STA442 Assignment 2
Zhuojing Qiu
16/10/2019

Question1. Would Better Schools Give You A Substantially Better Math Achievement?

Introduction

We analyzed the MathAchieve dataset from the MEMSS package. There are in total 6 variables in the dataset, including *School, Minority, Sex, SES, MathAch,* and *MEANSES.* Our interest was to analyze the differences in mathematics achievement between these schools, so we treated *School* as our random effect.

Methods

First, we tested the assumption under mixed model, that is, the normality of our data.

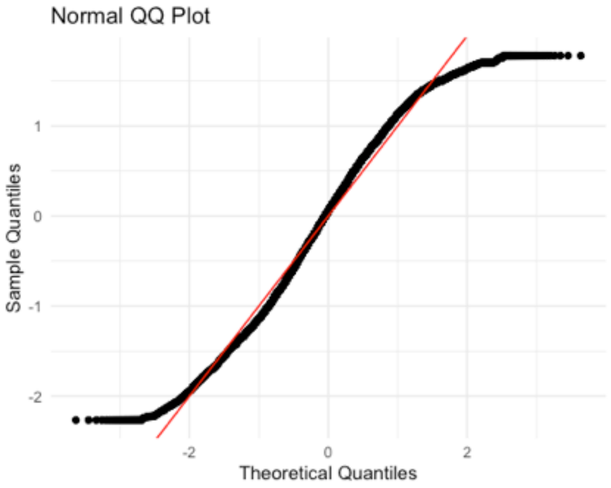


Figure 1.Evaluating Normality of *MathAch*

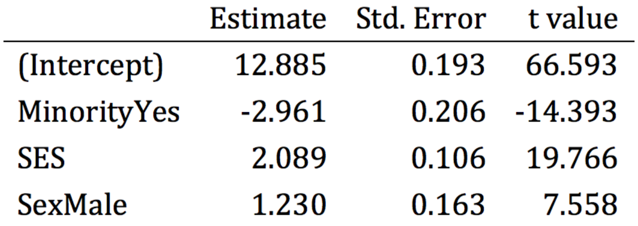
From the normality test, we could see that there was definitely some concerns with the two tails. The tails were heavier than those in a normal distribution. But the data appears not to be skewed relative to the normal distribution. We would proceed anyways.

We fitted a linear mixed model with random intercept for *School* and fixed effects for *Minority, Sex* and *SES*. Below is our fitted model.

$$MathAch_{(ij)} = \beta_0 + \beta_1 I_{MinorityYes(ij)} + \beta_2 I_{SexFemale(ij)} + \beta_3 x_{SES(ij)} + \beta_4 x_{School(ij)} + U_{(i)}$$

Where Ui is the random intercept for each school representing school i's deviation from the population average.

Results



	Estimate	Std. Error	t value
(Intercept)	12.885	0.193	66.593
MinorityYes	-2.961	0.206	-14.393
SES	2.089	0.106	19.766
SexMale	1.230	0.163	7.558

Table 1. Estimation of fixed effects in linear mixed model of MathAchieve dataset

Groups	Variance
School (Math Achievement Variance Between Schools)	3.674
Residual (Math Achievement Variance Within Schools)	35.909

Table 2. Estimation of random effects in linear mixed model of MathAchieve dataset

Our model gave us the estimation of both fixed effects and random effects in linear mixed model of MathAchieve dataset. Based on Table 2, we calculated the between-group variance to be 0.093 and within-group variance to be 0.907. In other words, variations caused by differences between schools is 9.3% and variations caused by differences within schools is 90.7%.

We then tested another model assumption, that is, the normality of the random effects.

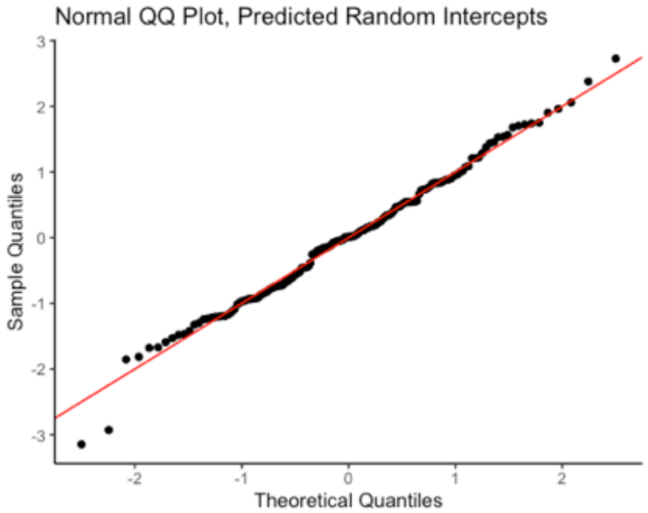


Figure 2. Evaluating Normality of Predicted Random Intercepts

Although it still shows a difference in the tails, the normality of predicted random intercepts is considerably better now.

