COVID-19 Mortality Analysis for 50 States

Dongdong Lu

1 Introduction

This year has seen the tragedy of COVID-19 pandemic. By the time of December 6, more than 270, 000 people has died from it just in United States. It has been noticeable within the US, different state has significant differences in their geographics, social economics, politics, and social welfare. Some states are richer, and some are poorer. Some are blue and some are red. Some are colder and some are warmer. The comparison goes on and on. So it is natural to ask how would these demographic factors affecting the mortality rate of COVID-19 for each state? I did the search on google scholar and have found almost all the mortality analysis are based on patient's physical conditions. And no such research on the state level has been published.

In this project, I first compile the economic, climate, geographic, population, and mortality data from governmental websites into one single big dataset. Then use "rstanarm" package to analyze 7 different models. The first 4 models are using just the poverty and political affiliation. Then the last 3 models are using comprehensive spectrums of covariates. Finally I discussed my findings about relationship between mortality and several key covariates such as population density, state politics, temperature and summarized the directions for future work.

2 Dataset

Dataset	Resource
WHO Flumart Output	World's Health Organization
COVID-19 Mortality	Centers for Disease Control and Prevention
2020 Presidential Election Results	New York Times
Temperature, Latitude and Longitude Data for 50 states	Climate.GOV
2020 US State by Race	World Population Review
Gross Domestic Product (GDP) Summary	Bureau of Economic Analysis
Crime Data Explorer (CDE)	Federal Bureau of Investigation

Table 1 Datasets and Resources

RESPONSE VARIABLE	PREDICTORS
COVID MORTALITY RATE PER 100, 000 RESIDENTS (FOR EACH STATE)	population density, GDP per capita, Biden's support rate in 2020 presidential election, Latitude and longitude of the state's capitol, White and asian percentage, poverty rates, violent crime rate, the state's average elevation and elevation

Table 2 Response Variable and Predictors

3 Data Exploration

Here is another graph showing how the mortality rate is increasing along with the poverty rate. Notice that we have 3 blue states: MA, NJ, and NY with very high mortality rate. If those states are removed. The density distribution will look more like normal distribution.

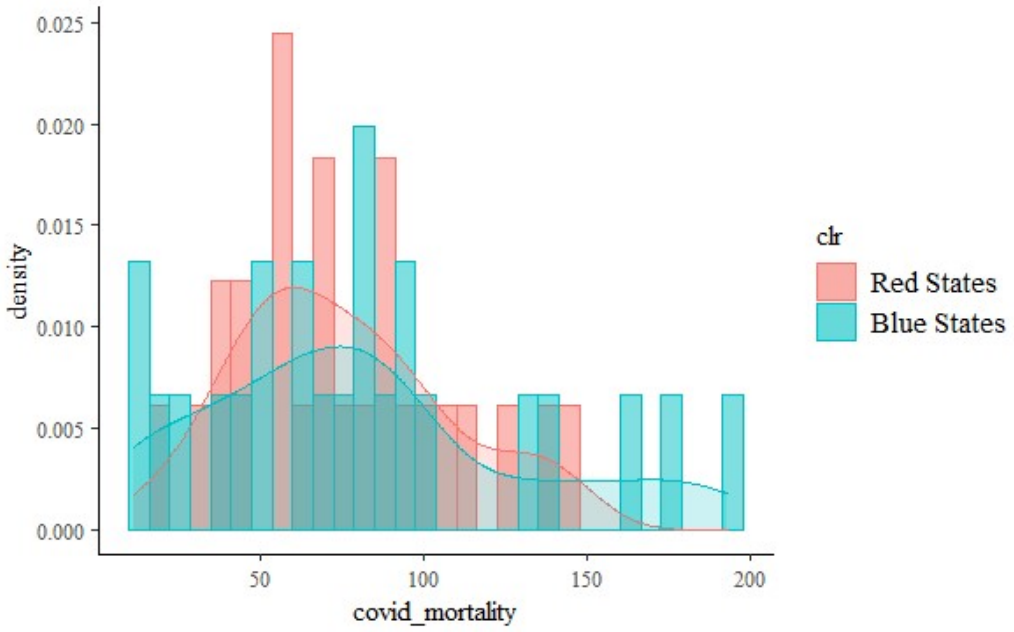


Figure 1 Histogram for Each State's Mortality

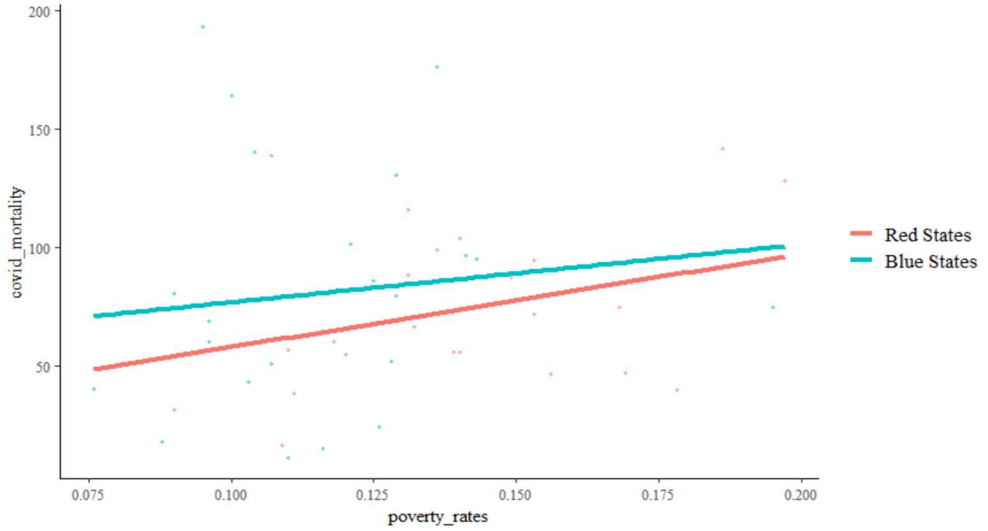


Figure 2 Scatterplots of Mortality for Blue and Red States

Three observations: 1. COVID-19 mortality rate tend to increase along the poverty rate. The medical service is very expensive in the US. Health care inequality follows systematically from economic inequality. 2. Blue states have a higher mortality rate than red states. This might be because the COVID-19 hit those blue states first and blue states typically have higher population density than red states which made the transmission easier. 3. The mortality rate is increasing faster along with poverty rate for red states. Why is this true? The answer may be found in the following quote:

> "*Programs such as Medicaid are governed by each state,* ***allowing more conservative states to limit access to lifesaving coverage based on income levels*** *...*
>
> – Healthcare in America (published before COVID)

4 Method

I am using the Bayesian normal regression to study what attributes of each state affect the mortality rate.

$$Y \sim \beta_0 + X_1\beta_1 + X_2\beta_2 + \cdots + X_p\beta_p + \epsilon,\ \epsilon \sim N(0, \sigma^2).$$

Unlike the frequentist approach, the Bayesian approach requires us to provide our own prior beliefs about the distributions. To not bring my personal bias, I used the default non-informative priors and the "rstanarm" package has automatically adjusted the priors to ones with better performance.

Our response variable is the mortality rate and our predictor variables are basic demographic information. I assume a normal distribution for the response since the data exploration suggested the real distribution might be very similar with normal distribution. The long tail of the mortality might be a good hint of using "log" transformation, however this will add the difficulty to interpretation. Therefore, after weighing potential gains over losses, I choose not to use the "log" regression. By default setting of "rstanarm", the program runs 4 chains with 2000 iterations.

For racial information, I did not add black or Hispanic population since adding them would give me collinearity issues for the numerical calculation, since the population ratio adds up to one. However, information regarding black or Hispanic population can be inferred from their white/asian counterpart.

4.1 Model Components

MODEL	PREDICTORS & USAGE
1	red
2	Poverty_rates
3	red + poverty_rates
4	red * poverty_rates
5	popu_density + gdp + biden_rates + latitude + longitude + white + asian + violent_crates + flu_mortality + elevation + temp +red * poverty_rates (using all predictors)
6	popu_density + gdp + biden_rates + latitude + white + asian + flu_mortality + temp +red * poverty_rates (deleting longitude, violent_crates, and elevation)
7	popu_density + biden_rates + latitude + white + asian + temp +red * poverty_rates (deleting flu_mortality and gdp)

Table 3 Predictors' Usage for 7 Models

4.2 Priors (for the final model 7)

	MEAN	DEFUALT	ADJUSTED
$\boldsymbol{\beta_0}$: intecept	0	2.5	3.942407e-01
$\boldsymbol{\beta_1}$: population density	0	2.5	9.411657e-03
$\boldsymbol{\beta_2}$: Biden support rate	0	2.5	1.818917e+01
$\boldsymbol{\beta_3}$: Latitude	0	2.5	8.170732e+02
$\boldsymbol{\beta_4}$: Percent of White	0	2.5	1.881733e+03
$\boldsymbol{\beta_5}$: Percent of Asian	0	2.5	2.156467e+01
$\boldsymbol{\beta_6}$: Average Temp	0	2.5	2.064025e+02
$\boldsymbol{\beta_7}$: Poverty Rates	0	2.5	3.680630e+03
$\boldsymbol{\beta_8}$: Red*Poverty	0	2.5	1.422794e+03

Table 4 Default Priors and Adjusted Priors

4.3 Model Comparison

We compare our 7 contending models using an approximation to Leave-One-Out (LOO) cross-validation by implementing the "loo" function in the **loo** package.

From the below result, we can know the seventh model is preferred as it has the highest expected log predicted density (elpd_loo) and lowest value of the LOO Information criterion (looic). Since the difference is significantly larger than the standard error, the preference of the seventh model over others is strong.

MODEL	ELPD_DIFF	SE_DIFF
7	0.0	0.0
6	-1.3	0.6
5	-6.9	0.9
2	-14.4	7.6
3	-14.5	7.3
1	-14.6	7.0
4	-15.5	7.3

Table 5 Expected Log Predicted Density and Standard Variances

4.4 Convergence

All the “RHAT” values converged to 1 which implies the MCMC convergence.

	MCSE	RHAT	EFFICIENT SAMPLES
(INTERCEPT)	2.5	1	2799
POPU_DENSITY	0	1	3182
BIDEN_RATES	0	1	3626
LATITUDE	0	1	2757
WHITE	1.1	1	2941
ASIAN	2.6	1	2754
TEMP	0	1	2491
RED	1	1	1995
POVERTY_RATES	5.6	1	2433
RED:POVERTY_RATES	7.3	1	1935
SIGMA	0.1	1	2852
MEAN_PPD	0.1	1	4004
LOG-POSTERIOR	0.1	1	1393

Table 6 Rhat Values and Efficient Samples

4.5 Posterior Checks

As we can see from the below graph, the replicated data look very similar with the actual values. This is implying that no serious problem found from this perspective. The density curve is also very similar withe the actual data.

