STA442 Homework2

SongQi Wang 1003439442

2019-10-16

MathAchieve

Introduction

In the data set MathAchieve (MEMSS package), there are 7185 observations. We want to see the substantial differences between schools and their behavior.

Method

It is easey to see that factors Minority (levels yes and no), and the variable SES (socio-economic status) are clearly fixed effects. We used Linear mixed models, and school is treated as a random effect:

$$
\begin{aligned}
Y_{ij} \mid U &\sim N(\mu_{ij}, \sigma^2) \\
\mu_{ij} &= X_{ij}\beta + U_i \\
U_i &\sim N(0, \sigma_U^2)
\end{aligned}
$$

where:

- Y_{ij} is the individual's MathAchieve j in the school i
- $X_{ij}\beta$ contains the intercept, whether the individual is Minority, individual's gender, and individual's socio-economic status.
- U_i is the random effect of different schools.

Results

The results of the fixed effects are summarized in table 1. We ckeck whether it appears that there are substantial differences between schools from the result of random effects. We get $\sigma_U^2 = 3.674$ and $\sigma^2 = 35.909$. So the intraclass correlation coefficient or the proportion of variance explained by *school* is $\frac{\sigma_U^2}{\sigma^2+\sigma_U^2} = \frac{3.674}{35.909+3.674} \approx 9.281\%$, which is very small. Therefore, the substantial differences between schools are very small.

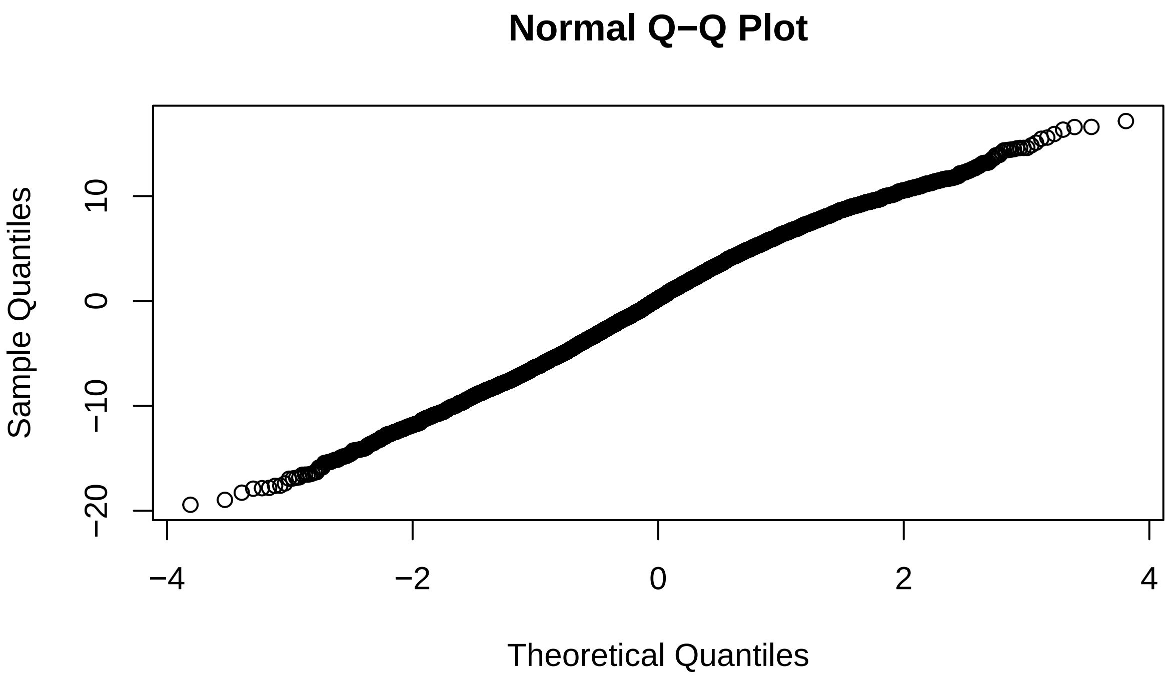
Table 1: Estimation of fixed effects in linear mixed model