Summary

We carried out an analysis treating school as a random effect, Minority and SES as fixed effects in order to compare the differences within schools and differences between students from different schools. From the results, only 9.3% of the variance in the data is due to variance between schools. Therefore, math achievement varies more within schools than between them. That is, there does not appear to be any substantial differences between schools.

Question2. Is There an Easiest Way To Complete a Drug Treatment?

Introduction

We analyzed the Treatment Episode Data Set using an R version of the dataset, which can be downloaded from pbrown.ca/teaching/appliedstats/data/drugs.rds page. The original dataset was provided by The Treatment Episode Data Set-Discharges (TEDS-D).

Our main focus was to investigate the relationship between the chance of a young person completing their drug treatment and different substances that the individual is addicted to. Specifically, we would like to find out if 'hard' drugs (Heroin, Opiates, Methamphetamine, Cocaine) are being more difficult to treat than alcohol or marijuana.We also investigated the relationship between different states and the treatment program completion rate, in order to see if American states have more effective treatment programs than other states.

Since our interest was to find out the effect of substance and states, we controlled the remaining variables, in particular, gender, race ethnicity, town, age, education level and whether homeless or not to be more accurate.

Methods

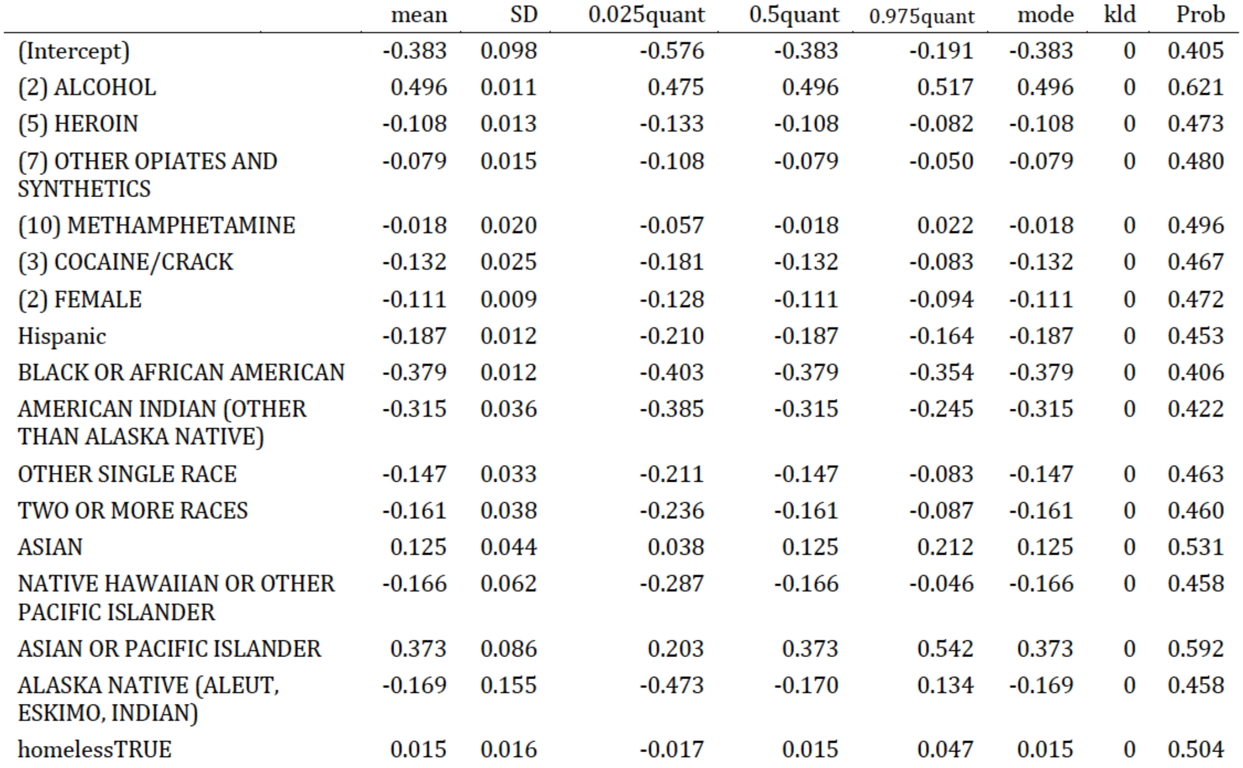
We fitted a binomial model with the predictors SUB1, GENDER, raceEthnicity, homeless, STFIPS and TOWN, using Bayesian inference with INLA. Since we were using Bayesian here, a prior was needed for all unknown quantities. For example, we chose a PC prior to give 5% chance that the random effect standard deviation of STFIPS is greater than 0.1, then we set another prior for the standard deviation of TOWN.