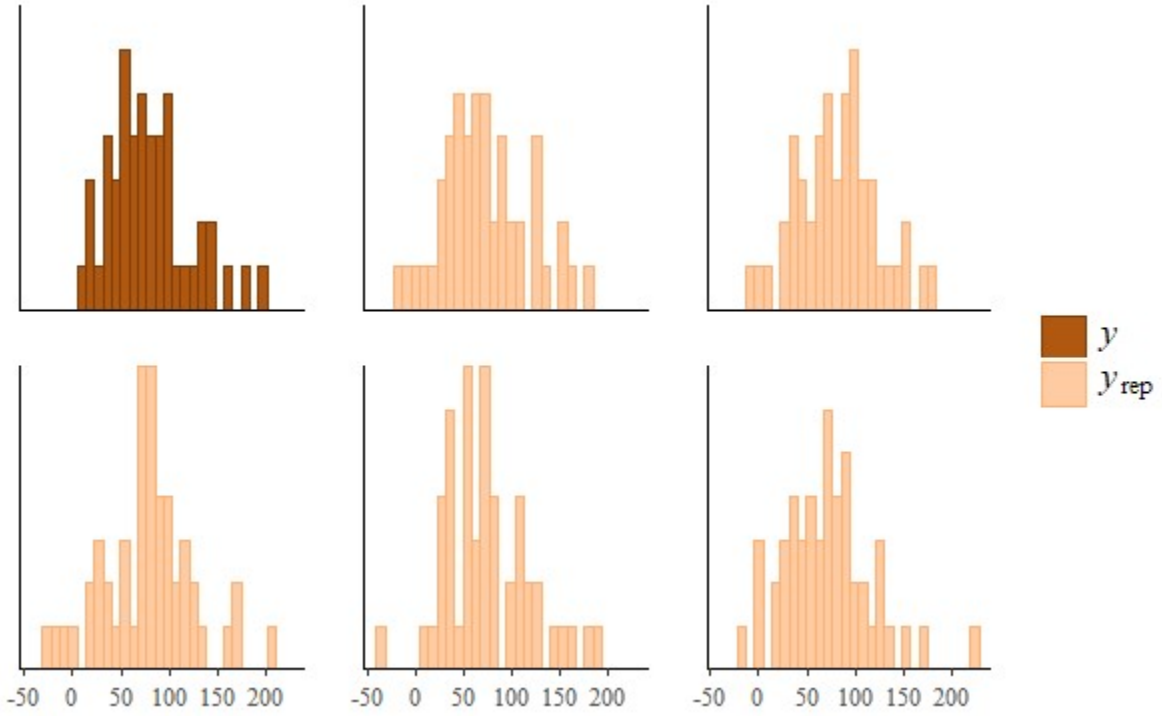


Figure 3 Histogram of Replicated Data

The replicated data has roughly the same mean, but the standard deviation (46) seems to be a little bit higher than the actual data (42). This might suggest using a different model than the normal regression. But the difference is not so big that we need to sacrifice interpretability to use transformations of the response.

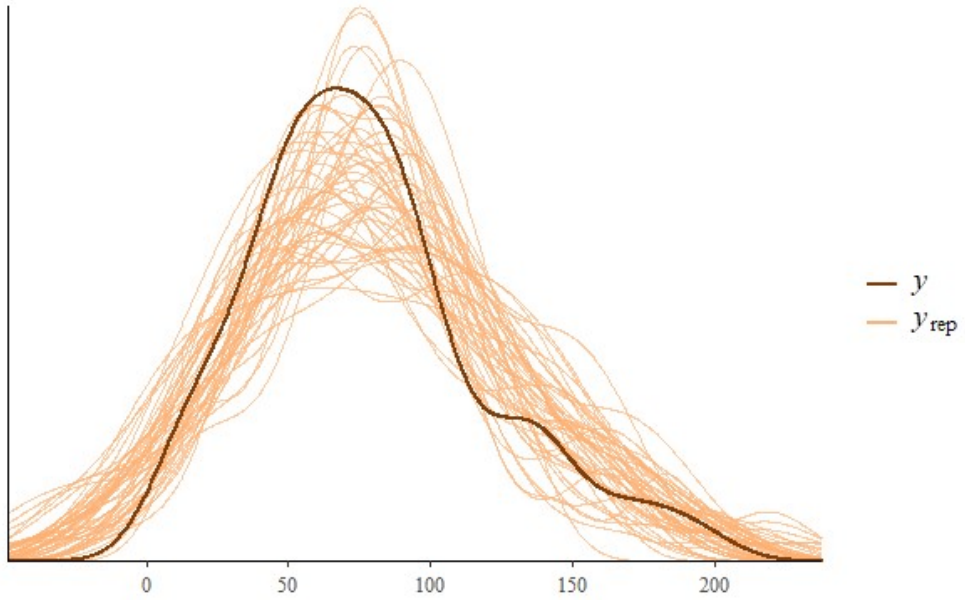


Figure 4 Density Plot of Replicated Data

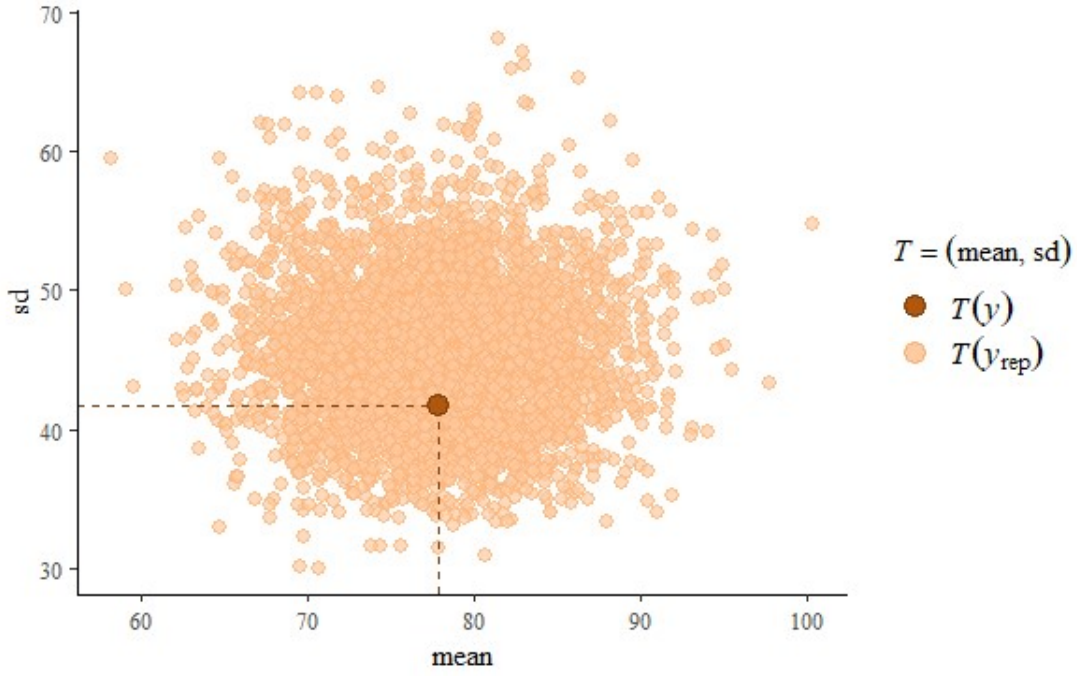


Figure 5 Mean and SD for Actual and Replicated Data

The LOO-PIT quantiles which compares standardized PIT values to the standard normal distribution look normal.

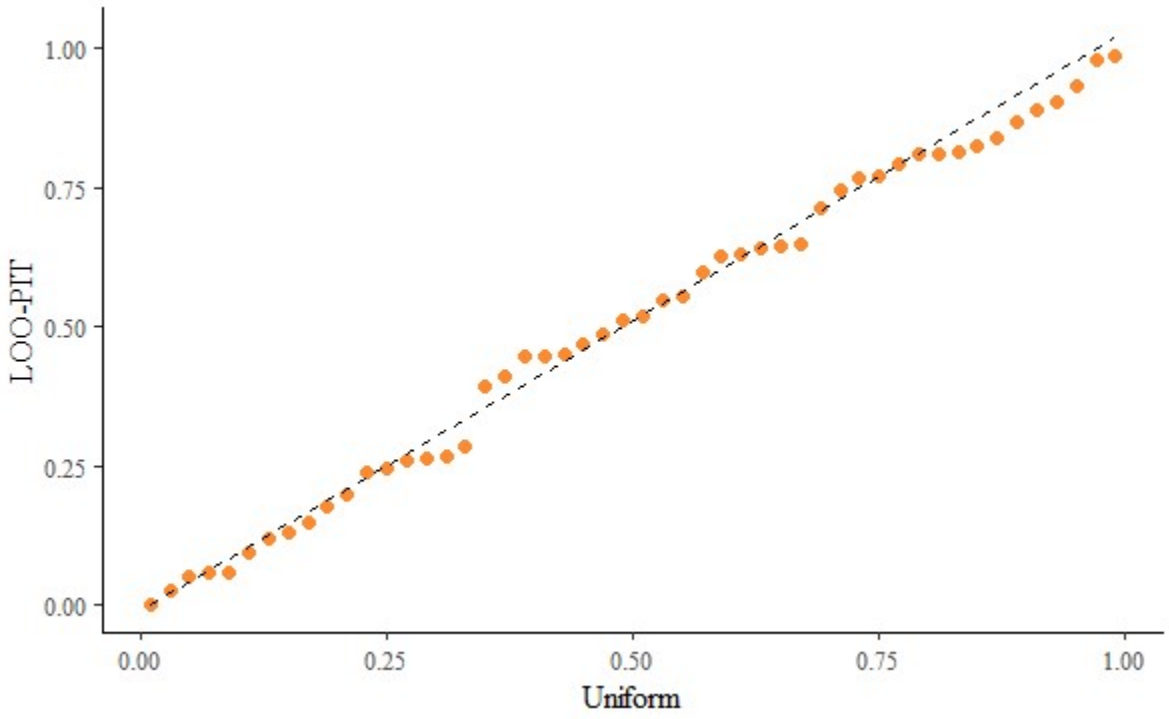


Figure 6 LOO-PIT Quantiles Plot

4.6 Posterior Results

According to the following table, our finalized model is:

$$mortality = 387.2 + 0.1 * population_{density} - 0.1 * \text{Biden}_{\text{rate}} - 4.7 * \text{latitude} - 131.5\,\text{percent\_of\_white} - 4.7 * \text{latitude} - 189.3 * \text{percent\_of\_asian} - 6.3 * \text{temperature} + 241.3 * \text{poverty\_rate} + 16.3 * \text{red} * \text{poverty\_rate}$$