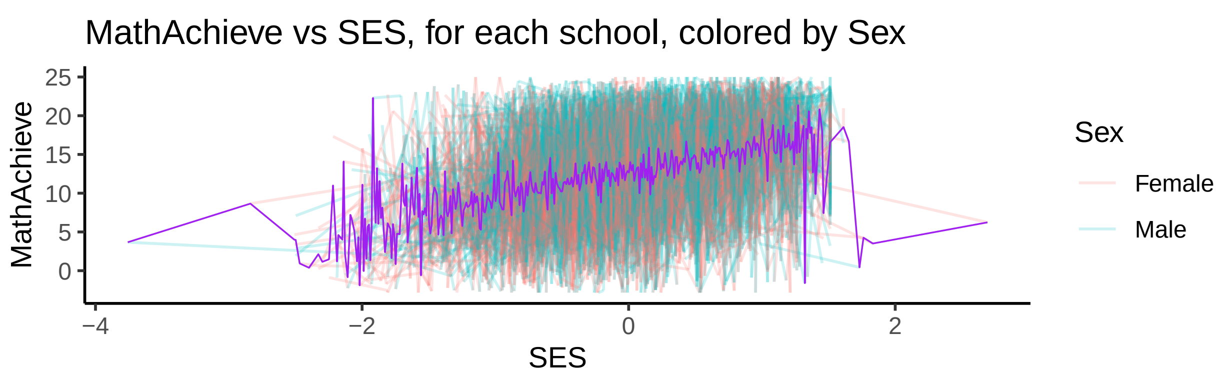
	Estimate	Std. Error	t value
Intercept	12.885	0.193	66.593
Minority	-2.961	0.206	-14.393
SexMale	1.230	0.163	7.558
SES	2.089	0.106	19.766

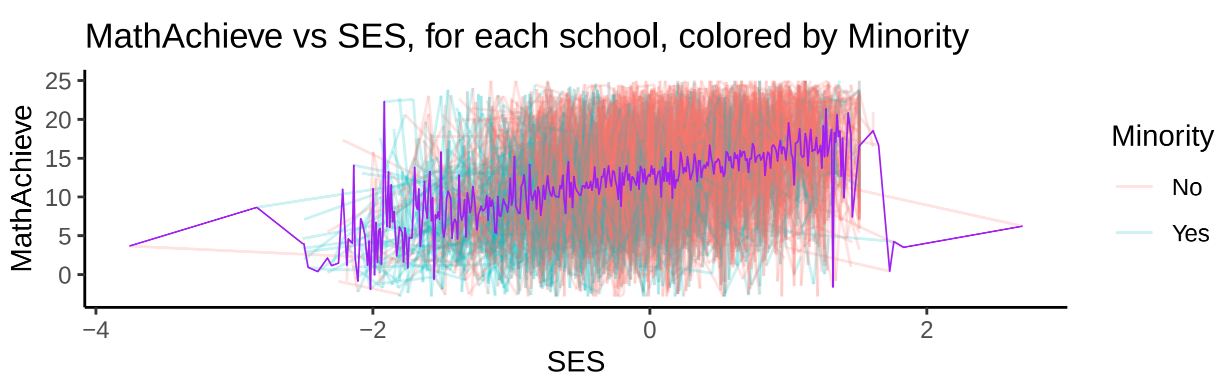
Other obsivations show that our medol is proper and fit the data very well.

From the QQ plot, we can see the normality of our model is satisfied.

From two line plots of the datasets. We can see the Mathachieve of male students is little bit higher, and the Minority has a lower Mathchieve. These two plots are consistent with the estimation of fixed effects.







Drugs treatment program

Itroduction

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. Based on this data set, following hypoheses are discussed in this report.

1.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.

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

Methods

Since we were dealing with the success rate of the treatment, we uesd logistic regression model. STFIPS and TOWN are treated as random effects.

$$
\begin{aligned}
Y_{ij} &\sim Bernoulli(\pi_i) \\
\ln(\frac{\pi_i}{1-\pi_i}) &= \mu + X_{ij}\beta + U_i + V_i \\
U_i &\sim N(0, \sigma_U^2) \\
V_i &\sim N(0, \sigma_V^2)
\end{aligned}
$$

where:

- π_i is the individual's treatment success rate.
- $X_{ij}\beta$ contains the intercept, individual's primary addiction, age, gender and ethnicity.
- U_i is the random effect of STFIPS.
- V_i is the random effect of TOWN.

