APSTA-GE 2123 Assignment 2

Yuyue Hua

1 The Impact of Medicaid Expansion on Voter Participation

```
library(brms)
```

```
## Loading required package: Rcpp

## This version of Shiny is designed to work with 'htmlwidgets' >= 1.5.
##     Please upgrade via install.packages('htmlwidgets').

## Loading 'brms' package (version 2.12.0). Useful instructions
## can be found by typing help('brms'). A more detailed introduction
## to the package is available through vignette('brms_overview').

##
## Attaching package: 'brms'

## The following object is masked from 'package:stats':
##
##     ar
```

```
library(haven)
#cat('options(contrasts = c(unordered = "contr.treatment", ordered = "contr.treatment"))', file = "~/.R
#unzip("100.00019026_supp.zip")
oregon <- as_factor(read_dta(file.path("19026_supp", "Data", "individual_voting_data.dta")))
#table(oregon$treatment) # this indicates who won the Medicaid lottery
#Sys.setenv(LANG = "en")
```

1.1 Priors and Prior Predictive Distribution with brms

```
#get_prior(registered_1 ~ treatment+numhh_list, data = oregon, family =  bernoulli)

prior1 <- brm(registered_1 ~ treatment + numhh_list, data = oregon, family = bernoulli, seed = 2020,samp
                prior(normal(0, 1.5), class = "b") +
                prior(normal(0, 3), class = "Intercept") )
```

```
## Compiling the C++ model

## Start sampling
```

```
ppe<-pp_expect(prior1,nsamples=1100)
summary(colMeans(ppe))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4920  0.4920  0.4920  0.4937  0.4959  0.5023
```

We can see that the prior predicted probability of being registered to vote is centered at around 0.5 with minimum of 0.496 and maximum of 0.516. Although it might be worth trying to put more uncertainty on the priors, there are no weird values in this prior predictive distribution.

1.2 Posterior Distribution

```
post1 <- brm(registered_1 ~ treatment + numhh_list, data = oregon, family = bernoulli, seed = 2020,prio
               prior(normal(0, 1.5), class = "b") +
               prior(normal(0, 3), class = "Intercept"))
```

```
hypothesis(post1, "treatment > 0")
```

```
## Hypothesis Tests for class b:
##        Hypothesis Estimate Est.Error CI.Lower CI.Upper Evid.Ratio
## 1 (treatment) > 0     0.03      0.02        0     0.05      20.62
##   Post.Prob Star
## 1      0.95    *
## ---
## 'CI': 90%-CI for one-sided and 95%-CI for two-sided hypotheses.
## '*': For one-sided hypotheses, the posterior probability exceeds 95%;
## for two-sided hypotheses, the value tested against lies outside the 95%-CI.
## Posterior probabilities of point hypotheses assume equal prior probabilities.
```

```
#draws <- as.matrix(post1)
#mean(draws[,"b_treatment"]>0)
```

From the output of hypothesis command, we can see that about 95% of treatment coefficients from the posterior distribution are greater than zero.

1.3 Alternative Model

```
post2 <- brm(registered_1 ~ treatment + numhh_list+age_list, data = oregon, family = bernoulli,seed = 2
               prior(normal(0, 1.5), class = "b") +
               prior(normal(0, 3), class = "Intercept"))
```

```
loo_subsample(post1,post2,reloo=T)
```

```
## Warning: Different subsamples in 'post2' and 'post1'. Naive diff SE is
## used.
```

```
## Output of model 'post1':
##
## Computed from 4000 by 400 subsampled log-likelihood
## values from 74922 total observations.
##
##          Estimate   SE subsampling SE
## elpd_loo -50956.7 43.9            0.7
## p_loo         3.9  0.1            0.9
## looic    101913.4 87.8            1.4
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
##
## Output of model 'post2':
##
## Computed from 4000 by 400 subsampled log-likelihood
## values from 74922 total observations.
##
```

```
##          Estimate   SE subsampling SE
## elpd_loo -50696.9 49.4            2.0
## p_loo         7.0  0.1            1.7
## looic    101393.7 98.8            3.9
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
##
## Model comparisons:
##       elpd_diff se_diff subsampling_se_diff
## post2   0.0       0.0     0.0
## post1 259.8      66.1     2.1
```

The output of loo shows an increase in ELPD after age is added into the model.

