STA442 Assignment 2

Yiwen Yang (#1004244800)

Oct.15^{th}, 2019

1. Is High Math Score Only Achievable in Good Schools?

Setup of laboratory

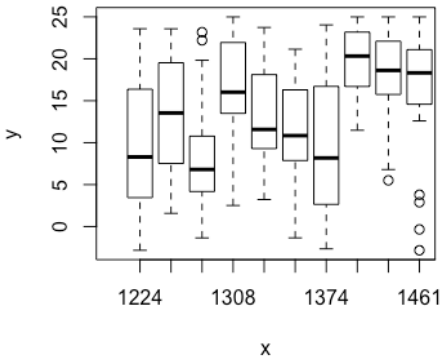
We have analyzed the MathAchieve data from the MEMSS package, where we measured the math achievement of students within each of the 160 different schools. Note that the number of math achievement scores we measured varies in numbers, in another word, the number of measurement is unequal among schools.

Our interest is to find out there are more potential differences between schools, (quality of education) or difference from performance of students within each school is larger.

Model and Interpretation

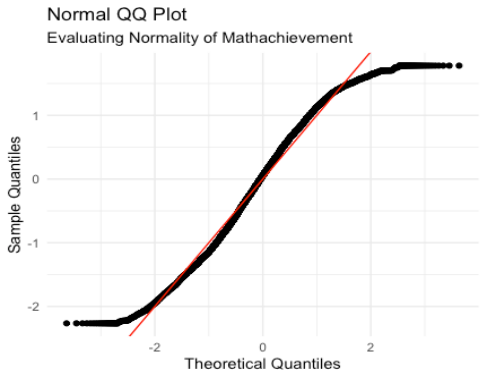
Since we would like to find out the difference in mathematical scores between specific schools, treating the variable School as a random effect is the plausible decision.

Plot 1 (Boxplot of MathAch across Different Schools)*



We first would like to take a look of math performance across schools to have a general idea of the interquartile range of data points. As we can see from the boxplot, the median and range of data varies very much across school and several data distribution within the school is quite widely spread. Note I only included 10 out of 160 schools in this plot, if all are included, the x-axis will not be observable.

Plot 2 (Evaluating Normality of MathAch)*



Usually we would fit the model and check normality of residuals, but here it's more convenient to just check the data itself because the residuals are now more complicated. From the normal Q-Q plot, there are definitely some concerns on the tail, thus p-value won't be right as it is a tail probability. However, we have to proceed anyways.

Down below is our fitted model with an additional random intercept for each school, U_i.

$$\boldsymbol{MathAch = \beta_0 + \beta_1 I_{minority\ yes} + \beta_2 I_{sex\ male} + \beta_3 \cdot SES + \beta_4\ MEANSES + U_i\ \ (i = 1\ldots,160)}$$

The model assumptions are:

$$Y_{ij}|U_i \sim^{ind} N\,(X_{ij}\beta + U_i\,, \sigma^2)$$

$$U_i \sim N(0, \sigma_u^2)$$

Y_{ij} is the math achievement for i^{th} school, j^{th} student.

X_{ij} is the fixed effect of covariates, including the intercept β_0.

U_i is individual i's (school) random effect, in another word, the between-group variability, its variance is denoted as σ_u^2.

σ^2 is the randomness within each observation, in another word, the within-group variability.

Table 1 (Estimated Fixed Effects for MathAchieve Dataset)*

	Estimated value	*Standard Error*	*t-value*	*P-value*
Intercept	*12.830*	*0.172*	*74.733*	*< 2e-16*
$I_{minority}$*(if yes, then I =1)*	*-2.731*	*0.203*	*-13.422*	*< 2e-16*
$I_{sex}(if\ male, then\ I = 1)$	*1.218*	*0.161*	*7.571*	*4.25e-14*
Social Economic Status	*1.926*	*0.108*	*17.761*	*< 2e-16*
Mean of Social Economic Status	*2.882*	*0.368*	*7.840*	*3.25e-13*

To interpret the model, $I_{minority\ yes}$, $I_{sex\ male}$ are dummy variables; and clearly, Social economic status and its mean has an effect on students' performances, so we controlled them as fixed effects. It turns out that each student has its own fitted model, which we can have better predictions. Hence, the slope for each student is assumed to be similar, but with different intercept. Surprisingly, every coefficient seems to be statistically significant by looking at their p-values.