Our model also contained random intercepts for STFIPS and *TOWN*, and fixed effects for *SUB1*, *GENDER*, *raceEthnicity* and *homeless*. Below is our fitted model.

$$ln\left(\frac{p}{1-p}\right)_{(ij)} = \beta_0 + \beta_1 x_{SUB1(ij)} + \beta_2 I_{GENDER\ FEMALE(ij)} + \beta_3 x_{raceEthnicity(ij)} + \beta_4 I_{homeless(ij)} + \beta_5 x_{STFIPS(ij)} + \beta_6 x_{TOWN(ij)} + U_{(i)} + V_{(i)}$$

Where *p* is chance of a young person completing their drug treatment, *Ui* is the random intercept for each state representing state i's deviation from the population average, *Vi* is the random intercept for each town representing town i's deviation from the population average

Results



	mean	SD	0.025quant	0.5quant	0.975quant	mode	kld	Prob
(Intercept)	-0.383	0.098	-0.576	-0.383	-0.191	-0.383	0	0.405
(2) ALCOHOL	0.496	0.011	0.475	0.496	0.517	0.496	0	0.621
(5) HEROIN	-0.108	0.013	-0.133	-0.108	-0.082	-0.108	0	0.473
(7) OTHER OPIATES AND SYNTHETICS	-0.079	0.015	-0.108	-0.079	-0.050	-0.079	0	0.480
(10) METHAMPHETAMINE	-0.018	0.020	-0.057	-0.018	0.022	-0.018	0	0.496
(3) COCAINE/CRACK	-0.132	0.025	-0.181	-0.132	-0.083	-0.132	0	0.467
(2) FEMALE	-0.111	0.009	-0.128	-0.111	-0.094	-0.111	0	0.472
Hispanic	-0.187	0.012	-0.210	-0.187	-0.164	-0.187	0	0.453
BLACK OR AFRICAN AMERICAN	-0.379	0.012	-0.403	-0.379	-0.354	-0.379	0	0.406
AMERICAN INDIAN (OTHER THAN ALASKA NATIVE)	-0.315	0.036	-0.385	-0.315	-0.245	-0.315	0	0.422
OTHER SINGLE RACE	-0.147	0.033	-0.211	-0.147	-0.083	-0.147	0	0.463
TWO OR MORE RACES	-0.161	0.038	-0.236	-0.161	-0.087	-0.161	0	0.460
ASIAN	0.125	0.044	0.038	0.125	0.212	0.125	0	0.531
NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER	-0.166	0.062	-0.287	-0.166	-0.046	-0.166	0	0.458
ASIAN OR PACIFIC ISLANDER	0.373	0.086	0.203	0.373	0.542	0.373	0	0.592
ALASKA NATIVE (ALEUT, ESKIMO, INDIAN)	-0.169	0.155	-0.473	-0.170	0.134	-0.169	0	0.458
homelessTRUE	0.015	0.016	-0.017	0.015	0.047	0.015	0	0.504

Table 3. Estimation of Fixed Effects in Binomial Regression Model of Drugs Dataset