2 Coronavirus in NYC

```
ROOT <- "https://raw.githubusercontent.com/nychealth"
NYC <- readr::read_csv(paste0(ROOT, "/coronavirus-data/master/case-hosp-death.csv"))
```

```
## Parsed with column specification:
## cols(
##   DATE_OF_INTEREST = col_character(),
##   CASE_COUNT = col_double(),
##   HOSPITALIZED_COUNT = col_double(),
##   DEATH_COUNT = col_double()
## )
```

```
NYC$day <- 1:nrow(NYC)
```

2.1 Negative Binomial Model

```
#get_prior(CASE_COUNT ~ poly(day,degree = 2, raw = FALSE), data = NYC, family =  negbinomial)

postnb <- brm(CASE_COUNT ~ poly(day,degree = 2, raw = FALSE), data = NYC, family =  negbinomial,seed=202
          prior(normal(3, 2), class ="b",coef="polydaydegreeEQ2rawEQFALSE1") +
          prior(normal(-2, 2), class ="b",coef="polydaydegreeEQ2rawEQFALSE2") +
          prior(normal(0, 4), class = "Intercept") +
          prior(exponential(1), class = "shape"))
```

2.2 Poisson Model

```
#get_prior(CASE_COUNT ~ poly(day,degree = 2, raw = FALSE), data = NYC, family =  poisson)
postpo <- brm(CASE_COUNT ~ poly(day,degree = 2, raw = FALSE), data = NYC, family =  poisson,seed=2020, p
          prior(normal(0, 2), class ="b",coef="polydaydegreeEQ2rawEQFALSE1") +
          prior(normal(-2, 2), class ="b",coef="polydaydegreeEQ2rawEQFALSE2") +
          prior(normal(0, 4), class = "Intercept") )
```

2.3 Model Comparison

```r
library(bayesplot)
```

```
## This is bayesplot version 1.7.1
```

```
## - Online documentation and vignettes at mc-stan.org/bayesplot
```

```
## - bayesplot theme set to bayesplot::theme_default()
```

```
##    * Does _not_ affect other ggplot2 plots
```

```
##    * See ?bayesplot_theme_set for details on theme setting
```

```r
loo(postnb,postpo,reloo=T)
```

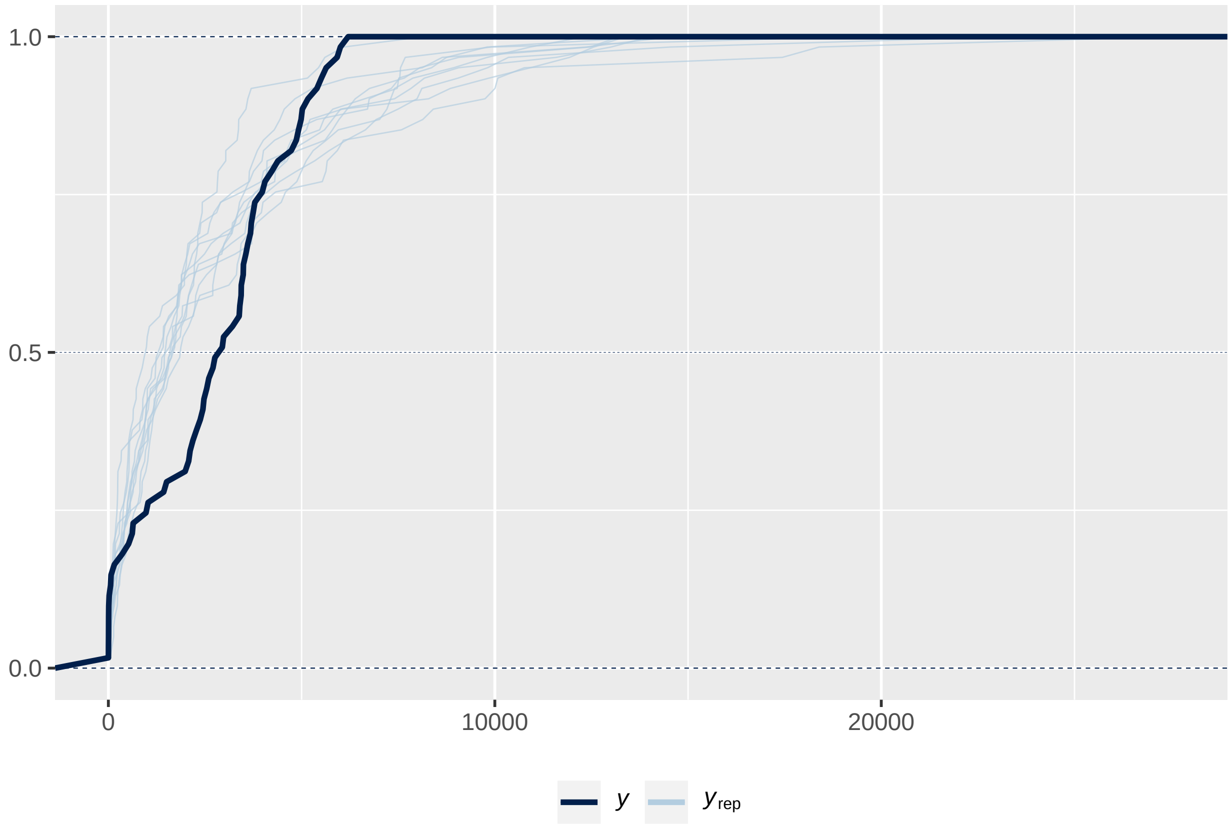
```
## No problematic observations found. Returning the original 'loo' object.
```

```
## 44 problematic observation(s) found.
## The model will be refit 44 times.
```

```
## Output of model 'postnb':
##
## Computed from 4000 by 61 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo   -513.1 10.6
## p_loo         4.4  1.0
## looic      1026.1 21.2
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
##
## Output of model 'postpo':
##
## Computed from 4000 by 61 log-likelihood matrix
##
##          Estimate     SE
## elpd_loo -11206.9 1294.4
## p_loo       982.8  139.6
## looic     22413.8 2588.8
## ------
## Monte Carlo SE of elpd_loo is 3.4.
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     57    93.4%   1
##  (0.5, 0.7]   (ok)        4     6.6%   97
##    (0.7, 1]   (bad)       0     0.0%   <NA>
##    (1, Inf)   (very bad)  0     0.0%   <NA>
##
## All Pareto k estimates are ok (k < 0.7).
## See help('pareto-k-diagnostic') for details.
##
## Model comparisons:
##        elpd_diff se_diff
## postnb      0.0       0.0
## postpo -10693.8    1296.1
```