To use Bayesian inference, we set following penalized complexity prior. The plots of prior and posterior show that our prior is reasonable, you can see the plots in Appendix:

$$
\begin{aligned}
P(\sigma_U > 0.81) &= 5\% \\
P(\sigma_V > 0.63) &= 5\%
\end{aligned}
$$

And the null hypoheses we tested are:

$$
H_0 \colon \beta_{Heroin} = \beta_{Opiates} = \beta_{Cocaine/Crack} = \beta_{Methamphetamine} = \beta_{Alcohol} = 0
$$
$$
H_0 \colon \sigma_U^2 = \sigma_V^2 = 0
$$

Table 2: Posterior means and quantiles for model parameters.

	0.5quant	0.025quant	0.975quant
(Intercept)			
(Intercept)	0.716	0.575	0.891
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	0.917	0.994
COCAINE/CRACK	0.855	0.814	0.898
GENDER			
FEMALE	0.893	0.878	0.909
AGE18-20			
AGE18-20	0.935	0.916	0.953
AGE15-17			
AGE15-17	0.926	0.905	0.947
AGE12-14			
AGE12-14	0.972	0.934	1.012
raceEthnicity			
Hispanic	0.832	0.812	0.851
BLACK OR AFRICAN AMERICAN	0.682	0.666	0.699
AMERICAN INDIAN (OTHER TH	0.728	0.679	0.781
OTHER SINGLE RACE	0.865	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	0.845	0.748	0.953
ASIAN OR PACIFIC ISLANDER	1.454	1.227	1.723
ALASKA NATIVE (ALEUT, ESK	0.845	0.624	1.145
homeless			
TRUE	1.005	0.973	1.037
SD			
STFIPS	0.688	0.556	0.866
TOWN	0.538	0.486	0.601

Results

The results of posterior means and quantiles for model parameters are summarized in table 2. All the model paremeters in the table are exponentialed values of β. The reference group in the model is marijuana, so it's exponentialed parameter equals to 1. As we can see in the table, the exponentialed parameters of Heroin, Opiates, Methamphetamine and Cocaine are less than 1 and the exponentialed parameter of Alcohol is greater than 1, which means the treatment of Alcohol have a higher success rate than marijuana, and the treatments of these 'hard' drugs have a lower success rate than alcohol and marijuana.