Table 2 (Estimated Random Effects for MathAchieve Dataset)*

Group	*Variance*
School (difference of MathAch between school)	*2.443*
Residual (difference of MathAch within school)	*35.900*

To determine whether difference between school or within school is greater, we need to calculate the following. The between-group variability, in another word, the difference between school is 6% (2.443/ (2.443+35.9)), and the within-

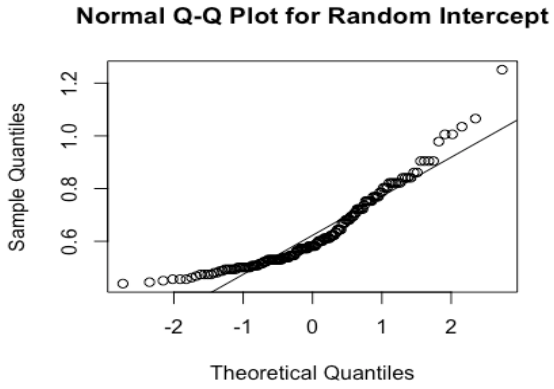
group variability, which is difference within school is 94%! Therefore, students' math performances are quite similar across schools, but the difference of scores within individual school are especially diversified.

Checking More Model Assumptions

We would like to check another assumption, the normality of U_i even though they are technically created by ourselves.

$$\boldsymbol{U_i \sim N(0, \sigma_u^2)}$$

Plot 3 (Evaluating Normality of Random Intercept for Each School)*



Not bad! The points do not deviate too much from the Q-Q line, and there is no obvious systematic deviation.

Comments

1. REML is unbiased for random effect, but it shrinks the parameter to zero.
2. The assumption of normality may not hold; therefore, the p-value may be inaccurate, resulting the parameter to be statistically insignificant. Thus, the correlation between fixed effect variables is quite strong (negatively and positively). Yet, our predominant concern is the random effect, so for fixed effect, we will just leave it as-is.

Conclusion

From our analysis, we treated the variable School as the random effect in order to determine variability of math performance of students between schools or within particular school is more distinctive. From the variance of random effects, the proportion of variance explained by difference between school is very low, only 6%; whereas the difference of performance within school explains 94%. To conclude, the dissimilarity of math achievement per student within school is much higher than between school, this is fairly reasonable since the quality of education and the material being taught is indifferent, but the level of education being received and mentality on test day varies immensely for individuals.

Appendix

Please refer to the R-code at the very last several pages of this document.

2. Analysis of Probability of Becoming Drug Addicts and The Treatment Efficacy among American States

Summary of Result

We have analyzed the result of The Treatment Episode Data Set – Discharges (http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/35074) to explore the probability for a youngster who is addicted to one of the several different drugs, to complete a treatment. We are also keen to determine the following question, are some American states have specifically effective treatment with high completion rate whereas other states have problematic treatment with very low completion rate?

Out of the six substances (Marijuana, Alcohol, Heroin, Opiates and Synthet, Methamphetamine, Cocaine) we are interested in, the odds for a youngster to complete the treatment associated with the above substances is 1.000,1.642, 0.898, 0.924, 0.982, 0.876 respectively. As we can see from these odds, except for marijuana and alcohol, all of the other substances have a completing probability lower than 1, which on a probability scale is lower than 50%, indicating they are harder to treat.

On the other hand, some American states do have particularly effective treatment such as Florida and Nebraska. In most states, the efficacy is roughly indifferent. The remaining states' treatments are very problematic, for example, Florida and Mexico have a much lower rate of completing the treatment. Some of reasons maybe because Virginia have a large proportion of marijuana addicts since many of them are ancestors of cannabis farmers. Overall, America should get more attention towards young drug addicts and states which has problematic treatments.