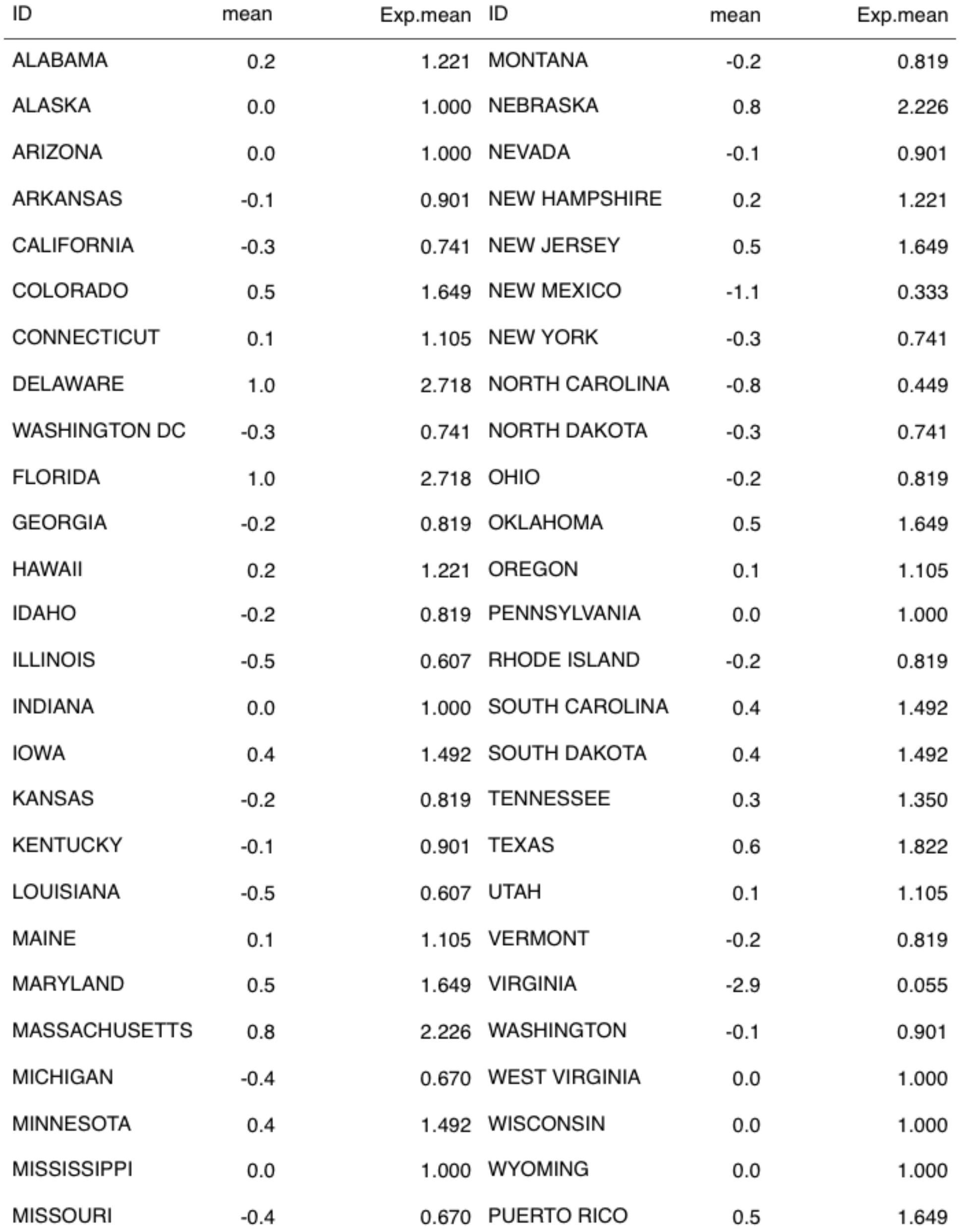
We got a good summary of the fixed effects parameters estimates from our model, including all the information to construct point estimates and 95% min-width credible intervals. Because we were looking at the marginal posteriors, that is, each parameter one at a time, the credible

completing a drug treatment by calculating $\frac{e^{\beta_1}}{1+e^{\beta_1}}$, where β_1 is the coefficients in the mean column.

interval was still correct. To better interpret the coefficients, we found the probability of

In addition, the substance MARIJUANA/HASHISH was treated as our reference group while fitting the model, it would have a mean of 0 and a corresponding probability of 0.5.

After comparing the probability, which represents the chance of an individual completing their drug treatments, we saw that alcohol and marijuana have a relatively higher probability than Heroin, Opiates, Methamphetamine and Cocaine. In other words, if a young person is addicted to either alcohol or marijuana, then the chance of that person completing their drug treatment is approximately 62.1% and 50% respectively; if a young person is addicted to other ‘hard’ drugs, then the chance of that person completing their drug treatment is generally below 50%.