Table 7 Posteriors

	MEAN	**SD**	**10%**	**50%**	**90%**
$\boldsymbol{\beta_0}$: intercept	387.2	121.2	232.4	387.4	541.2
$\boldsymbol{\beta_1}$: population density	0.1	0.0	0.1	0.1	0.1
$\boldsymbol{\beta_2}$: Biden support rate	-0.1	0.9	-1.3	-0.1	1.0
$\boldsymbol{\beta_3}$: Latitude	-4.7	1.6	-6.7	-4.7	-2.7
$\boldsymbol{\beta_4}$: Percent of White	-131.5	63.7	-212.8	-131.9	-50.7
$\boldsymbol{\beta_5}$: Percent of Asian	-189.3	144.1	-371.2	-189.2	-9.4
$\boldsymbol{\beta_6}$: Average Temp	-6.3	2.4	-9.3	-6.4	-3.4
$\boldsymbol{\beta_7}$: Poverty Rates	241.3	284.2	-106.8	239.6	603.1
$\boldsymbol{\beta_8}$: Red*Poverty	16.3	341.0	-410.0	14.0	442.8
$\boldsymbol{\beta_9}$**: SIGMA**	28.7	3.3	24.8	28.3	33.1

4.7 Interpretations

1. For a state with 6 million population, 50 residents per square mile, 50% Biden support rate, 40-degree latitude, 80% percent white population and 5% percent Asian population, 7.3 degree average temperature, 10 percent people living under poverty, and do not support Trump, we expect there are roughly 63 people per 100, 000 residents has died from COVID-19 by now. Notice this imaginary state's demographic information is made up intentionally to mimic Colorado. And the actual mortality number for Colorado is 60 which is very close to our imaginary's state's number.
2. Holding other demographics variable constant：
    a. if the above imaginary state turns to support Trump, we would expect additional 1.6 deaths per 100, 000 residents, but this expectation has a higher amount of uncertainty since the density plot of corresponding coefficients straddles on zero.
    b. If the above imaginary state reduces its poverty rate to 5%, we would expect 50 (instead of 62) people died from COVID, which is 12 lives less. This prediction has a high amount of certainty since the major area of the distribution of corresponding coefficient are in the positive area (density-plot of the corresponding coefficient is positively skewed).

5 Conclusions

COVID mortality rate is positively influenced by population density, GDP, violent crime rates. COVID mortality is negatively influenced with Biden support rate, latitude, asian/white population, temperature

Before doing this research, I thought the COVID-19 is a respiratory disease is very similar with influenzas in terms of its transmission pathways and the fact that it is disproportionally affecting the seniors and patients with pre-conditions. So, the mortality rate of COVID-19 and influenzas should have a very higher correlation and the influenzas' mortality rate would be a good predictor for COVID-19. However, after doing this research, I found the correlation between them is very weak (0.07) and influenzas' mortality rate is not a good predictor for the COVID-19 mortality. Instead, there is strong evidence that the temperature and poverty matters. This might due to the fact COVID-19 transmit easier indoors and warmer state's resident tend to spend more time outside. As for the poverty, it goes without saying that how expensive the medical service in the US is and people living under poverty are least likely to receive adequate health care at the early stage of the infection which eventually adds to their odds of mortality.

Directions for Future Research

1. Use individual demographic information rather then the collective information for the whole state. This would give us a much higher precision in drawing the conclusions about each patient's COVID-19 mortality odds.
2. Implement the generalized linear model and consider the log transformation of the response. As I noted earlier in the model section, the variance for the replicated data

is a little bit higher than the actual observations which may indicate the deficiency of linear regression model.

3. Remove outliers: As we can see from the histogram, several states have particularly higher mortality rate. We can remove 5 out 50 states as our outliers so that the normal regression may work better.

Final Project – Code Part

```
# load packages
library(ggplot2)
library(bayesplot)
theme_set(bayesplot::theme_default())

library(rstanarm)

#load data
library(readxl)
data <- read_excel("C:/Users/Don/Desktop/Bayes Project/data.xlsx")

# created the "blue" indicator
data$blue = abs(1-data$red)

# define models
post1 <- stan_glm(covid_mortality ~ red, data = data,
                  family = gaussian(link = "identity"),
                  seed = 1)
post2 <- update(post1, formula = . ~ poverty_rates)
post3 <- update(post1, formula = . ~ red + poverty_rates)
post4 <- update(post1, formula = . ~ red * poverty_rates)

# output coefficients
reg0 <- function(x, ests) cbind(1, 0, x) %*% ests
reg1 <- function(x, ests) cbind(1, 1, x) %*% ests
```

```
args <- list(ests = coef(post3))

# define categorical variables
data$clr <- factor(data$blue, labels = c("Red States","Blue States"))
lgnd <- guide_legend(title = NULL)

# create plot about the mortality over poverty rates
base2 <- ggplot(data, aes(x = poverty_rates, color = clr), scale_color_manual(values = c("blue","red"))) +
  geom_point(aes(y = covid_mortality), shape = 21, stroke = .2, size = 1) +
  guides(color = lgnd, fill = lgnd) +
  theme(legend.position = "right")

# add parallel lines
base2 +
  stat_function(fun = reg0, args = args, aes(color = "Blue States"), size = 1.5) +
  stat_function(fun = reg1, args = args, aes(color = "Red States"), size = 1.5)

#define coefficients
reg0 <- function(x, ests) cbind(1, 0, x, 0 * x) %*% ests
reg1 <- function(x, ests) cbind(1, 1, x, 1 * x) %*% ests
args <- list(ests = coef(post4))

# add non-parallel lines
base2 +
  stat_function(fun = reg0, args = args, aes(color = "Blue States"), size = 1.5) +
  stat_function(fun = reg1, args = args, aes(color = "Red States"), size = 1.5)

# Color by red/blue states
ggplot(data, aes(x=covid_mortality, color=clr, fill=clr)) +
```

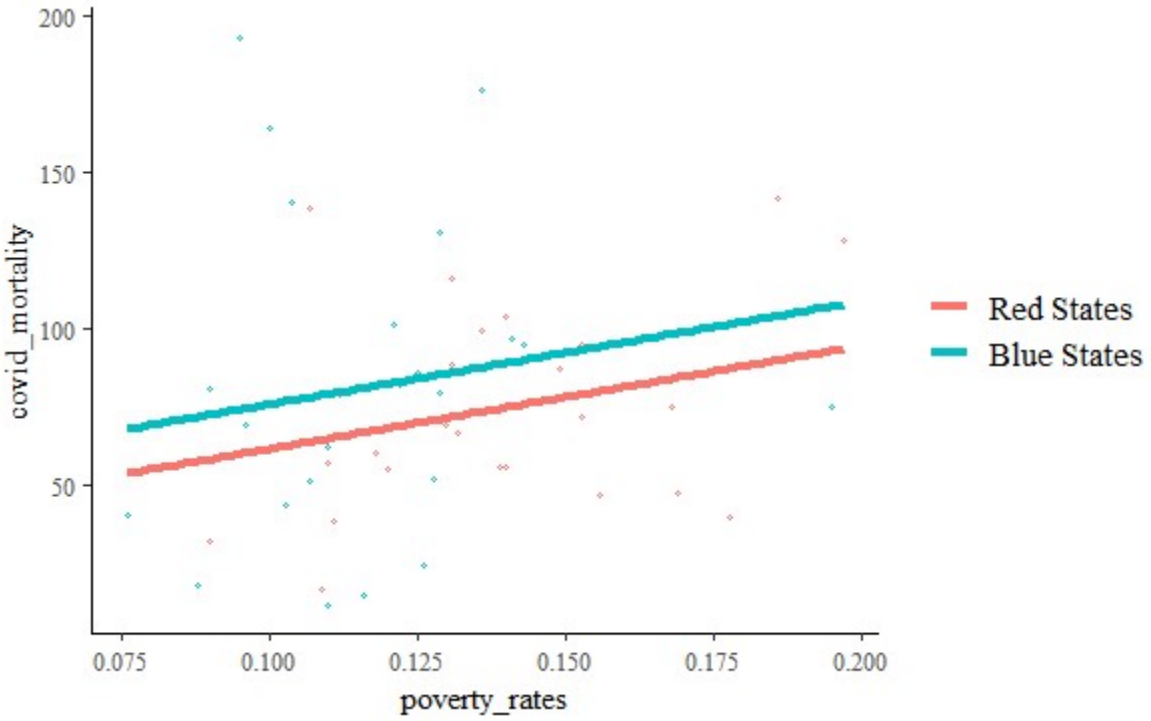
```
 geom_histogram(aes(y=..density..), alpha=0.5,
                position="identity")+
 geom_density(alpha=.2)


library(ggplot2)
library(bayesplot)
theme_set(bayesplot::theme_default())


library(rstanarm)


library(readxl)
data <- read_excel("C:/Users/Don/Desktop/Bayes Project/data.xlsx")



data$blue = abs(1-data$red)


# fit a new batch of models using more predictors
post5 <- update(post1, formula = . ~ popu_density + gdp + biden_rates + latitude + longitude + white + asian + violent_crates + flu_mortality + elevation + temp +red * poverty_rates)


#  delete longitude and violent crime rates
post6 <- update(post1, formula = . ~ popu_density + gdp + biden_rates + latitude + white + asian + flu_mortality + temp +red * poverty_rates)


# delete flu rates
post7 <- update(post1, formula = . ~ popu_density  + biden_rates + latitude + white + asian + temp +red * poverty_rates)


# generate summaries
summary(post5)
```

