Matrix)
library(lme4)

#QQ plot
data("MathAchieve", package = "MEMSS")
head(MathAchieve)
qqnorm(MathAchieve$MathAch)
qqline(MathAchieve$MathAch,col='red',lwd=2)

#summary table
mem <- lme(MathAch ~ Minority + SES + Sex, random = ~1 |School, data = MathAchieve)
memtable <- summary(mem)$tTable[, -3]
knitr::kable(Pmisc::lmeTable(mem), digit =3)




 ####################################

#Question2

install.packages("INLA", repos = c(getOption("repos"),INLA= "https://inla.r-inla-download.org/R/stable")
install.packages("Pmisc", repos = "http://r-forge.r-project.org")
install.packages("Hmisc")
install.packages("data.table")
install.packages("nlme")
download.file("http://pbrown.ca/teaching/appliedstats/data/drugs.rds",
              "drugs.rds")
xSub = readRDS("drugs.rds")
table(xSub$SUB1)
attributes(xSub)
table(xSub$STFIPS)[1:5]

table(xSub$TOWN)[1:2]
```

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

library(Matrix)
library(parallel)
library(sp)
library(INLA)
library(lattice)
library(survival)
library(Formula)
library(ggplot2)
library(Hmisc)
library(Pmics)

ires = inla(y ~ AGE +GENDER + raceEthnicity + homeless +
              f(STFIPS, hyper=list(prec=list(
                 prior='pc.prec', param=c(0.9, 0.15)))) +
               f(TOWN, hyper=list(prec=list(
                 prior='pc.prec', param=c(0.7, 0.1)))),
            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)),
Pmisc::mdTable(toPrint, digits = 3, mdToTex = TRUE,
               guessGroup = TRUE, caption = "Posterior means and quantiles for model parameters.")
knitr::kable(toPrint,digits = 3)

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)
```