STA442 HW2

Depeng Ye 1002079500

05/10/2019

1. Math

Introduction

The MathAchieve dataset comes from the MEMSS package has 7185 observations (students) on 6 catagories of information including "School", "Minority", "Sex", "SES", "MathAch". and "MEANSES". This dataset was collected to analysis how well the students will perform in their Math Achievement considering the 5 investigated aspects. The study was designed and performed with the hypothesis that the factor "School" is a random effect when fitting a model of students' Mathematics Achievement with "Sex", "Minority", and "SES" as fixed effects. Namely, our study is trying to investigate the reliability of the statement: different students from identical schools have a substantial distinguish in Math Avhievements. The assumptions are that "Minority" (levels yes or no) is factor, "SES"(socio-economic status) is fixed effect, and we could also consider "Sex" as a fixed effect.

Method

We want to fit a linear mixed effect model, which regards school as a random effect, to see how the result of such random model tells us.
Before fitting a linear mixed effect model, we need to test the normality. Two normality tests (Jarque-Bera Test and Anderson-Darling test) was performed in the appendix. Both of the p-value were less than 2.2×10^{-16} which is significant enough to say that "MathAch" follows normal distribution.
The method used to fit the model is the lme function in the nlme package. We are fitting a model of "MathAch" regarding factor "Minority" and variable "SES" to see how significnat the influence of "school". The mathematical formula of the model is provided as follows:

$$Y_{ij} = X_{ij}\beta_i + U_i + Z_{ij}$$

where Y_{ij} refers to "MathAch"; X_{ij} refers to fixed effect "Minority", "SES" and "Sex" with coefficients β_i; $U_i \sim N(0, \sigma^2)$ is the random effect "School"; and Z_{ij} refers to random error. All vectors are three dimensional.

Result

Results of the linear mixed effect model has been attached to the appendix. Based on the result, we want to determine whether or not "School" is a random effect. Hence we want to compare the variance of students' Math Achievements within and across schools. According to Table 2, the variance of students' Math Achievement within the same school (represented by σ) is approximately 5.99 while the variance between different schools (represented by τ) is 1.91. It is obvious that the variance of Math Achievements within schools is significantly higher than the variance of students' Math Achievement between different schools. A plot of the fitted model is attached, though it might seem messy and not quite useful.

Conclusion

Considering the results of data analyzed in the section above, we could draw a conclusion that there is a substantial volatility in Math Achievement within school than between different schools. Hence, it is a proper approach to make factor "School" as a random effect in this study we have performed.

Appendix: Code, Tables and Plots

```
data("MathAchieve", package = "MEMSS")
knitr::kable(head(MathAchieve), caption = "Overview of MathAchieve Data")
```

Table 1: Overview of MathAchieve Data

School	Minority	Sex	SES	MathAch	MEANSES
1224	No	Female	-1.528	5.876	-0.428
1224	No	Female	-0.588	19.708	-0.428
1224	No	Male	-0.528	20.349	-0.428
1224	No	Male	-0.668	8.781	-0.428
1224	No	Male	-0.158	17.898	-0.428
1224	No	Male	0.022	4.583	-0.428

```
#Normality Testing
jarque.bera.test(MathAchieve$MathAch)
```

```
##
##  Jarque Bera Test
##
## data:  MathAchieve$MathAch
## X-squared = 292.98, df = 2, p-value < 2.2e-16
```