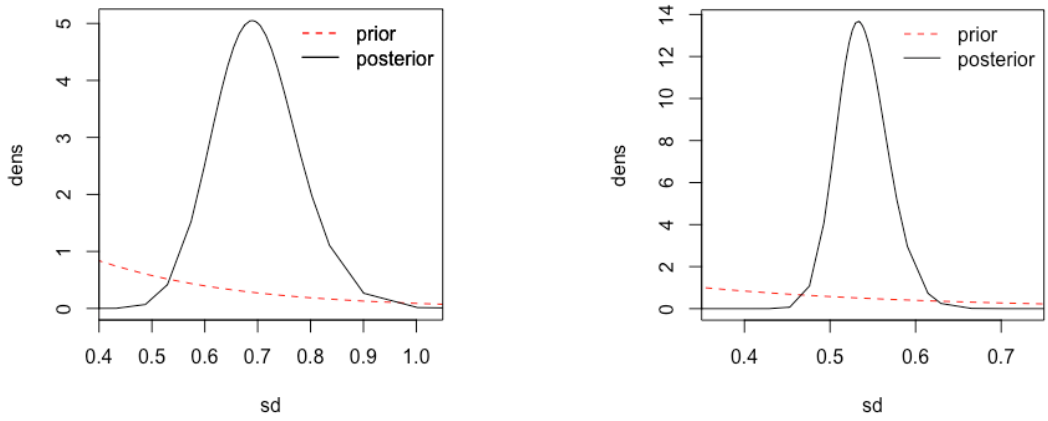
Introduction

We interpreted the result of annual discharges from substance abuse treatment from TEDS-D. The R version of the data is named as 'drugs.rds', which can be downloaded from http://pbrown.ca/teaching/appliedstats/data/drugs.rds page. Our first main focus is to find out the probability for an individual to complete a treatment who is specifically addicted to one of the following substances, marijuana, alcohol, heroin, opiates and synthet, methamphetamine and cocaine. Secondly, we would like to discover whether some American states have

more effective treatment for drugs. Furthermore, several variables are clearly fixed covariates, in particular, sub1(type of drug), gender, ethnicity and homelessness in order to receive the most accurate outcome of the analysis.

Methods and Interpretation

The following analysis is based on Bayesian inference, the reason why we choose Bayesian instead of frequentist is because all the available information is included in the model and we can see the distribution for each parameter, which are highly interpretable. At the same time, we added some external information, as known as prior, is the distribution we assume the data to have. We add this information into our analysis to have more precision. We treated STFIPS and TOWN as random effects because the treatment efficacy deviate among each other.



The two graphs from left to right are the two priors we set for σ_u and σ_v and the corresponding posteriors respectively. On the left, our prior for σ_u is interpreted as $P(\sigma_u > 0.9) = 0.05$, and on the right σ_v is $P(\sigma_v > 0.6) = 0.05$. We used the technique called **penalized complexity prior for precision.** Ideally, the prior should be given by field expert, but since we do not have this person here, we have to adjust the prior several times in order to make it reasonable on the above 'posterior and prior' graphs. Different prior will give different posterior distributions because **posterior $\propto$ likelihood $\times$ prior**, this is why choosing a good prior is particularly important. Note that, we initially have $\sigma_u \sim$ exp (30), but it not plausible at all since P ($\sigma_u > 0.1$) $\neq 0.05$. Its posterior and prior graph is in the appendix.[1]

Some properties of posteriors are, if the posterior is normally distributed, then its mean, mode and median will be the same. However, the exact distribution of posterior is usually unknown due to prior. Furthermore, posterior probability is the conditional probability given sample, so it is indeed a random variable. To evaluate its uncertainty, credible interval is more appropriate rather than confidence interval.

When we are dealing with probabilities and random effects, logistic generalized linear mixed model is the most appropriate. Down below is our fitted model.

$$\boldsymbol{logit(\lambda_{ij}) = X_{ijk}\beta + U_i + V_j \ (i = 1\ldots52,\ j = 1\ldots266,\ k = 1\ldots293772)}$$

The model assumptions are:

$$Y_{ijk} \sim^{ind} Bernoulli(\lambda_{ij})$$

$$logit(\lambda_{ij}) = X_{ijk}\beta + U_i + V_j$$

$$U_i \sim N(0, \sigma_u^2)$$

$$V_j \sim N(0, \sigma_v^2)$$

Y_{ijk} is the probability of completing the treatment for i^{th} state, j^{th} town, k^{th} person.

X_{ijk} has indicator variables for substance, gender, ethnicity and homelessness.

U_i is individual i's (state) random effect, its variance is denoted as σ_u^2.

V_j is individual j's (town) random effect, its variance is denoted as σ_v^2.

Note: $logit(\lambda_{ij})$ is equivalent to the term log odds. *K* is different for every town, so I used the total number of observation.