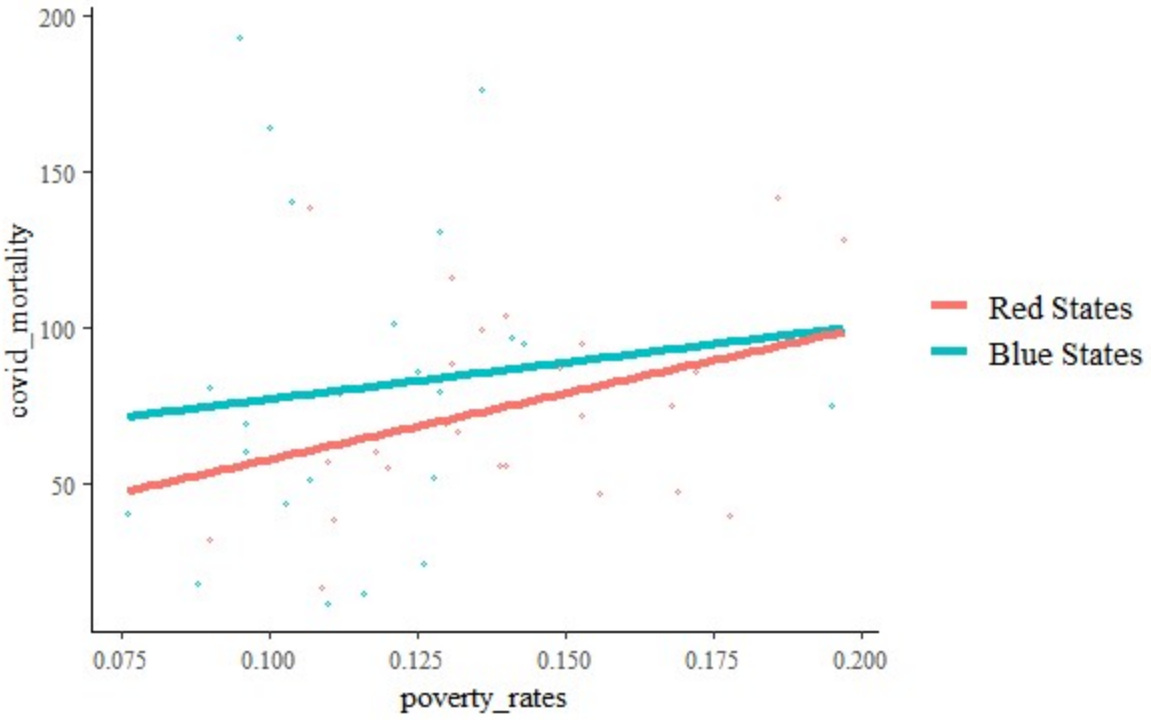
```
summary(post6)
summary(post7)

# get prior info
prior_summary(post7)$prior

# leave-one-out comparions
library(loo)

loo1 <- loo(post1, cores = 2)
loo2 <- loo(post2, cores = 2)
loo3 <- loo(post3, cores = 2)
loo4 <- loo(post4, cores = 2)
(comp <- loo_compare(loo1, loo2, loo3, loo4))



loo5 <- loo(post5, cores = 2)
loo6 <- loo(post6, cores = 2)
loo7 <- loo(post7, cores = 2)

(comp <- loo_compare(loo1, loo2, loo3, loo4,loo5,loo6, loo7))

# Posterior predictive errors
pp_check(post1, plotfun = "hist", nreps = 5)
pp_check(post2, plotfun = "hist", nreps = 5)
pp_check(post3, plotfun = "hist", nreps = 5)
pp_check(post4, plotfun = "hist", nreps = 5)
pp_check(post5, plotfun = "hist", nreps = 5)
pp_check(post6, plotfun = "hist", nreps = 5)
```

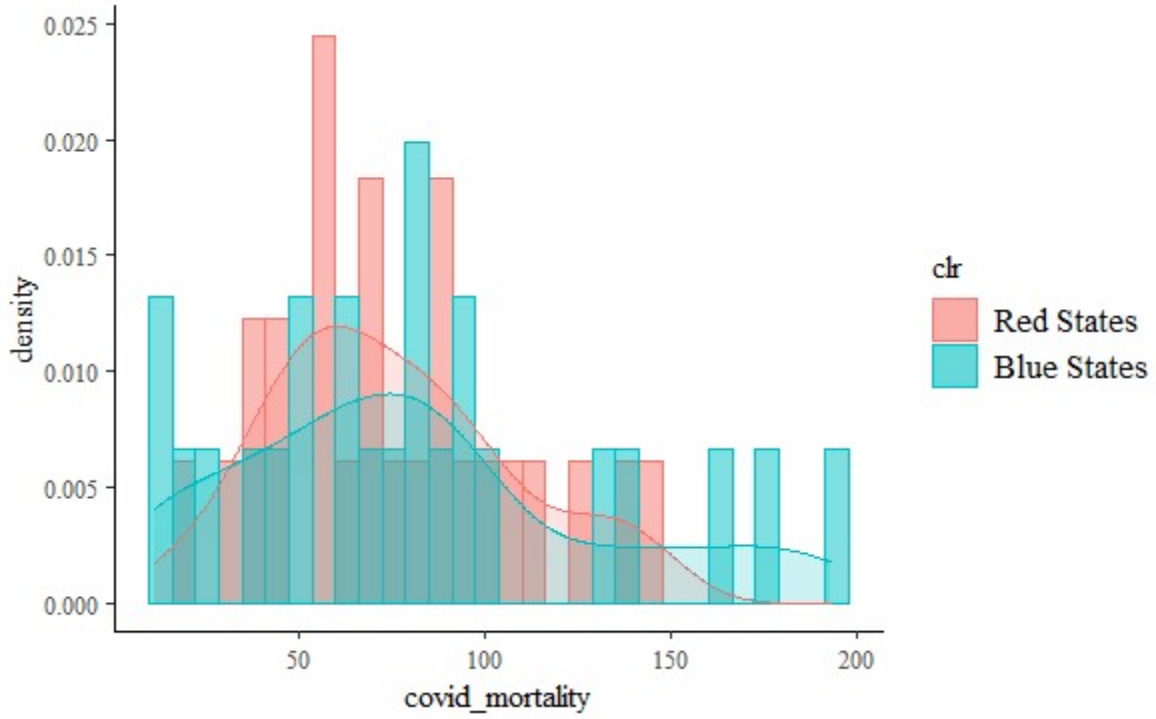
```
pp_check(post7, plotfun = "hist", nreps = 5)

# do differen kinds of posterior checks
pp_check(post5, plotfun = "stat", stat = "mean")
pp_check(post7, plotfun = "stat", stat = "mean")
pp_check(post7, check = "distributions")
pp_check(post7, check = "residuals")
pp_check(post7, check = "scatter")
pp_check(post7, plotfun = "dens_overlay")

y= data$covid_mortality
yrep = as.matrix(posterior_predict(post7,draws = 4000))

loo7 <- loo(post7, save_psis = TRUE, cores = 2)
psis7 <- loo7$psis_object
lw <- weights(psis7)

color_scheme_set("orange")
ppc_loo_pit_overlay(y, yrep, lw = lw)

# leave-one-out quantiles
ppc_loo_pit_qq(y, yrep, lw = lw)
keep_obs <- 1:50
ppc_loo_intervals(y, yrep, psis_object = psis7, subset = keep_obs)

# check posterior predictive mean and variances
pp_check(post5, plotfun = "stat_2d", stat = c("mean", "sd"))
pp_check(post7, plotfun = "stat_2d", stat = c("mean", "sd"))
```

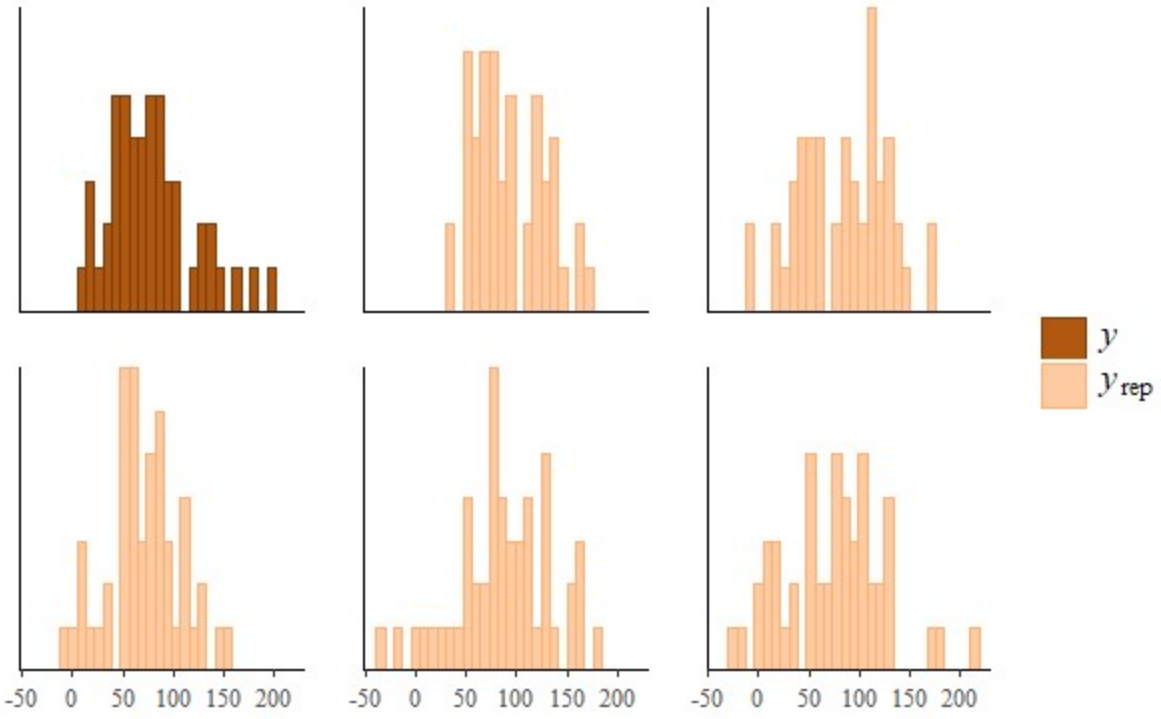
```
# Compare predictions
Poverty_rate <- seq(from = 0.075, to = 0.20, by = 0.00255)
y_blue <- posterior_predict(post5, newdata = data.frame(data[-22],red = rep(0,50)))
y_red <- posterior_predict(post5, newdata = data.frame(data[-22],red = rep(1,50)))
dim(y_red)

par(mfrow = c(1:2), mar = c(5,4,2,1))
boxplot(y_red, axes = FALSE, outline = FALSE, col = "red", ylim = c(0,200),
      xlab = "Poverty Rates", ylab = "COVID Mortality", main = "Red States")
axis(1, at = 1:ncol(y_red), labels = Poverty_rate, las = 3)
axis(2, las = 1)
boxplot(y_blue, outline = FALSE, col = "blue", axes = FALSE, ylim = c(0,200),
      xlab = "Poverty Rates", ylab = NULL, main = "Blue States")
axis(1, at = 1:ncol(y_red), labels = Poverty_rate, las = 3)
```