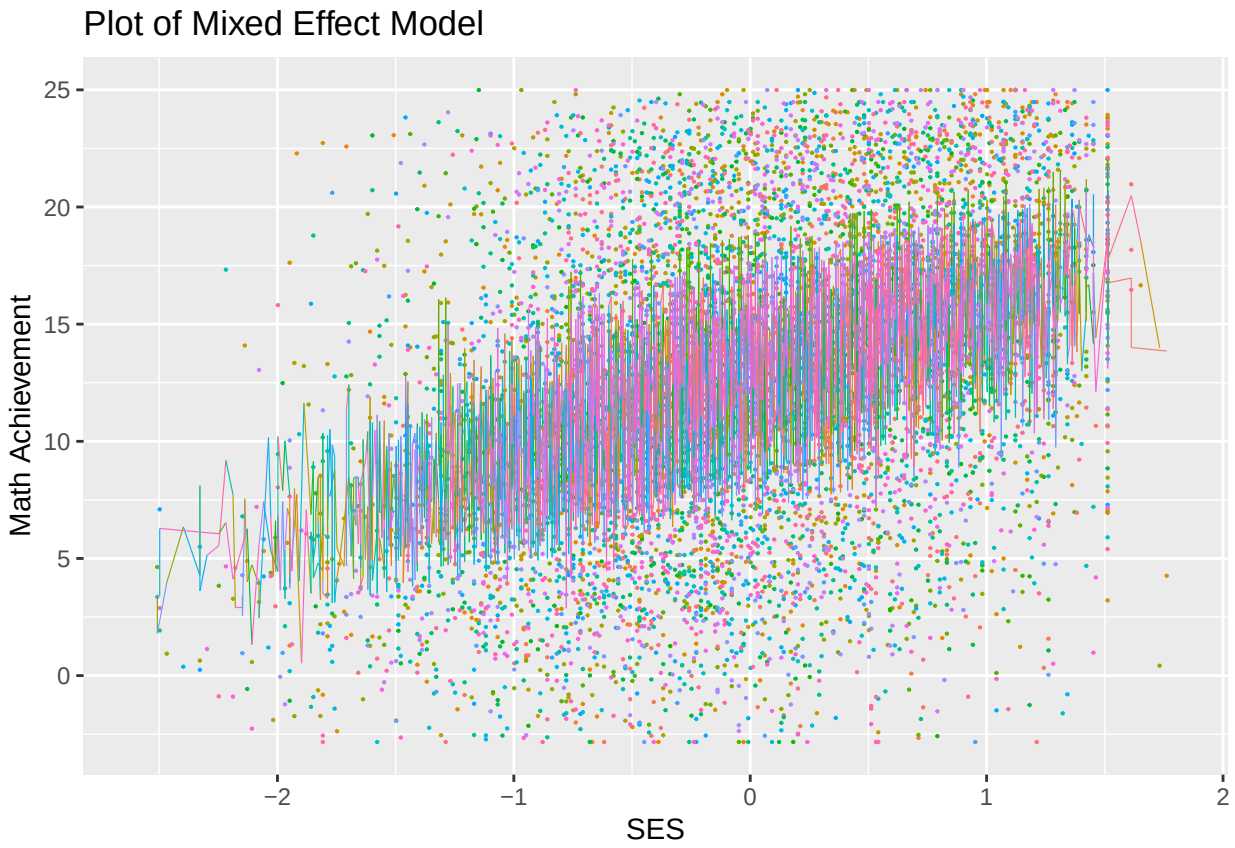
```
ad.test(MathAchieve$MathAch)
```

```
##
##  Anderson-Darling normality test
##
## data:  MathAchieve$MathAch
## A = 43.272, p-value < 2.2e-16
```

```
#Fit the LME and print the summary
MathFit = lme(MathAch ~ Minority + SES + Sex, random = ~1|School, method = "REML",
              data = MathAchieve)
knitr::kable(Pmisc::lmeTable(MathFit), digits = 3, caption = "Summary of Linear
             Mixed Effect Model")
```

Table 2: Summary of Linear Mixed Effect Model

	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


Plot of Mixed Effect Model
Math Achievement
25
20
15
10
5
0
−2
−1
0
1
2
SES

2. Drug

Introduction

source: https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/35074
We are having a dataset TEDS-D(Treatment Epiode Data Set - Discharge), which is a national census data system of annual discharges from absuse treatment facilities throughout the U.S. It provides annual data on the number and characteristic of persons discharged from public and private substance abuse treatment programs that received public fundings. This data has 422478 observations of 9 variables including "completed", "SUB1"(the type of drug taken), "GENDER", "AGE", "STFIPS"(the state subject is located in), "raceEthnicity", "homeless", EDUC"(education), and"TOWN".
Our first goal is to investigate the validity of two statements, the first one is:" 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". The second statement is:"Some American states have particularly effective treatment programs whereas other states have programs which are highly problematic with very low completion rates."