Priors:

$$\sigma_u \sim exp(3.32)$$

$$\sigma_v \sim exp(5)$$

Table 1 (Posterior means and quantiles for fixed effect and random effect SDs.)*

	0.5 quantile	*0.25 quantile*	*0.975 quantile*
Intercept	*0.682*	*0.546*	*0.850*
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	*0.876*	*0.834*	*0.920*

Standard deviation	*0.5 quantile*	*0.025 quantile*	*0.975 quantile*
STFIPS (American states)	*0.581*	*0.482*	*0.698*
TOWN (American towns)	*0.537*	*0.482*	*0.597*

The above data have transferred to odds, since it is directly proportional to probability, we will just interpret the odds. The two quantiles provide a measure of uncertainty in a 95 percent credible interval fashion where there is a 95% chance the true parameter will be captured by this credible interval. The substances are being treated as dummy variable with value 0 or 1. The odds of completing a treatment for a youngster who is addicted to alcohol is 1.642. Marijuana is not shown in the table as it is considered as reference group, so its mean is 1. **From the table, odds of alcohol is higher than the 'hard' drugs, so it is much easier to treat. However, odds of marijuana is indifferent to the remaining drugs, we might need extra data to determine whether it is easier to treat.** Interestingly, we had set two very different priors for states and town, but it turns out the standard deviations of posterior distributions are roughly the same.

Table 2 (Posterior means and quantiles for States as random effect.)*

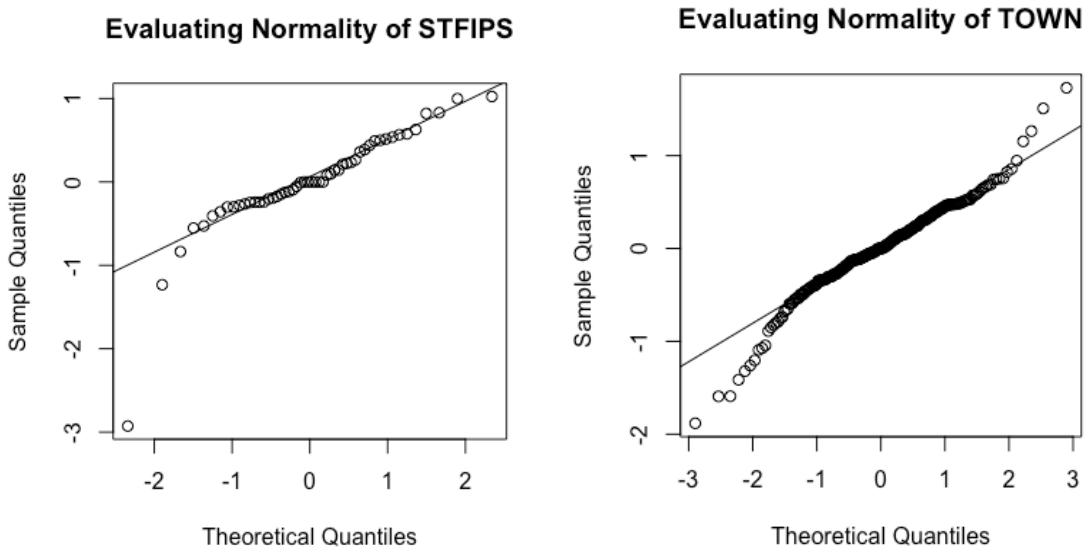
ID	mean	0.025q	0.975q	ID	mean	0.025q	0.975q
ALABAMA	0.2	-0.3	0.8	MONTANA	-0.2	-1.0	0.7
ALASKA	0.0	-0.9	0.8	NEBRASKA	0.8	0.4	1.2
ARIZONA	0.0	-1.4	1.4	NEVADA	-0.1	-0.8	0.6
ARKANSAS	-0.1	-0.7	0.5	NEW HAMPSHIRE	0.2	-0.3	0.7
CALIFORNIA	-0.3	-0.6	0.0	NEW JERSEY	0.5	0.2	0.8
COLORADO	0.5	0.1	1.0	NEW MEXICO	-1.2	-2.0	-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	-0.9	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	-1.1	0.6	PENNSYLVANIA	0.0	-1.4	1.4
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	-0.1	-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	-2.6
MASSACHUSETTS	0.8	0.4	1.3	WASHINGTON	-0.1	-0.5	0.3
MICHIGAN	-0.4	-0.7	0.0	WEST VIRGINIA	0.0	-1.4	1.4
MINNESOTA	0.4	0.0	0.9	WISCONSIN	0.0	-1.4	1.4
MISSISSIPPI	0.0	-1.4	1.4	WYOMING	0.0	-1.4	1.4
MISSOURI	-0.4	-0.7	-0.1	PUERTO RICO	0.6	-0.1	1.3