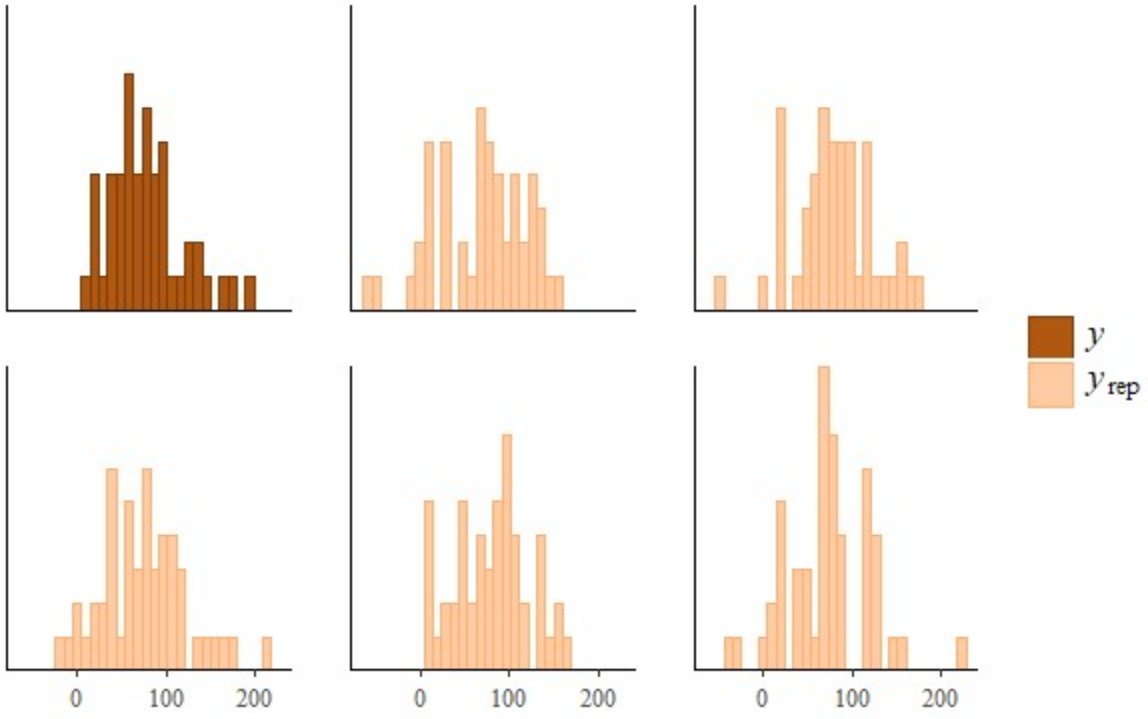
Final Project – Output

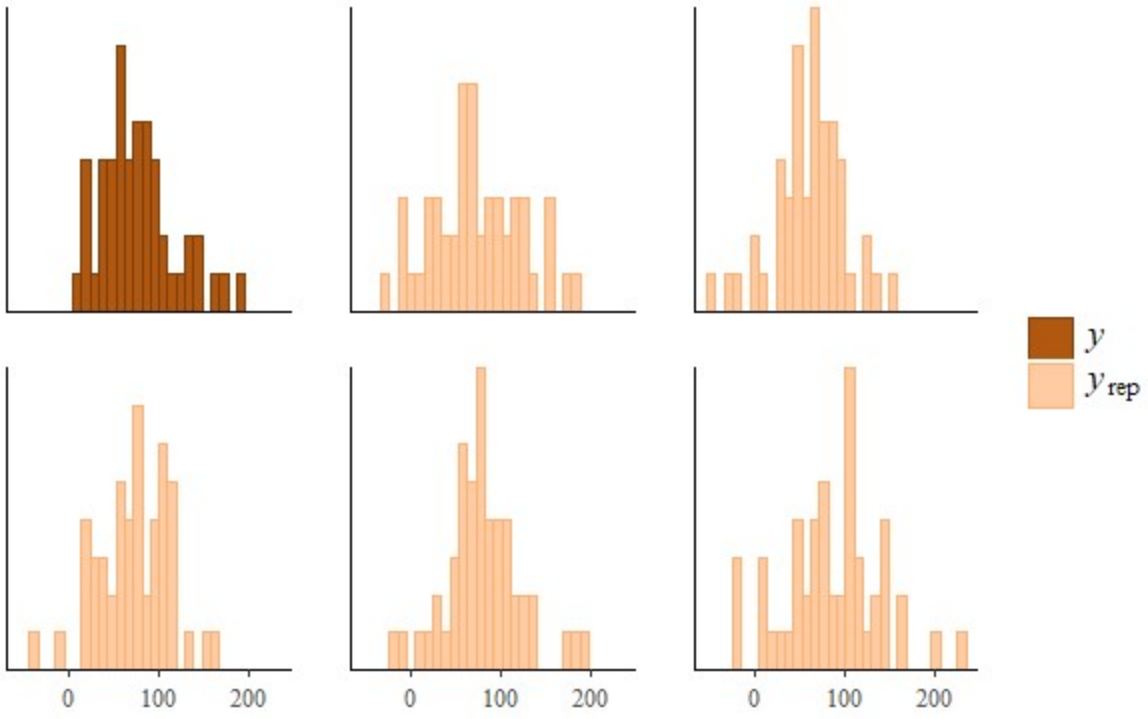


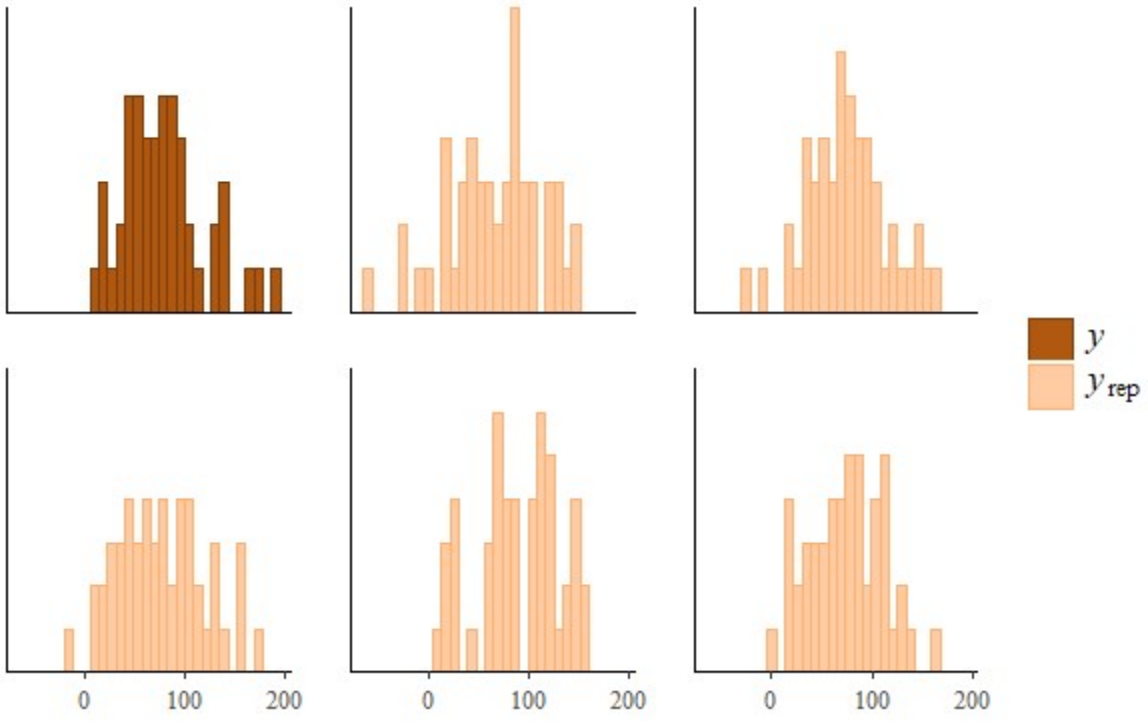


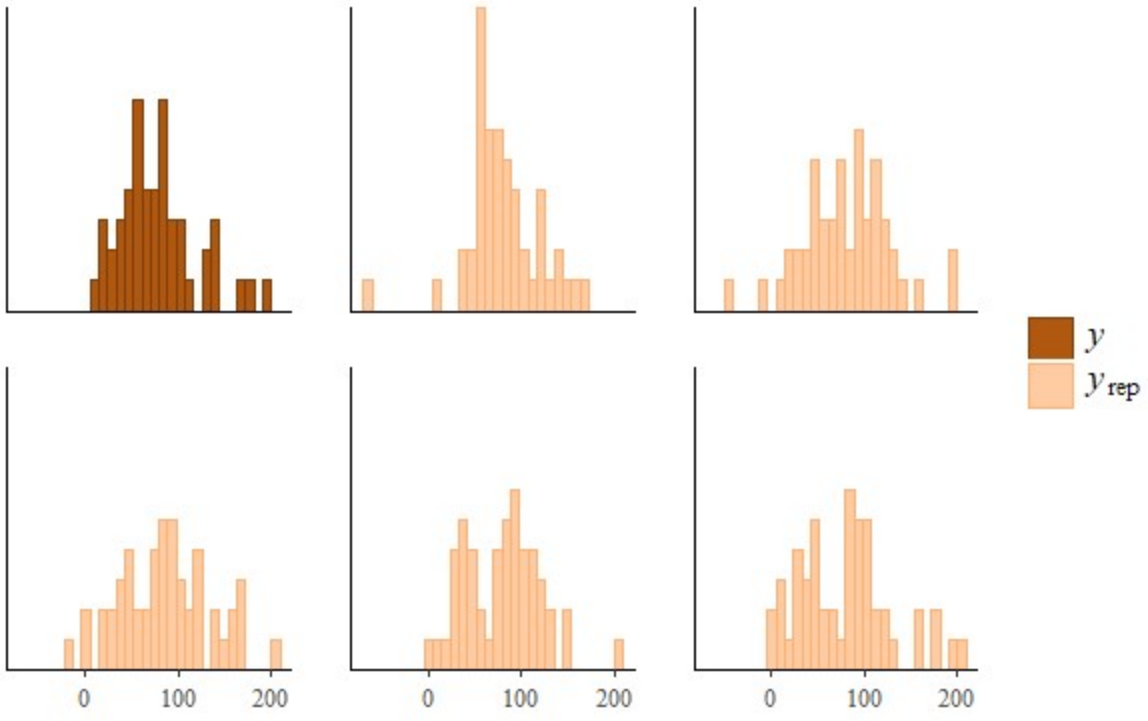

0.025
0.020
0.015
0.010
0.005
0.000
density
50
100
150
200
covid_mortality
clr
Red States
Blue States