Method

Generally INLA is used to measure how well the two statements are made. We processed the "completed" variable of the data using 1 for TRUE and 0 for FALSE, labled the new column "y".

First statement: Want to test whether harder drugs are harder to treat than alcohol and marijuana. We use log-odds:

$$\log \frac{\mu}{1-\mu} = \sum_{j=1}^{5} X_{ij}\beta_j$$

. Fitted the model using INLA package: fitting a model ires1 of "y" with respect to fixed effect "GENDER", "AGE", "raceEthniity", and "SUB1", and random effect "STFIPS". INLA model is

$$logit(\lambda_{it}) = \eta_{it} = X_{it}\beta + U_i$$

where U_i is the random effect STFIPS (i.e. states) and $X_{it}\beta$s are refering to different drug the subjects are addicted to. The prior used in this INLA model is $c(0.1, 0.05)$ which is given in the instruction of this assignment.

Second statement: In the second statement there are two random effects: "STFIPS" (states) and "TOWN". Hence we need to fit a INLA model with two random effects. In the appendix we have fixed a INLA named ires2 with the same fixed effect as before in ires1, but two Random effects "STFIPS" and "TOWN". THe INLA model is $logit(\lambda_{it}) = \eta_{it} = X_{it}\beta + U_i + V_i$ where U_i is the random effect STFIPS, and V_i refers to the random effect TOWN. We used the penalized complexity prior of $c(0.77, 0.05)$ for a proper fit meaning that $P(\sigma_u > 0.77) = 0.05$. A table inlcuding informations of the effect of drug treatment in each states were included as Table7 in the Appendix.

Result

First statement: With the fitted INLA model ires1 we generated Table 6 as the summary of posterior means and quantiles for model parameters. Because Marijuana is used as a reference group, then 0.5 quantile of Marijuana is 1. Also notice that alcohol and marijuana have the 0.5 quantile greater than or equal to 1 while others are less than 1. As a result, the odds of successfully treat a subject addicted to alcohol or marijuana will be a lot higher than successful treatment done on any people who is addicted to "harder" drugs.

Second statement: Based on Table 7, there are different means in different states. For example, in states like Alabama, Colorado, Texas, Utah, etc. the mean of the effect is positive. While in states like Nevada, New Mexico, Virginia, Michigan, etc. the average effect is negative. It is also worth to mention that there are some states, for example, Alaska, Wisconsin, Mississippi, etc. the mean effect of drug treatment is zero.

Conclusion

First statemnt: It is easy to conclude from the table and the analysis in the **Result** part that Alcohol and Marijuana are two kind os addiction that are a lot easier to get rid of when proper treatment has been applied. The other "hard" drugs when compared to these two, are more difficult to treat.

Second statement: Based on the mean of effects that are illustrated in the **Result** section, it is safe for us to conclude that some states in the U.S. are having effective treatment programs while others might not be so effective. Even further, there are states that are having negatively effective treatment programs which leads to an increase in the involved population of drug abuse. Some actions need to be done to fix such problem in the future.