The above table is the random intercept for each state on a log odds level. Just by looking at the mean, the increase or decrease in log odds contributed by these random intercepts are approximately between -0.5 to 0.5, so the contributions to **probability** are rather trivial. However, some states such as Virginia and New Mexico has a -2.9 and -1.2 contribution to log odds correspondingly, on an odds scale, the **odds** decreased by 95% and 70%, resulting the probability of completing the treatment much lower than other states. Some states do have better treatment efficacy such as Florida and Nebraska, their contribution to **odd** is roughly 170% and 125% percent in increase. For a discrete example, a Hispanic female who is addicted to heroin has a probability of 70% of completing the treatment, but if she is in Virginia, the probability dramatically dropped to 12%. If she is in Florida, the probability increases to 87%.

Checking More Model Assumption

$$U_i \sim N(0, \sigma_u^2) \qquad\qquad V_j \sim N(0, \sigma_v^2)$$

Plot 3 (Evaluating Normality for STFIPS and TOWN)*



As we can see, both of the normal Q-Q plot for the random intercept have a generally plausible fit, where the data points are not far away from the qqline, and there are no systemic deviations either. For the plot on the right, there are some concerns on the left side of the tail, but overall, a good fit.

Comments and Limitation

1. I only included substances and standard deviation of random effects in Table*1. Please refer to the full table in Appendix for more information.[2]
2. INLA is only able to give marginal posterior distribution, but joint posterior cannot be given.
3. Even though REML is unbiased for random effect, but since this dataset is large and it contains many fixed effect parameters (losing 1 degree of freedom for each fixed effect parameter being estimated), the parameter will suffer shrinkage and become biased. Furthermore, REML is not suitable for model comparison.

Conclusion

Our predominant interest is to discover the probability of completing treatment for different drugs as well as to discover the difference of treatment efficacy across American states. We treated the American states as a random effect and added external information to have more precision in the analysis. By Bayesian inference, hard drugs are indeed more difficult to treat compare to Alcohol, but marijuana is as difficult to treat, surprisingly, probably due to the fact they are easier to obtain. The treatment efficacy varies across different states, most of them are indifferent in terms of treatment efficacy, but some of them such as Florida has outstanding treatments, which leads to higher completing rate. However, for Virginia, the completing rate for an addict dropped by roughly 50% in general, it is quite shocking for a wealth state to have such a problematic treatment for drug addicts.

Reference

1. Faraway, Julian J, Xiao feng Wang, and Yu Yue Ryan (2018). Bayesian Regression Modeling with INLA. Chapman and Hall/CRC.
2. Patrick Brown. http://pbrown.ca/teaching/appliedstats/slides/bayes.pdf

Appendix

Code for Analyzing Math Achievement

```
install.packages("lme4")
library(lme4)
install.packages('nlme')
library(nlme)
install.packages("rmarkdown")
library(rmarkdown)
install.packages('MEMSS')


data("MathAchieve", package = "MEMSS")
head(MathAchieve)
Math = tibble(MathAchieve)

plot(MathAchieve$School,MathAchieve$MathAch,xlim= c(0,10.1))
# Here, the label of school is obsecured, which is a good thing because the label doesn't have a
# mathematical meaning.
# It looks like the variance is roughly constant.
```

```
# Check the normality of MathAch.
MathAchieve%>%
  dplyr::select(MathAch) %>%
  arrange(MathAch) %>%
  mutate_at("MathAch",funs( (. - mean(.)) / sd(.))) %>%
  mutate(q = qnorm(seq(1:7185) / (1 + 7185))) %>%
  ggplot(aes(x = q,y = MathAch)) +
  theme_minimal() +
  geom_point() +
  geom_abline(slope = 1,intercept = 0,colour = "red") +
  labs(title = "Normal QQ Plot",
       subtitle = "Evaluating Normality of Mathachievement",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles")


# random intercept
Mixed_model_1 glmmPQL(MathAch ~as.factor(Minority)+ as.factor(Sex) + SES + MEANSES,
                      random = ~1|School,family = 'gaussian',data = MathAchieve)
knitr::kable(summary(Mixed_model_1)$coef, digits=3)


# Test normality of Ui
random_intercept<- attr(ranef(Mixed_model_1, condVar = TRUE)[[1]], "postVar")
qqnorm(random_intercept,main='Normal Q-Q Plot for Random Intercept')
qqline(random_intercept)
```