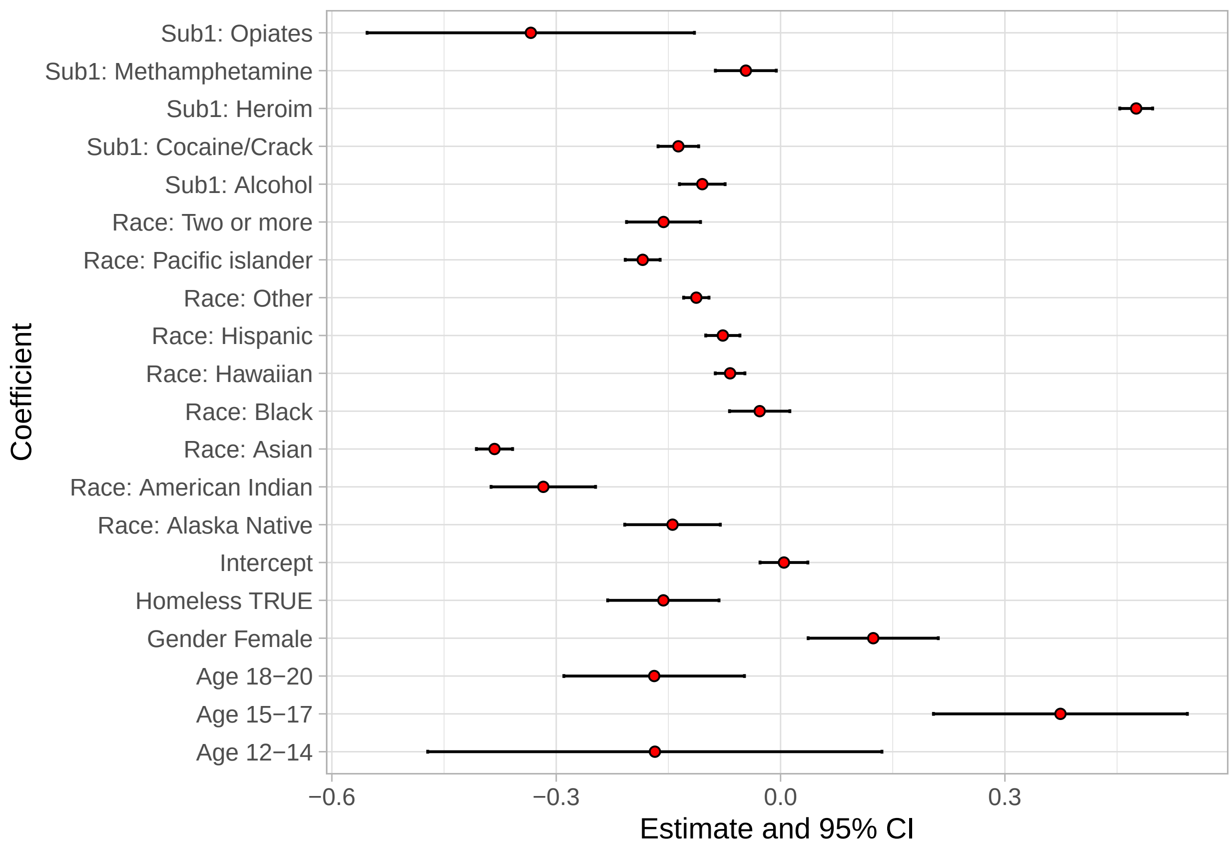


Figure 1: Estimate and 95% CI

Additionally, we can plot the credible interval of these parameters to see them more clearly.

Table 3: The random effects of each US state and town

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

The random effects of each US state and town are summarized in table 3. The higher the random effect of a state is, the better the treatment programs a state has. The random effect of Virginia is -2.9, which means the treatment programs in Virginia are much less effective than other states. Delaware and Florida have the random effects of 1. Their treatment programs have a higher success rate.

Conclusions

Baes on the model we have, we can conclude that

1.The chance of a young person completing their drug treatment does depend on the substance the individual is addicted to. 'Hard' drugs (Heroin, Opiates, Methamphetamine, Cocaine) are more difficult to treat than alcohol or marijuana.

2.Some American states have particularly effective treatment programs, such as Delaware and Florida. And some other states' programs are highly problematic with very low completion rates, such as Virgina.

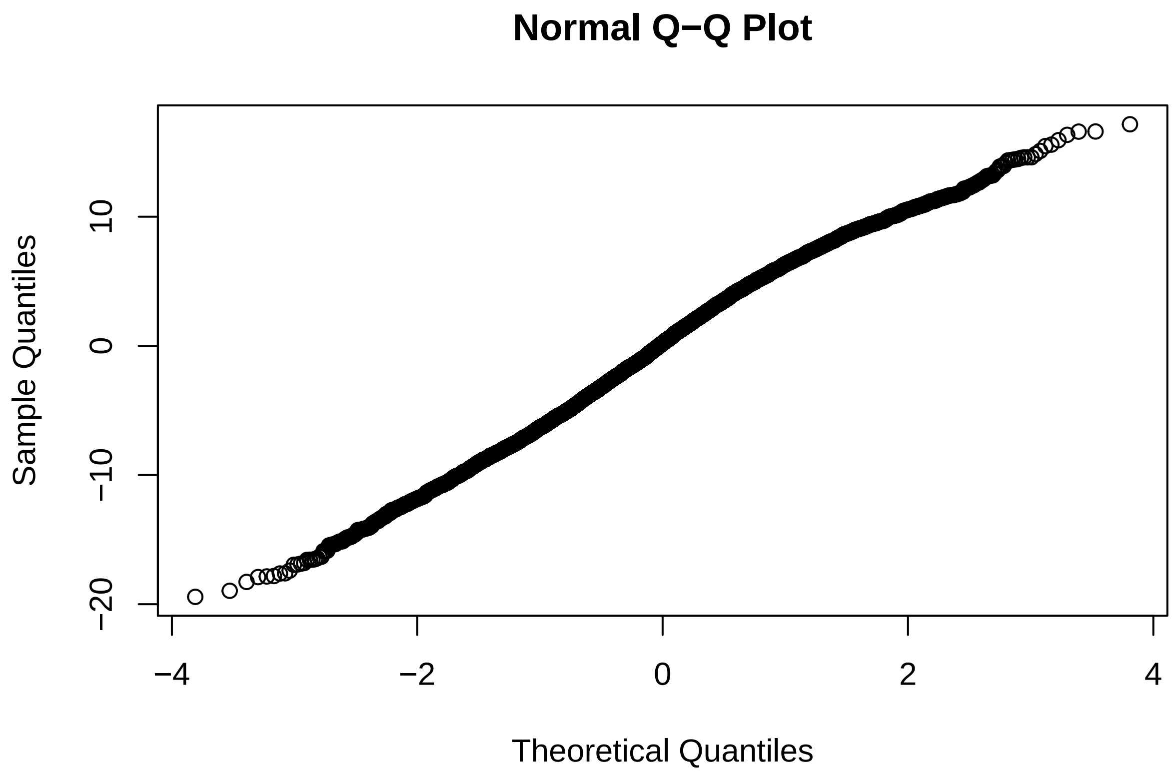
Table 4: Estimation of fixed effects in linear mixed model