Appendix: Code, Tables and Plots

```
xSub = readRDS("drugs.rds")

knitr::kable(table(xSub$SUB1), caption = "Overview of variable SUB1")
```

Table 3: Overview of variable SUB1

Var1	Freq
(4) MARIJUANA/HASHISH	188406
(2) ALCOHOL	97013
(5) HEROIN	58511
(7) OTHER OPIATES AND SYNTHETICS	45609
(10) METHAMPHETAMINE	21606
(3) COCAINE/CRACK	11333

```
knitr::kable(table(xSub$STFIPS)[1:5], caption = "Overview of variable STFIPS")
```

Table 4: Overview of variable STFIPS

Var1	Freq
(1) ALABAMA	616
(2) ALASKA	1360
(4) ARIZONA	4479
(5) ARKANSAS	1508
(6) CALIFORNIA	48065

```
knitr::kable(table(xSub$TOWN)[1:2], caption = "Overview of variable TOWN")
```

Table 5: Overview of variable TOWN

Var1	Freq
ABILENE, TX	42

Var1	Freq
AKRON, OH	1078

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

library("INLA")
```

```
## Loading required package: Matrix
```

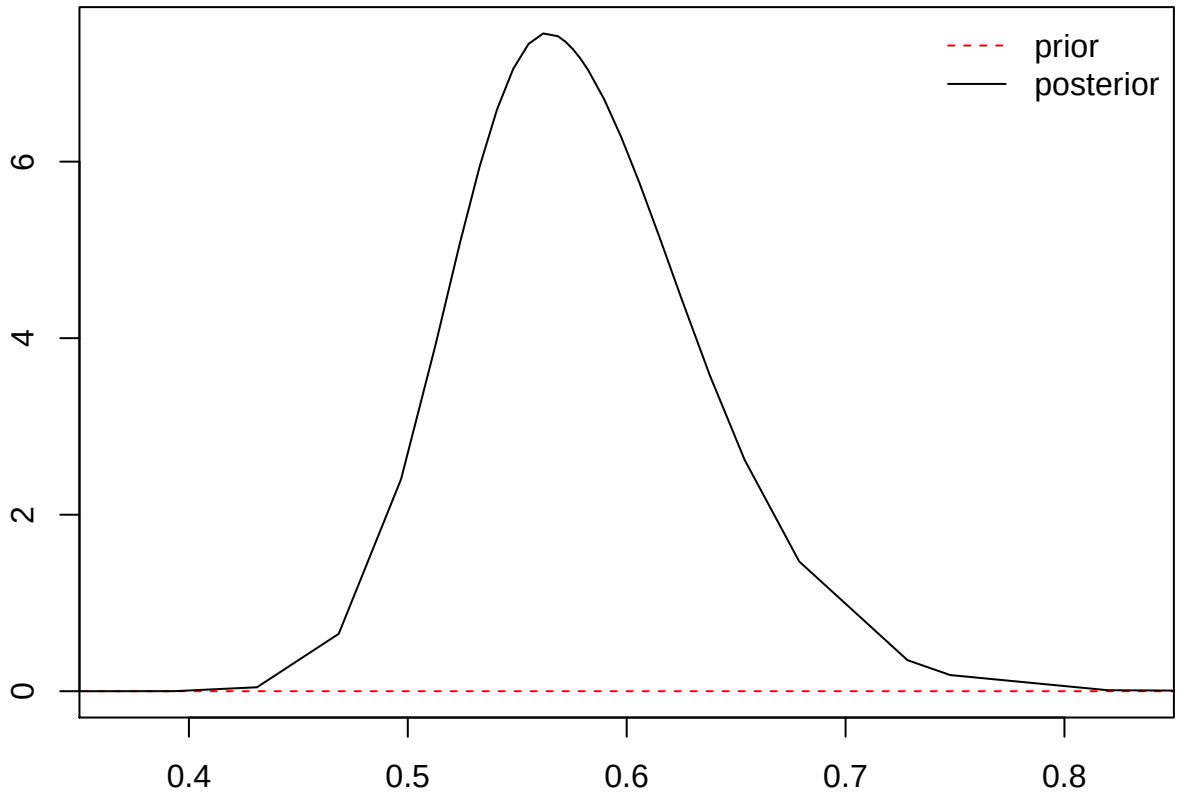
```
## Loading required package: sp
```