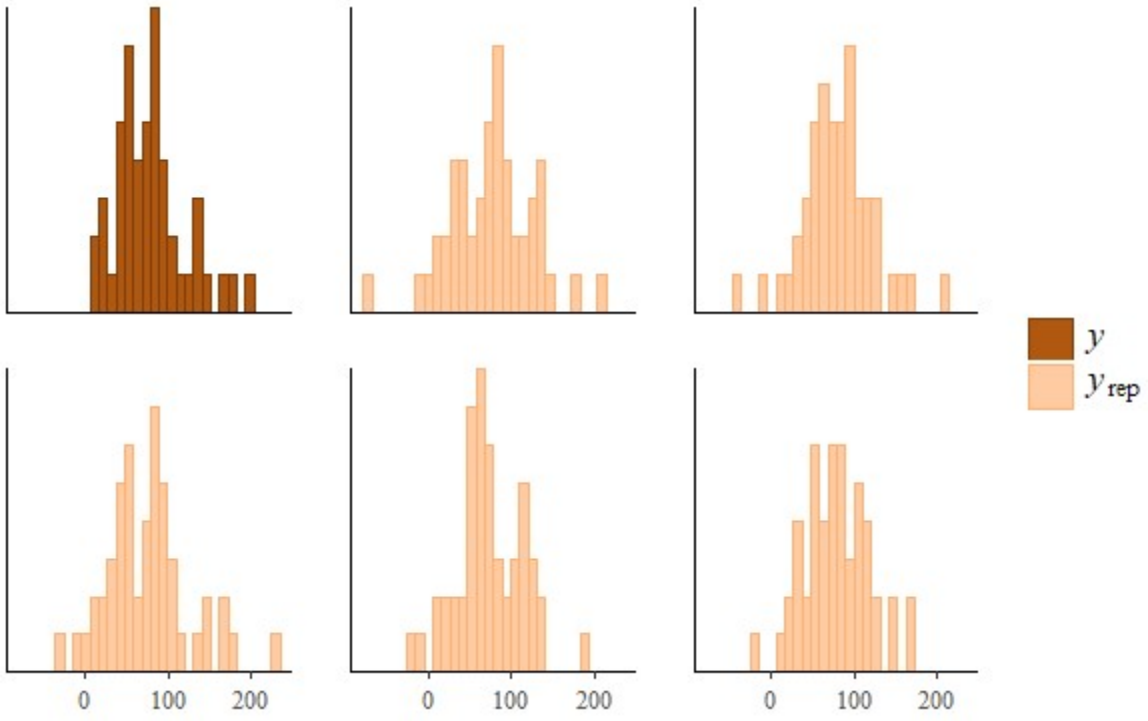

-50
0
50
100
150
200
y
y rep

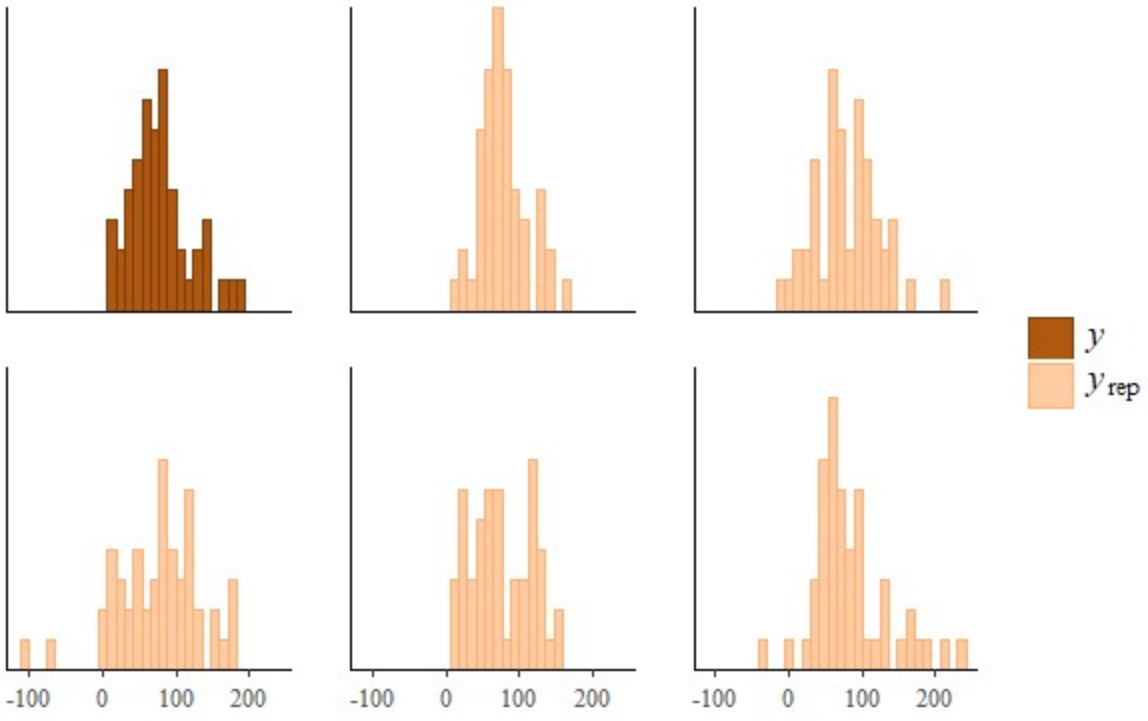

0
100
200
y
y rep

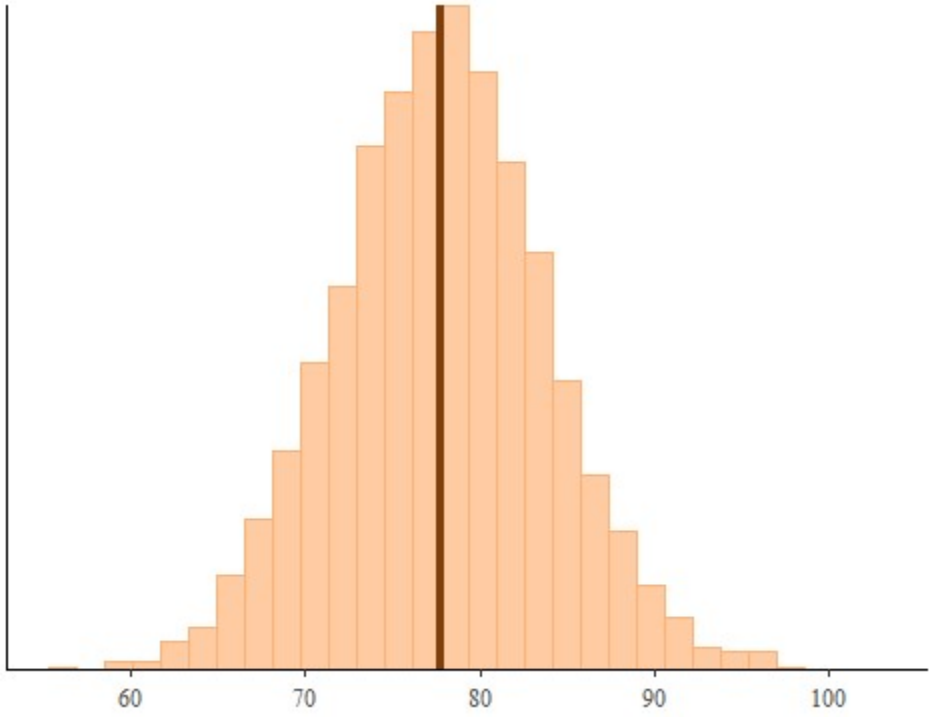

0
100
200
y
y rep

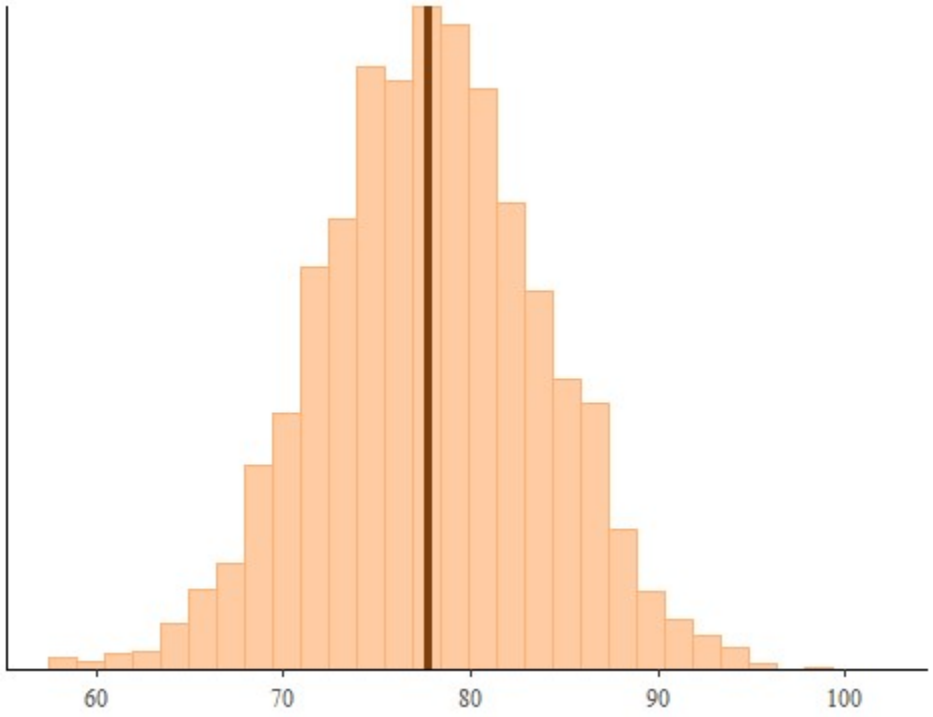

0
100
200
y
y rep

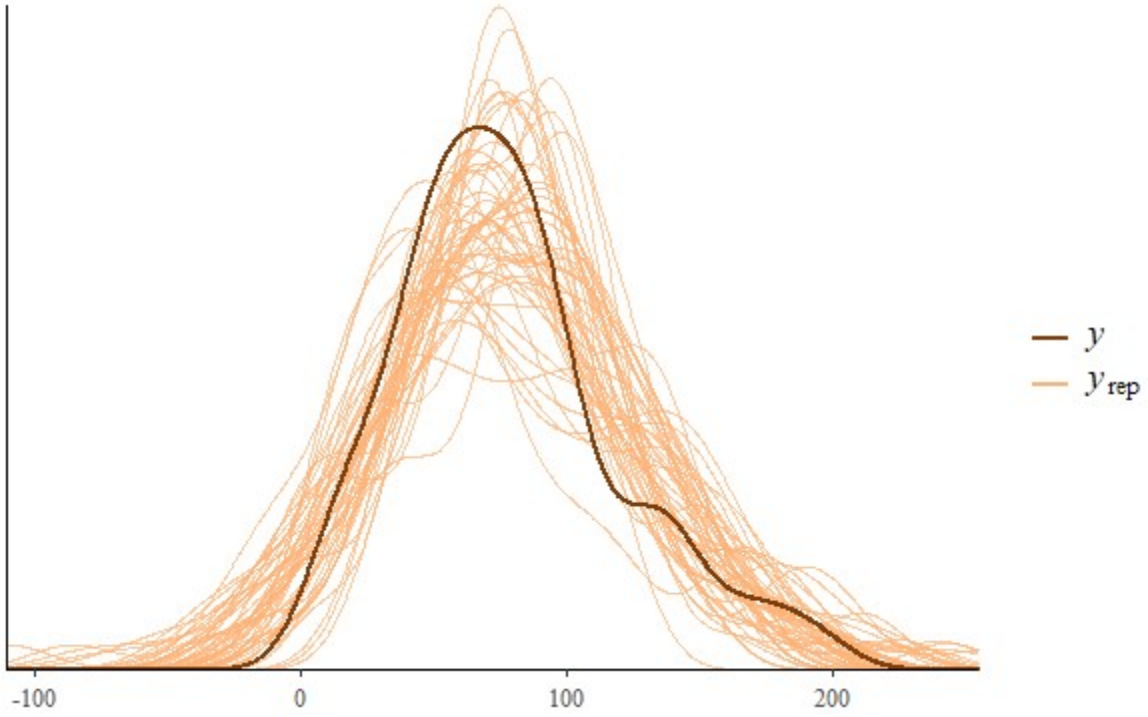

0
100
200
y
y rep

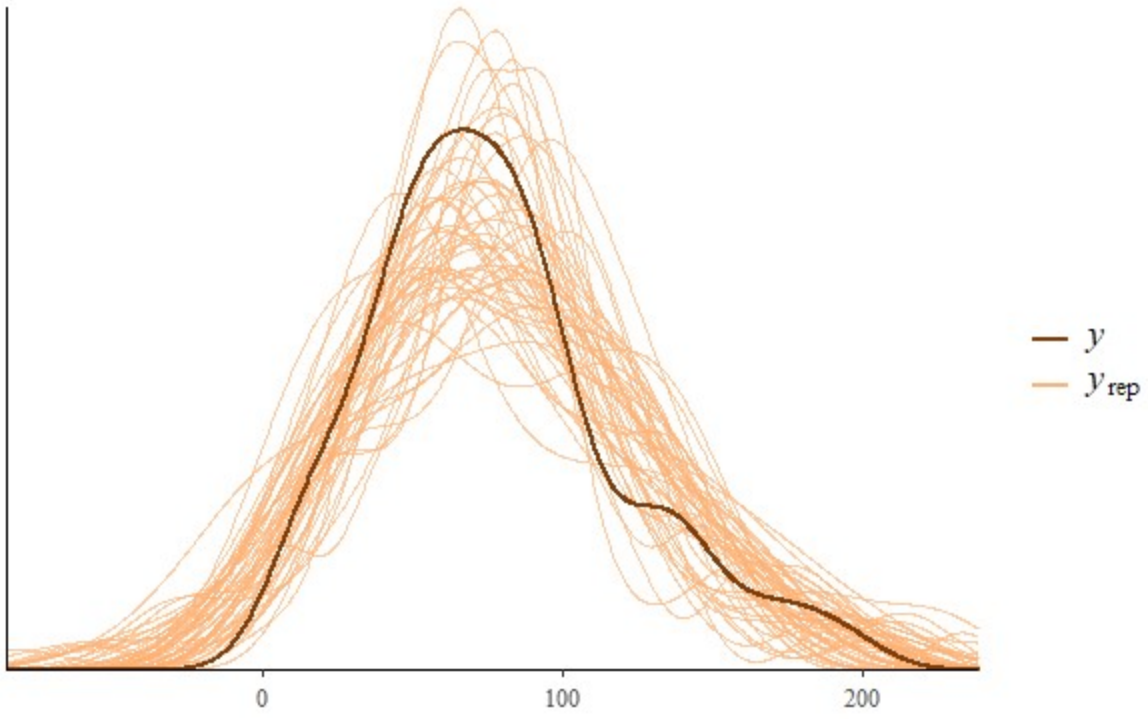

y
y rep
0
100
200

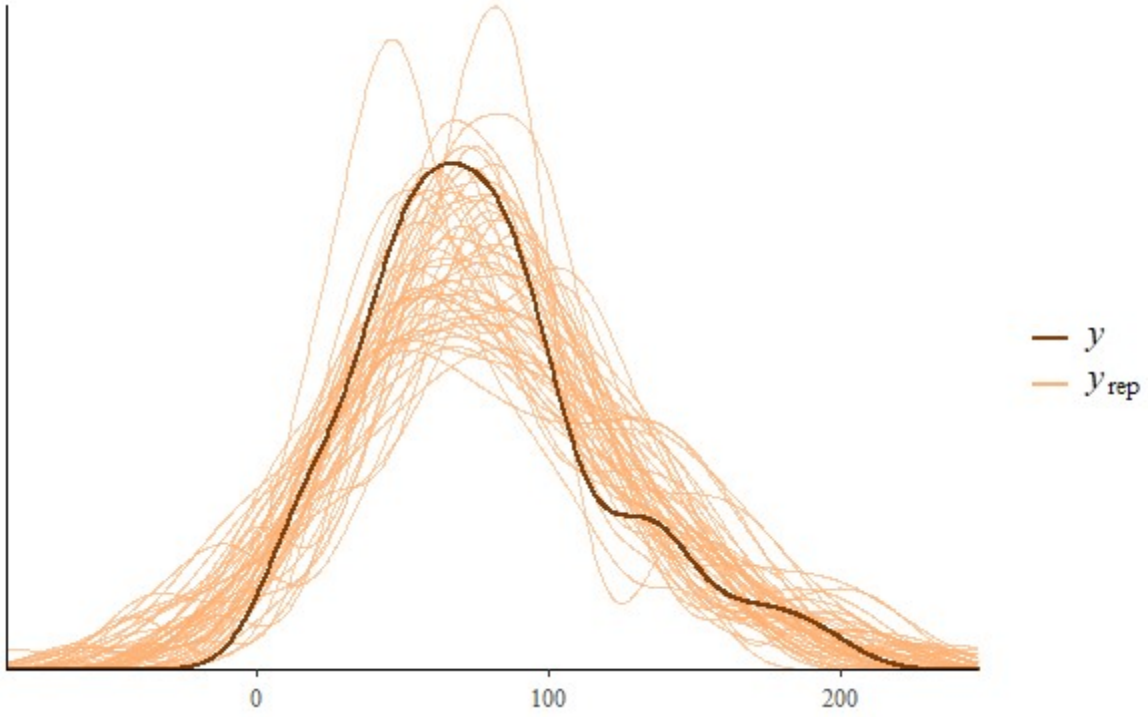

y
y rep
-100
0
100
200