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

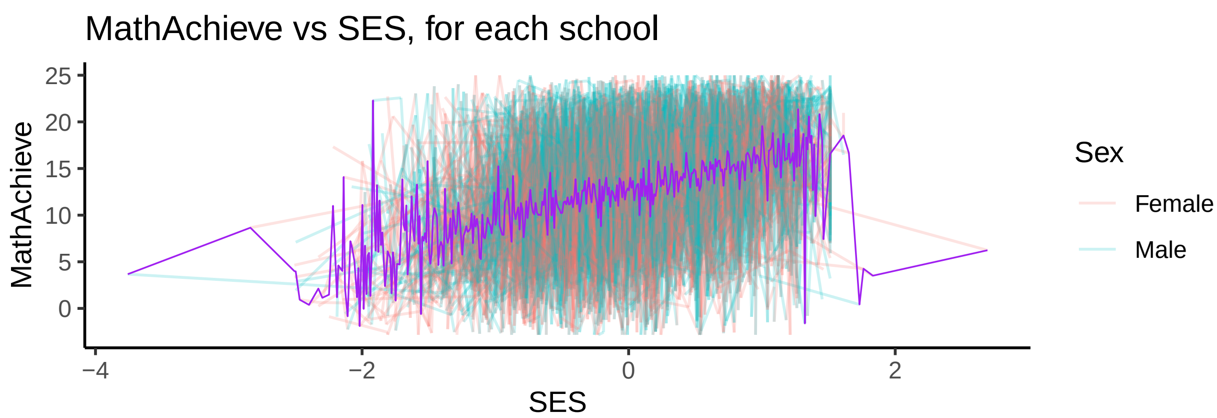
Appendix

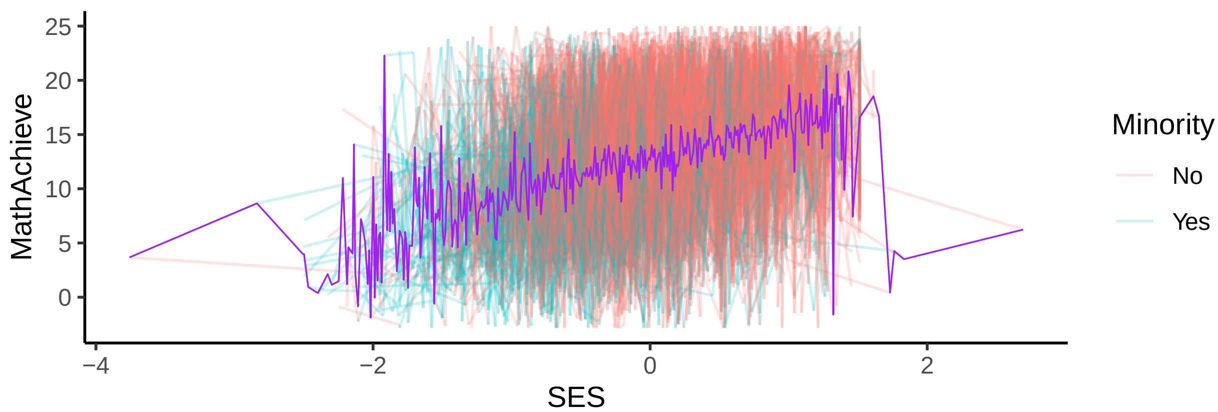
```
# MathAchieve

data("MathAchieve", package = "MEMSS")
model1 = lmer(MathAch ~ Minority + Sex + SES + (1 | School), data = MathAchieve)
fix_table1 = summary(model1)$coef

colnames(fix_table1) <- c("Estimate","Std. Error","t value")
rownames(fix_table1) <- c("Intercept",
                          "Minority",
                          "SexMale",
                          "SES")

knitr::kable(fix_table1, digits = 3, caption = "Estimation of fixed effects in linear mixed model") %>%
```

```
qqnorm(resid(model1))
```