```
## Loading required package: parallel
```

```
## This is INLA_19.09.03 built 2019-09-03 09:07:31 UTC.
## See www.r-inla.org/contact-us for how to get help.
## To enable PARDISO sparse library; see inla.pardiso()
```

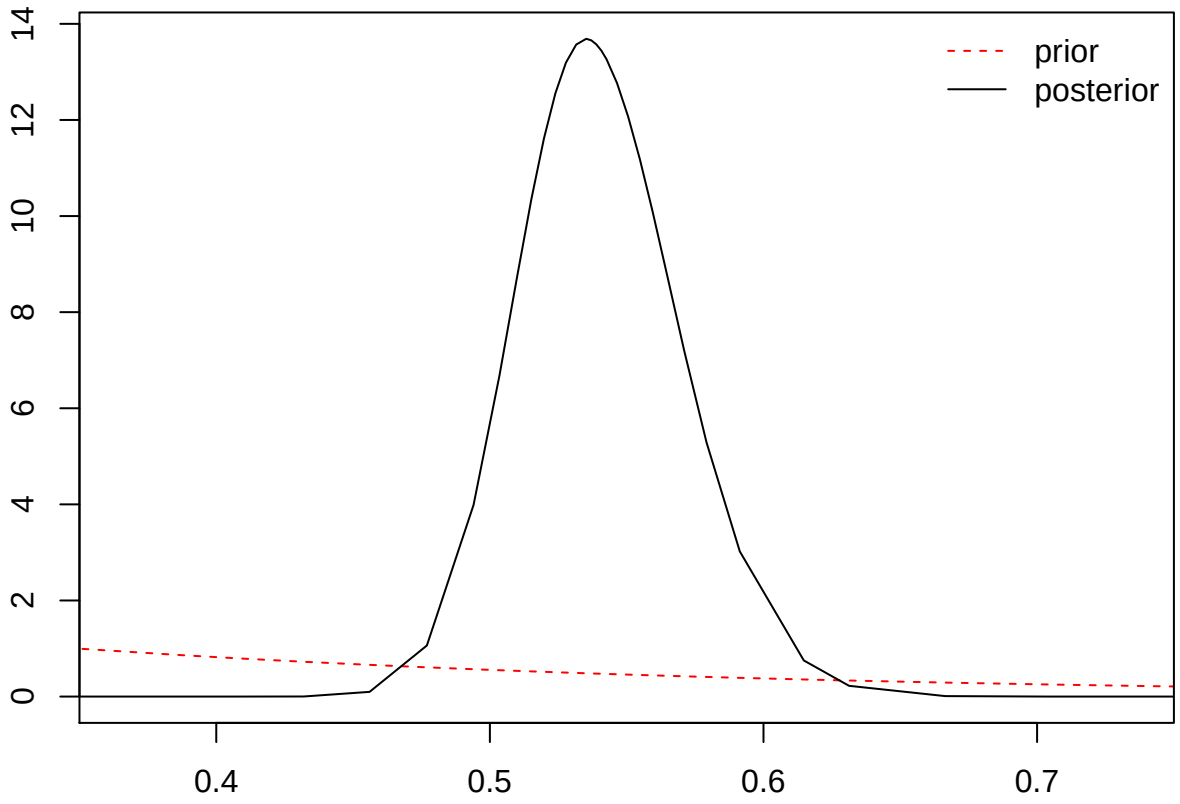
```
ires1 = inla(y ~ SUB1 + GENDER + raceEthnicity + homeless + AGE +
               f(STFIPS, hyper=list(prec=list(
                 prior='pc.prec', param=c(0.1, 0.05)))) +
               f(TOWN),
             data=forInla, family='binomial',
             control.inla = list(strategy='gaussian', int.strategy='eb'),
             control.family = list(link = "logit"))


sdState = Pmisc::priorPostSd(ires1)
par(mar = rep(2,4))
do.call(matplot, sdState$STFIPS$matplot)
do.call(legend, sdState$legend)
```