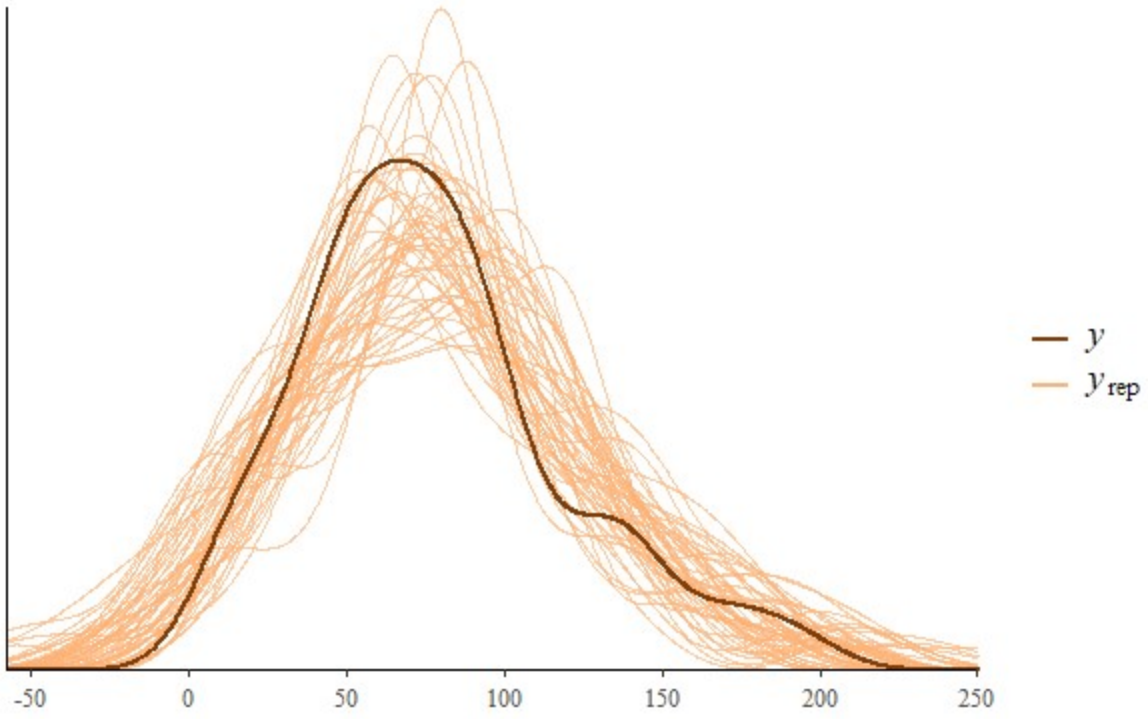

60
70
80
90
100
T = mean
$T(y_{rep})$
$T(y)$

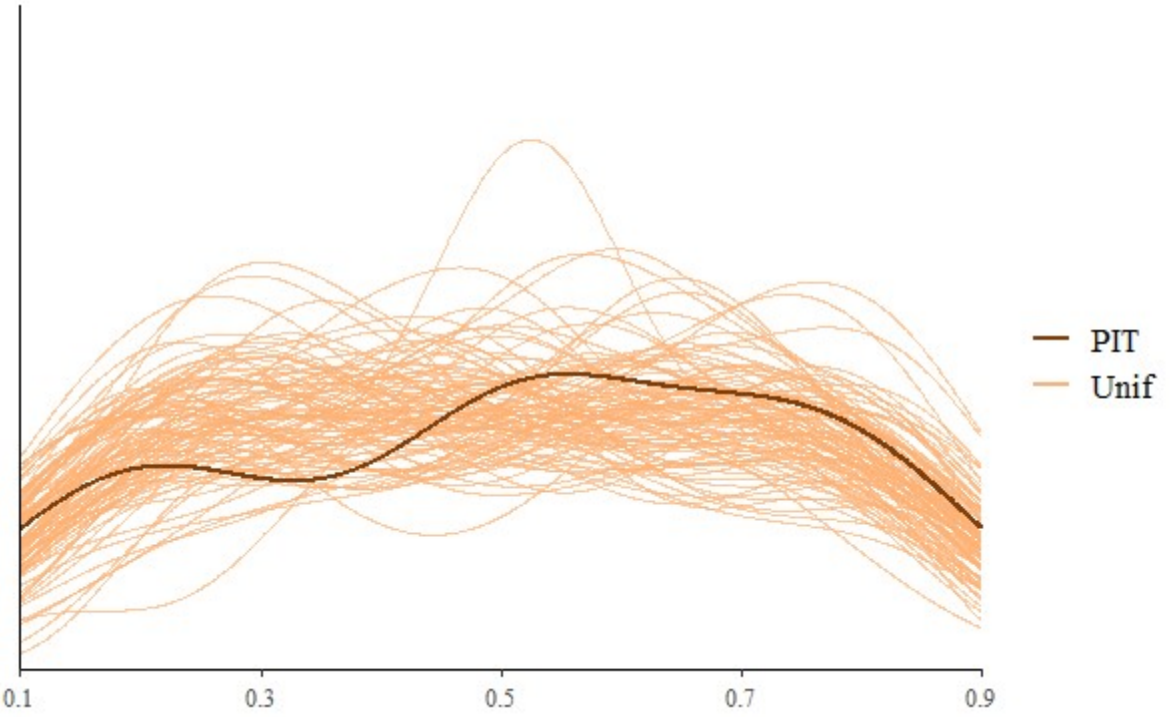

60
70
80
90
100
T = mean
$T(y_{rep})$
$T(y)$

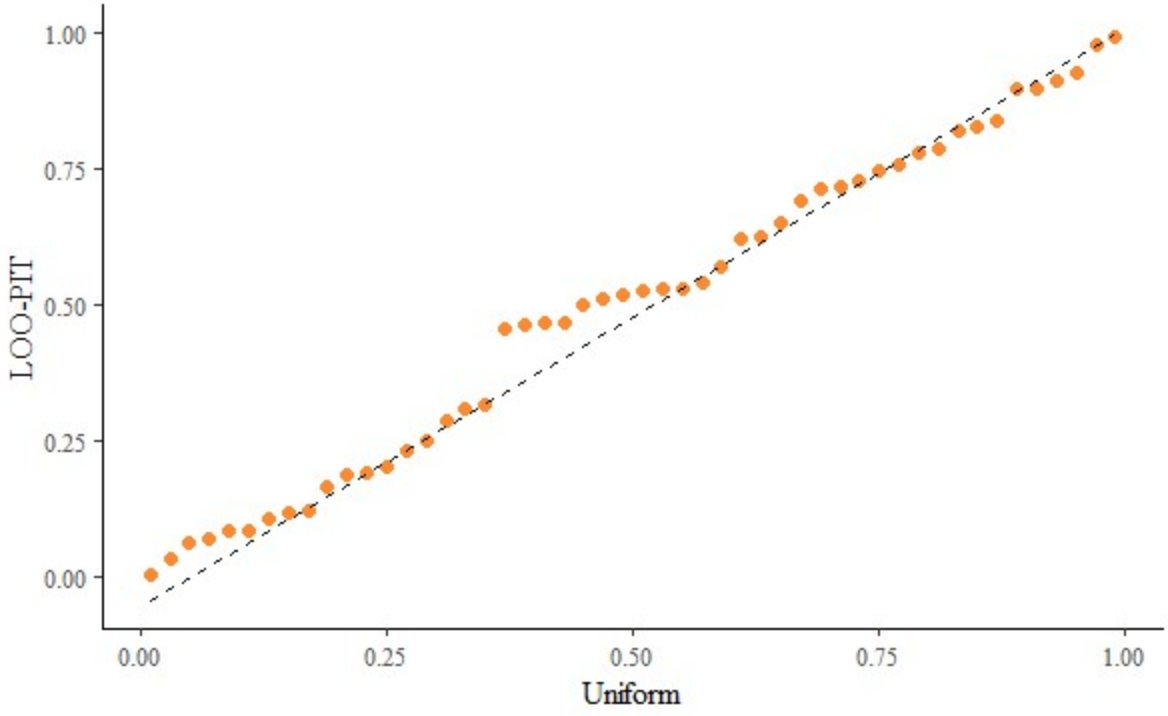

-100
0
100
200
y
y rep

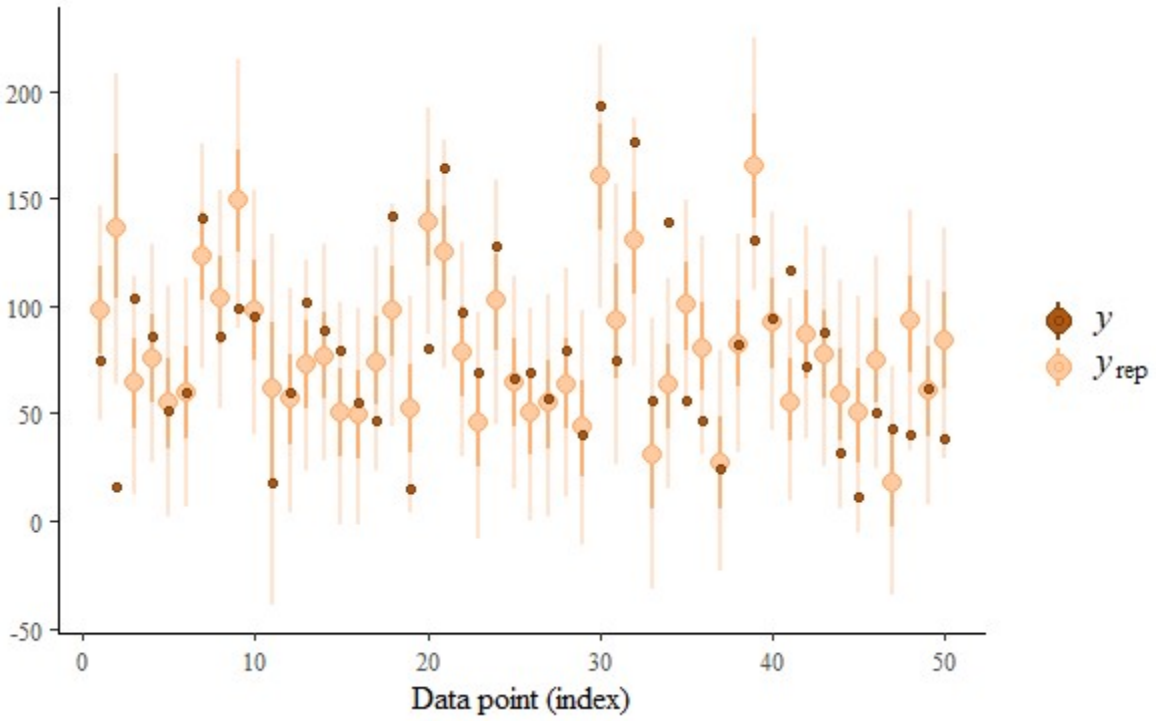

0
100
200
y
y rep

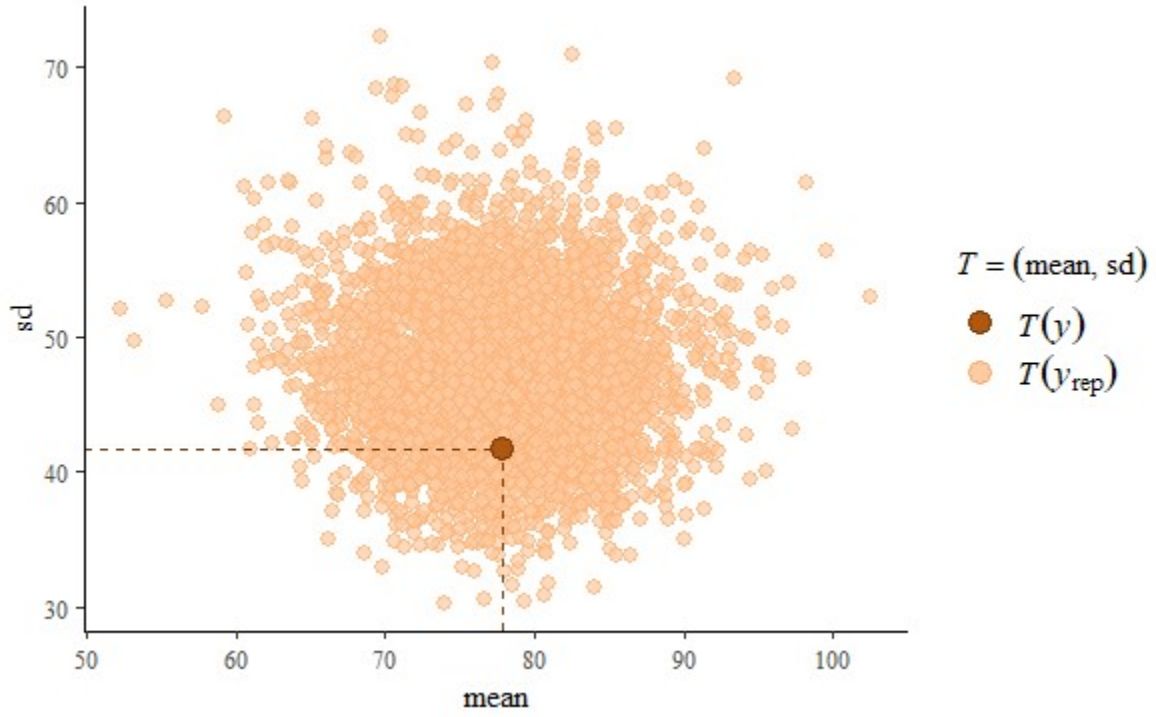

y
y rep
0
100
200


y
y rep
-50
0
50
100
150
200
250

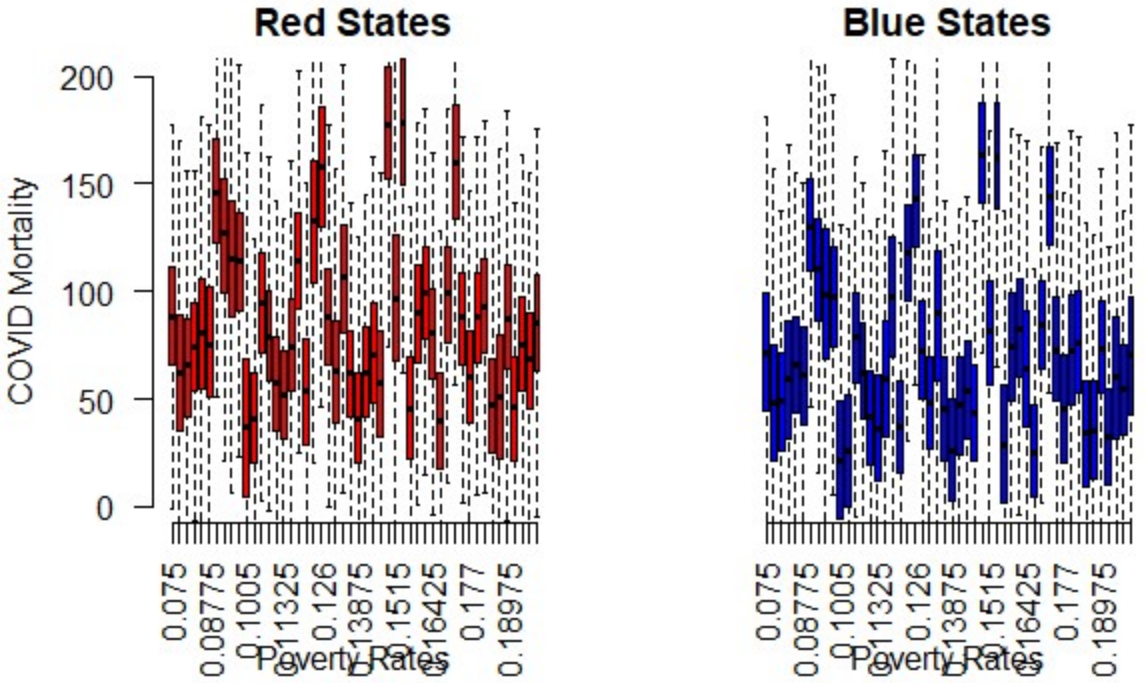

PIT
Unif
0.1
0.3
0.5
0.7
0.9


LOO-PIT
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Uniform


200
150
100
50
0
-50
0
10
20
30
40
50
Data point (index)
y
y_{rep}


70
60
50
40
30
sd
50
60
70
80
90
100
mean
T = (mean, sd)
T(y)
$T(y_{rep})$


sd
mean
70
60
50
40
30
60
70
80
90
100
$T = (\text{mean, sd})$
$T(y)$
$T(y_{\text{rep}})$


Red States
Blue States
COVID Mortality
200
150
100
50
0
0.075
0.08775
0.1005
0.11325
0.126
0.13875
0.1515
0.16425
0.177
0.18975
Poverty Rates