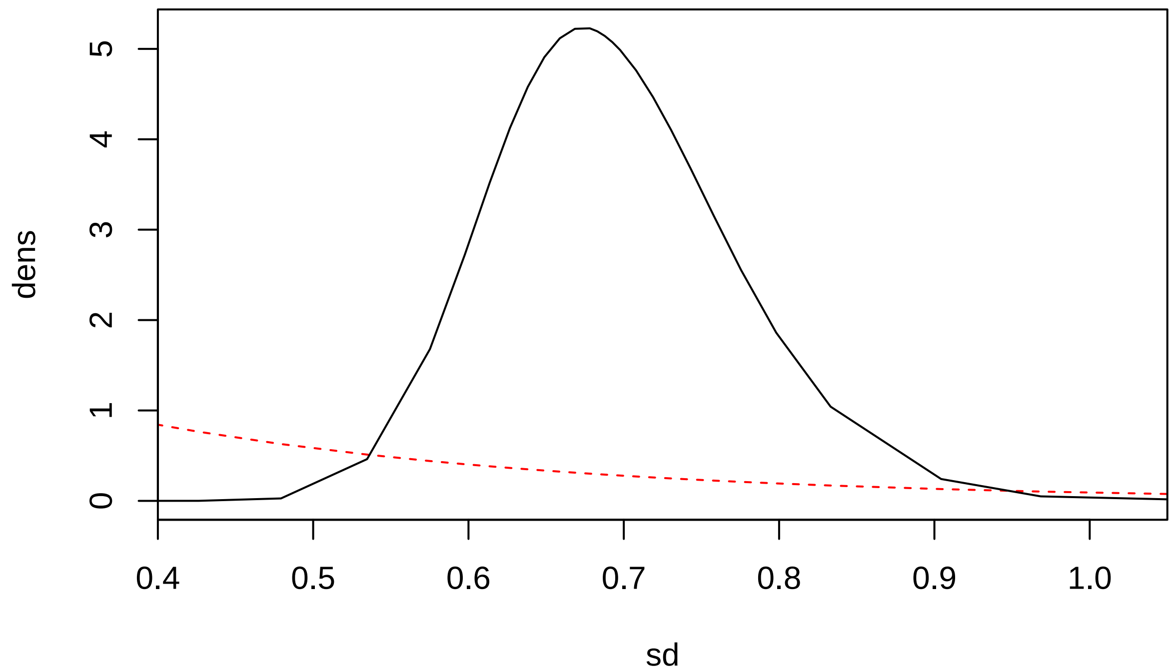



```r
# ggplot(aes(MathAchieve, x = SES,y = MathAch,group = School, color = Minority, pch = Sex)) +
#   theme_classic() + geom_point()

a = ggplot(MathAchieve, aes(x = SES,y = MathAch, group = School, color = Sex)) +
  theme_classic() +
  geom_line(alpha = 0.2) +
  geom_line(data = MathAchieve %>% group_by(SES) %>% summarise(MathAch = mean(MathAch)),
            aes(x = SES,y = MathAch,group = 1),
            colour = "Purple",
            size = 0.3) +
  labs(x="SES", y="MathAchieve")

b = ggplot(MathAchieve, aes(x = SES,y = MathAch, group = School, color = Minority)) +
  theme_classic() +
  geom_line(alpha = 0.2) +
  geom_line(data = MathAchieve %>% group_by(SES) %>% summarise(MathAch = mean(MathAch)),
            aes(x = SES,y = MathAch,group = 1),
            colour = "Purple",
            size = 0.3) +
  labs(x="SES", y="MathAchieve")

cowplot::plot_grid(a + labs(title = "MathAchieve vs SES, for each school"),b, nrow = 2)
```