```
ires2 = inla(y ~ SUB1 + GENDER + raceEthnicity + homeless + AGE +
               f(STFIPS, hyper=list(prec=list(
                 prior='pc.prec', param=c(0.77, 0.05)))) +
               f(TOWN,hyper=list(prec=list(
                 prior='pc.prec', param=c(0.77, 0.05)))),
             data=forInla, family='binomial',
             control.inla = list(strategy='gaussian', int.strategy='eb'),
             control.family = list(link = "logit"))

sdState1 = Pmisc::priorPostSd(ires2)
do.call(matplot, sdState1$TOWN$matplot)
do.call(legend, sdState1$legend)
```




```
toPrint = as.data.frame(rbind(exp(ires1$summary.fixed[,
                                                    c(4, 3, 5)]),
                              sdState$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.")
```

```
ires1$summary.random$STFIPS$ID = gsub("[[:punct:]]|[[:digit:]]",
                                      "", ires1$summary.random$STFIPS$ID)
ires1$summary.random$STFIPS$ID = gsub("DISTRICT OF COLUMBIA",
                                      "WASHINGTON DC", ires1$summary.random$STFIPS$ID)
toprint = cbind(ires1$summary.random$STFIPS[1:26, c(1, 2, 4, 6)],
                ires1$summary.random$STFIPS[-(1:26), c(1, 2, 4, 6)])

colnames(toprint) = gsub("uant", "", colnames(toprint))
knitr::kable(toprint, digits = 1, format = "latex",
             caption = "means and quantiles of random effect STFIPS (State)")
```

Table 6: Posterior means and quantiles for model parameters.

	0.5quant	0.025quant	0.975quant
(Intercept)			
(Intercept)	0.717	0.593	0.865
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	0.899
GENDER			
FEMALE	0.893	0.878	0.909
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	0.793	0.921
ASIAN	1.132	1.037	1.235
NATIVE HAWAIIAN OR OTHER	0.845	0.749	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
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
SD			
STFIPS	0.575	0.483	0.701
TOWN	0.535	0.484	0.603

Table 7: means and quantiles of random effect STFIPS (State)

ID	mean	0.025q	0.975q	ID	mean	0.025q	0.975q
ALABAMA	0.2	-0.3	0.7	MONTANA	-0.2	-0.9	0.6
ALASKA	0.0	-0.8	0.8	NEBRASKA	0.8	0.4	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	-0.3	0.6
CALIFORNIA	-0.3	-0.5	0.0	NEW JERSEY	0.5	0.2	0.8
COLORADO	0.5	0.1	0.9	NEW MEXICO	-1.1	-1.8	-0.4
CONNECTICUT	0.1	-0.4	0.6	NEW YORK	-0.3	-0.6	0.0
DELAWARE	1.0	0.7	1.3	NORTH CAROLINA	-0.8	-1.1	-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	1.1
HAWAII	0.2	-0.6	1.0	OREGON	0.1	-0.2	0.4
IDAHO	-0.2	-0.9	0.6	PENNSYLVANIA	0.0	-1.1	1.1
ILLINOIS	-0.5	-0.8	-0.2	RHODE ISLAND	-0.2	-0.6	0.2
INDIANA	0.0	-0.8	0.7	SOUTH CAROLINA	0.4	0.1	0.7
IOWA	0.4	0.1	0.7	SOUTH DAKOTA	0.4	-0.3	1.2
KANSAS	-0.2	-0.5	0.1	TENNESSEE	0.3	-0.2	0.7
KENTUCKY	-0.2	-0.5	0.2	TEXAS	0.6	0.3	0.9
LOUISIANA	-0.5	-0.9	-0.1	UTAH	0.1	-0.5	0.6
MAINE	0.1	-0.6	0.9	VERMONT	-0.2	-1.0	0.6
MARYLAND	0.5	0.2	0.8	VIRGINIA	-2.9	-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	0.5	-0.1	1.2