Code for Analyzing drug treatment and effectiveness across different American states

```
download.file("http://pbrown.ca/teaching/appliedstats/data/drugs.rds", "drugs.rds")
install.packages("Pmisc", repos="http://R-Forge.R-project.org")
install.packages('INLA',repos=c(getOption('repos'),INLA= 'https//inla.r-inla-download.org/R/stable'),dept= TRUE)
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)

library("Pmisc")
library("INLA")
ires = inla(y ~ SUB1 + GENDER + raceEthnicity + homeless +
              f(STFIPS, hyper=list(prec=list( prior='pc.prec', param=c(0.9, 0.05)))) +
              f(TOWN,hyper=list(prec=list( prior='pc.prec', param=c(0.6, 0.05)))),
            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, ".*"),
```

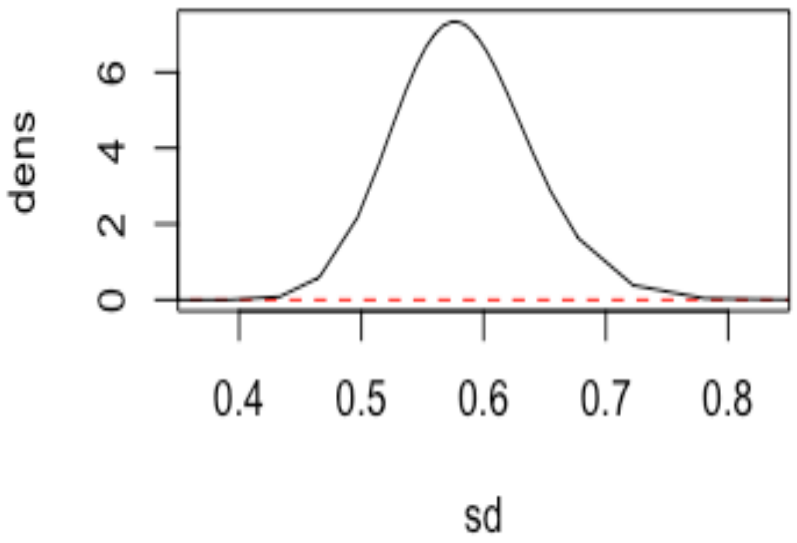
```
"\\1", rownames(toPrint)), category = substr(gsub(sss,
"", rownames(toPrint)), 1, 25), toPrint)
Pmisc::mdTable(toPrint, digits = 3, mdToTex = TRUE,
  guessGroup = TRUE, caption = "Posterior means and quantiles for model parameters.")


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")

random_intercept_Ui = ires$summary.random$STFIPS['mean']
Ui= unlist(random_intercept_Ui,use.names=FALSE)
qqnorm(Ui,main = 'Evaluating Normality of STFIPS')
qqline(Ui)

random_intercept_Vj = ires$summary.random$TOWN['mean']
Vj= unlist(random_intercept_Vj,use.names=FALSE)
qqnorm(Vj,main= 'Evaluating Normality of TOWN')
qqline(Vj)
```

(1) Initial prior



(2) Fixed effect full table

	mean	0.025q	0.975q
Intercept	0.681	0.546	0.850
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
Ethnicity Group			
Hispanic	0.829	0.810	0.849
BLACK OR AFRICAN AMERICAN	0.685	0.669	0.702
AMERICAN INDIAN (OTHER TH	0.729	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	0.846	0.750	0.955
ASIAN OR PACIFIC ISLANDER	1.451	1.225	1.720
ALASKA NATIVE (ALEUT, ESK	0.844	0.623	1.144
Homeless			
True	1.015	0.983	1.048
SD			
STFIPS	0.694	0.557	0.868
TOWN	0.537	0.484	0.599