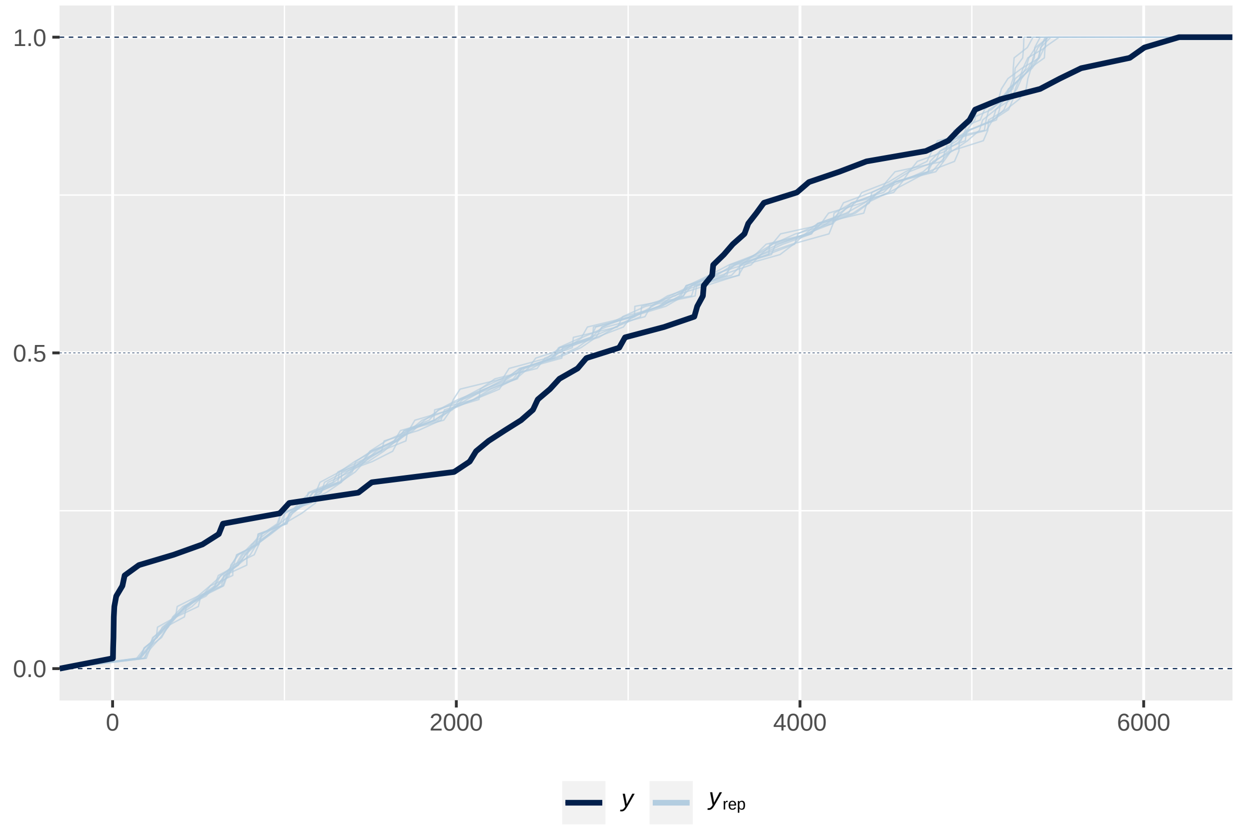
```
pp_check(postnb, type = "ecdf_overlay") + legend_move("bottom")
```

```
## Using 10 posterior samples for ppc type 'ecdf_overlay' by default.
```



```
pp_check(postpo, type = "ecdf_overlay") + legend_move("bottom")
```

```
## Using 10 posterior samples for ppc type 'ecdf_overlay' by default.
```