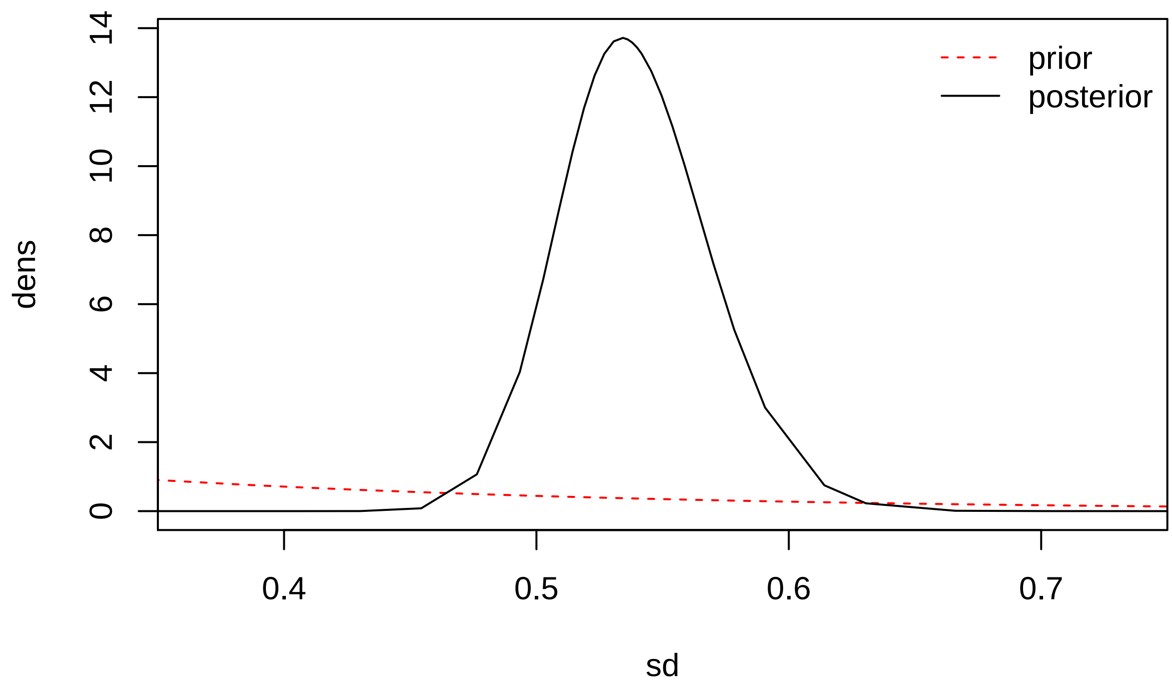





```
download.file("http://pbrown.ca/teaching/appliedstats/data/drugs.rds",
"drugs.rds")

xSub = readRDS("drugs.rds")

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

inla_formula = y ~ SUB1 + GENDER + AGE + raceEthnicity + homeless +
              f(STFIPS, model = "iid",
                prior='pc.prec',
                param=c(0.81, 0.05)) +
              f(TOWN, model = "iid",
                prior='pc.prec',
                param=c(0.63, 0.05))

ires = inla(inla_formula,
            data = forInla,
            family = 'binomial',
            control.inla = list(strategy='gaussian',
                                int.strategy='eb'))

sdState = Pmisc::priorPostSd(ires)

do.call(matplot, sdState$STFIPS$matplot)
```




```r
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_beta_mean = ires$summary.fixed[,1]
ires_beta_low = ires$summary.fixed[,3]
ires_beta_up = ires$summary.fixed[,5]

rownames(ires$summary.fixed) = c("Sub1: Opiates",
                                 "Sub1: Heroim",
                                 "Sub1: Cocaine/Crack",
                                 "Sub1: Alcohol",
                                 "Sub1: Methamphetamine",
                                 "Race: Two or more",
                                 "Race: Other",
                                 "Race: Hawaiian",
                                 "Race: Hispanic",
                                 "Race: Black",
```

Table 5: Posterior means and quantiles for model parameters.

	0.5quant	0.025quant	0.975quant
(Intercept)			
(Intercept)	0.716	0.575	0.891
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	0.917	0.994
COCAINE/CRACK	0.855	0.814	0.898
GENDER			
FEMALE	0.893	0.878	0.909
AGE18-20			
AGE18-20	0.935	0.916	0.953
AGE15-17			
AGE15-17	0.926	0.905	0.947
AGE12-14			
AGE12-14	0.972	0.934	1.012
raceEthnicity			
Hispanic	0.832	0.812	0.851
BLACK OR AFRICAN AMERICAN	0.682	0.666	0.699
AMERICAN INDIAN (OTHER TH	0.728	0.679	0.781
OTHER SINGLE RACE	0.865	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	0.845	0.748	0.953
ASIAN OR PACIFIC ISLANDER	1.454	1.227	1.723
ALASKA NATIVE (ALEUT, ESK	0.845	0.624	1.145
homeless			
TRUE	1.005	0.973	1.037
SD			
STFIPS	0.688	0.556	0.866
TOWN	0.538	0.486	0.601

```
                                  "Race: Pacific islander",
                                  "Race: Asian",
                                  "Race: American Indian",
                                  "Race: Alaska Native",
                                  "Homeless TRUE",
                                  "Gender Female",
                                  "Age 18-20",
                                  "Age 15-17",
                                  "Age 12-14",
                                  "Intercept")

ires_beta_plot = tibble(beta = ires_beta_mean,
                        coef = rownames(ires$summary.fixed),
                        cilower = ires_beta_low,
                        cilupper = ires_beta_up) %>%
  ggplot(aes( x = coef, y = beta)) +
  theme_light() +
  geom_errorbar(aes(ymin = cilower, ymax = cilupper),width = .1) +
  geom_point(pch = 21, colour = "black", fill = "red") +
  coord_flip() +
  labs(x = "Coefficient", y = "Estimate and 95% CI")

cowplot::plot_grid(
  ires_beta_plot + labs("Estimate and 95% CI")
  )
```