ID	mean	Exp.mean	ID	mean	Exp.mean
ALABAMA	0.2	1.221	MONTANA	-0.2	0.819
ALASKA	0.0	1.000	NEBRASKA	0.8	2.226
ARIZONA	0.0	1.000	NEVADA	-0.1	0.901
ARKANSAS	-0.1	0.901	NEW HAMPSHIRE	0.2	1.221
CALIFORNIA	-0.3	0.741	NEW JERSEY	0.5	1.649
COLORADO	0.5	1.649	NEW MEXICO	-1.1	0.333
CONNECTICUT	0.1	1.105	NEW YORK	-0.3	0.741
DELAWARE	1.0	2.718	NORTH CAROLINA	-0.8	0.449
WASHINGTON DC	-0.3	0.741	NORTH DAKOTA	-0.3	0.741
FLORIDA	1.0	2.718	OHIO	-0.2	0.819
GEORGIA	-0.2	0.819	OKLAHOMA	0.5	1.649
HAWAII	0.2	1.221	OREGON	0.1	1.105
IDAHO	-0.2	0.819	PENNSYLVANIA	0.0	1.000
ILLINOIS	-0.5	0.607	RHODE ISLAND	-0.2	0.819
INDIANA	0.0	1.000	SOUTH CAROLINA	0.4	1.492
IOWA	0.4	1.492	SOUTH DAKOTA	0.4	1.492
KANSAS	-0.2	0.819	TENNESSEE	0.3	1.350
KENTUCKY	-0.1	0.901	TEXAS	0.6	1.822
LOUISIANA	-0.5	0.607	UTAH	0.1	1.105
MAINE	0.1	1.105	VERMONT	-0.2	0.819
MARYLAND	0.5	1.649	VIRGINIA	-2.9	0.055
MASSACHUSETTS	0.8	2.226	WASHINGTON	-0.1	0.901
MICHIGAN	-0.4	0.670	WEST VIRGINIA	0.0	1.000
MINNESOTA	0.4	1.492	WISCONSIN	0.0	1.000
MISSISSIPPI	0.0	1.000	WYOMING	0.0	1.000
MISSOURI	-0.4	0.670	PUERTO RICO	0.5	1.649

Table 4. Estimation of Random Effects in Binomial Regression Model of Drugs Dataset

Table 4 provided us with the exponentiated mean of random effects parameter *STFIPS* (which represents the odds of an individual completing a drug treatment). We found that if an individual is taking a treatment program in Alabama or Iowa, controlling the remaining variables, this person would be 22% and 49% more likely to complete the treatment respectively. In other words, we would consider a treatment program to be effective if the corresponding state has an exponentiated mean that is greater than 1.

If the treatment program was provided by Alaska or Arizona, controlling the remaining variables, the probability of an individual completing the treatment would both remain the same. In other words, any states with an exponentiated mean equal to 1 would be considered to have a treatment program that is average, that is, neither particularly effective nor highly problematic.

If it was Arkansas or California's treatment program, controlling the remaining variables, an individual that is taking the treatment program would be 10% and 26% less likely to complete treatment program respectively. In other words, we would consider a treatment program to be problematic if the corresponding state has an exponentiated mean that is less than 1.

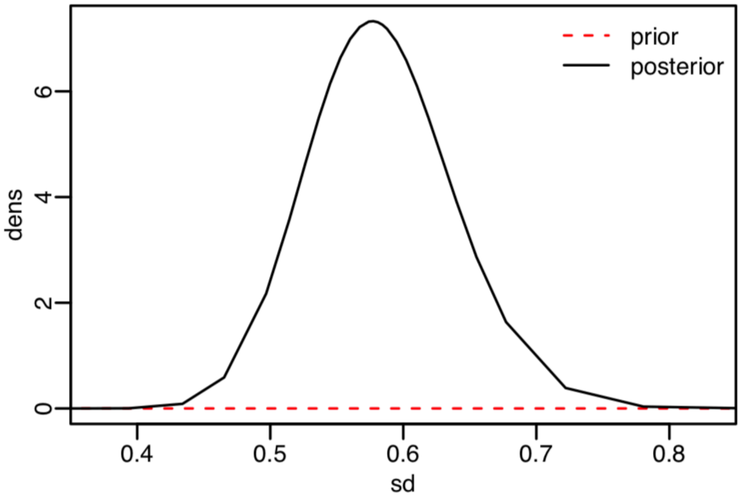


Figure 3. Posterior for State Standard Deviation

	0.5quant	0.025quant	0.975quant
SD			
STFIPS	0.581	0.482	0.698
TOWN	0.537	0.482	0.597

Table 5. Standard Deviation of Fixed Effects

We plotted out the posteriors and priors for standard deviation of states in order to compare the credible intervals to the confidence intervals. From the plot, we could get an interval estimate for standard deviation sigma that is approximately from 0.48 to 0.7. Reading off Table 5, we had

P(0.482 < sigma < 0.698) = 95%, which means that there is a 95% credible interval for sigma between 0.482 and 0.698. Therefore, we got two very close intervals showing that the biased was towards zero by design.