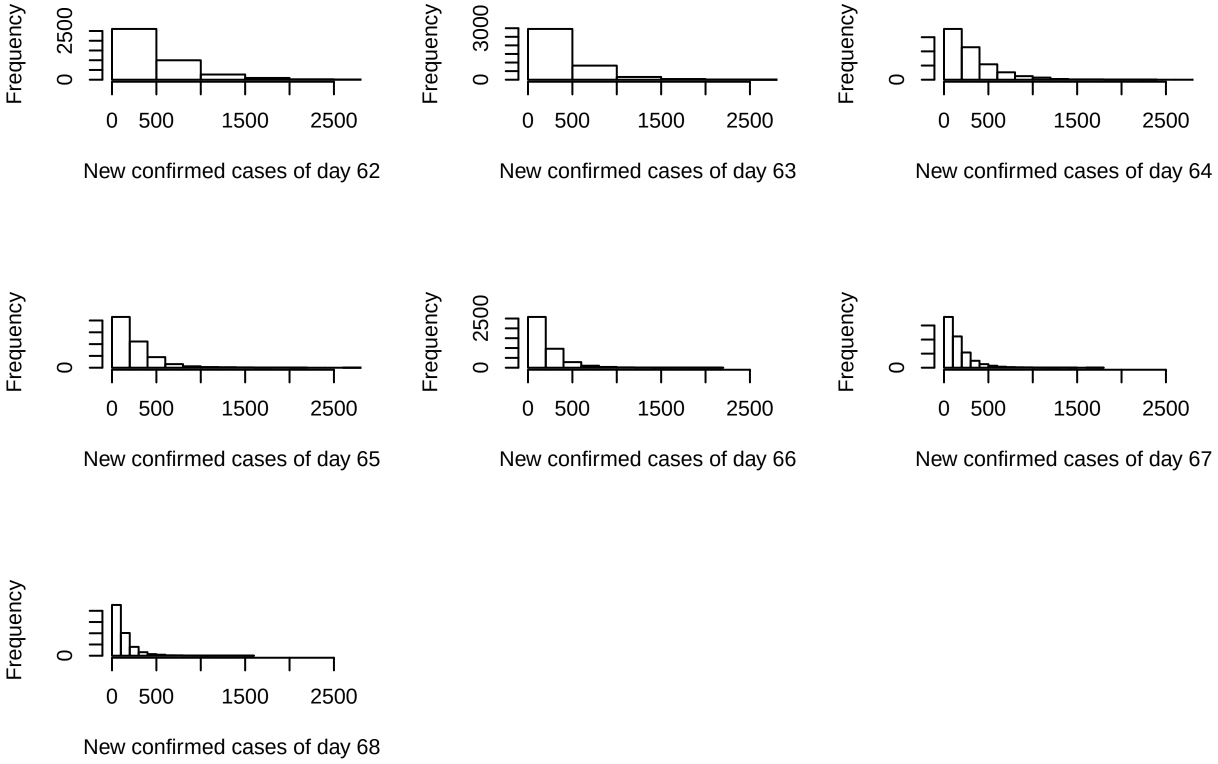
We can plot empirical CDF to see which model fits the data better. We can see that negative binomial model seems to capture the overall trend better. The expected log predictive density is further calculated and suggests that negative binomial model is preferable.

2.4 Posterior Prediction

```
#Create new data
n=dim(NYC)[1]
newday=as.data.frame(NYC$day[n] + 1:7)
names(newday)="day"

#Predict for next week
newcases<-posterior_predict(postnb,newdata=newday)
par(mfrow=c(3,3))
for (i in 1: 7){
  xlabel=paste0("New confirmed cases of day ",n +i)
  hist(newcases[,i],main="",xlab=xlabel,xlim=c(0,2700))
}
round(colMeans(newcases),2)
```

```
## [1] 477.10 384.01 311.29 254.33 199.84 161.92 127.86
```



The histograms for the next coming week shows a decreasing trend of the predicted new confirmed cases everyday with posterior distribution centering more towards left. But we can see that the average predicted values for new confirmed cases of the next 7 days are still quite large. All of them are greater than 100.