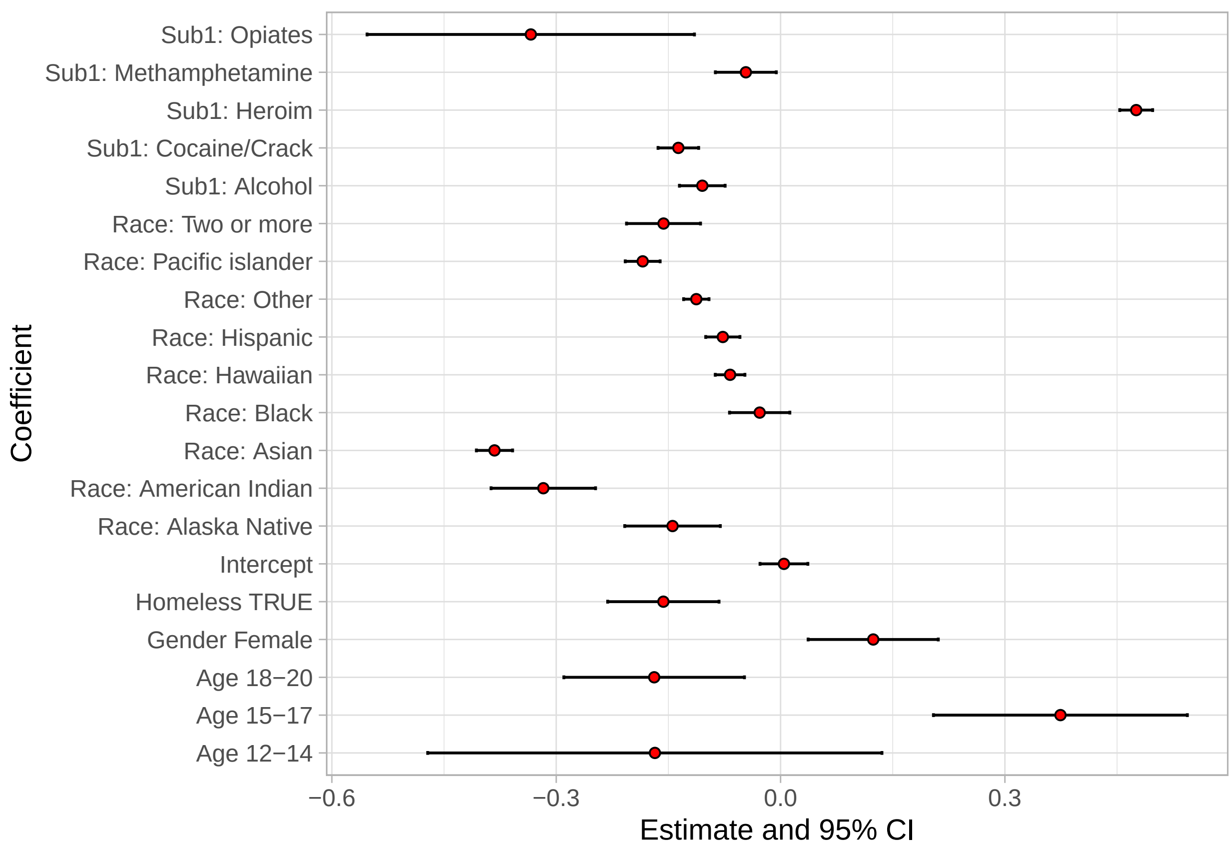


Table 6: The random effects of each US state and town

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

```r
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 = "The random effects of each US state and t
```