Conclusions

We carried out an analysis by fitting an mixed effects binomial regression model, where random effects were STFIPS and TOWN, and fixed effects are SUB1, GENDER, raceEthnicity and homeless.

From our results, we found that the chance of a young person with an addiction to ‘hard drugs’ is slightly lower than that with an addiction to alcohol and marijuana. That is, a young person with ‘hard drugs’ addiction appears to have more difficulty completing a drug treatment program. However, in this model, we did not have enough information to determine the correlation between substance and completion rate.

In addition, by looking at the exponentiated estimates of random effects parameter *STFIPS*, we found that different states have different effective levels of treatment programs. Among the 52 states in our dataset, Delaware and Florida have the most effective treatment programs that is 2.7 times more likely than average for an individual to complete. On the other hand, New Mexico has a treatment program that is only 33% of the average likely to be completed.

Therefore, we can say that some American states have particularly effective treatment programs whereas other states have programs which are considered highly problematic with very low completion rates.

Appendix

```
#Question1
library(tidyverse)
library(lme4)
library(nlme)
data("MathAchieve", package = "MEMSS")
head(MathAchieve)

# ModelAssumptions: normality of data
MathAchieve%>%
  dplyr::select(MathAch) %>%
  arrange(MathAch) %>%
  mutate_at("MathAch",funs( (. - mean(.)) / sd(.))) %>%
  mutate(q = qnorm(seq(1:nrow(MathAchieve)) / (1 + nrow(MathAchieve)))) %>%
  ggplot(aes(x = q,y = MathAch)) +
  theme_minimal() +
  geom_point() +
  geom_abline(slope = 1,intercept = 0,colour = "red") +
  labs(title = "Normal QQ Plot",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles")

# Fit the model
mod<-lmer(MathAch~1+(1|School)+Minority+SES+Sex, data=MathAchieve)
summary(mod)
random.effects(mod)
knitr::kable(summary(mod)$coef, digits=3)

# Test another model assumptions: normality of random effects
tibble(b = ranef(mod)$School[,1]) %>%
  mutate_at('b',funs((.-mean(.))/sd(.))) %>%
  arrange(b) %>%
  mutate(q=qnorm(seq(1:nrow(ranef(mod)$School))/(1+nrow(ranef(mod)$School)))) %>%
  ggplot(aes(x=q,y=b)) +
  theme_classic() +
  geom_point() +
  geom_abline(slope=1,intercept=0,colour='red') +
  labs(title = "Normal QQ Plot, Predicted Random Intercepts",
       x="Theoretical Quantiles",
       y="Sample Quantiles”)


#Question2
library(INLA)
download.file("http://pbrown.ca/teaching/appliedstats/data/drugs.rds", "drugs.rds")
xSub = readRDS("drugs.rds")
table(xSub$SUB1)
table(xSub$STFIPS)[1:5]
table(xSub$TOWN)[1:2]
forInla = na.omit(xSub)
forInla$y = as.numeric(forInla$completed)
logit_to_prob <- function(logit){
  odds <- exp(logit)
  prob <- odds /(1 + odds)
  return(prob)}

ires = inla(y ~ SUB1 + GENDER + raceEthnicity + homeless +
           f(STFIPS, hyper=list(prec=list(prior='pc.prec', param=c(0.1, 0.05)))) +
             f(TOWN), data=forInla, family='binomial', control.inla = list(strategy='gaussian',
int.strategy='eb'))
Print = as.data.frame(rbind(ires$summary.fixed))
Prob <- logit_to_prob(Print$mean)
Print_new <- cbind(Print, Prob)
colnames(Print_new) <- c('mean','SD','0.025quant','0.5quant','0.975quant','mode','kld','Prob')
rownames(Print_new) = gsub("SUB1", "", rownames(Print_new) )
rownames(Print_new) = gsub("GENDER", "", rownames(Print_new) )
rownames(Print_new) = gsub("raceEthnicity", "", rownames(Print_new) )
knitr::kable(Print_new, caption = 'Estimated parameters of fixed effects', digits=3)

sdState = Pmisc::priorPostSd(ires)
do.call(matplot, sdState$STFIPS$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$ID)
toprint = cbind(ires$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")
```