

APSTA-GE 2123 Assignment 2 Answer Key

May 4, 2020

1 The Impact of Medicaid Expansion on Voter Participation

```
library(brms)
options(mc.cores = parallel::detectCores())
library(haven)
unzip("100.00019026_supp.zip")
oregon <- as_factor(read_dta(file.path("19026_supp", "Data", "individual_voting_data.dta")))
oregon <- na.omit(oregon[, c("vote_presidential_2008_1", "numhh_list", "treatment", "prevote",
                             "english_list", "female_list", "age_list", "zip_hh_inc_list")])
```

1.1 Priors and Prior Predictive Distribution with brms

First, we can look at the output of `get_prior` to see what all the parameters are referred to as, in order to form our own prior distributions over them.

```
get_prior(vote_presidential_2008_1 ~ numhh_list + treatment, data = oregon, family = bernoulli)
```

```
##           prior      class                coef group
## 1                      b
## 2                      b numhh_listsignedselfupP1additionalperson
## 3                      b numhh_listsignedselfupP2additionalpeople
## 4                      b                      treatment
## 5 student_t(3, 0, 10) Intercept
##   resp dpar nlpar bound
## 1
## 2
## 3
## 4
## 5
```

```
priors <- prior(normal(0, 1), class = "b") +
  prior(normal(0, 1), class = "b", coef = "treatment") +
  prior(normal(0, 1), class = "Intercept")
```

Second, we can draw from those priors using the `brm` function with the `sample_prior = "only"` argument.

```
draws <- brm(vote_presidential_2008_1 ~ numhh_list + treatment, data = oregon,
             family = bernoulli, prior = priors, sample_prior = "only")
mu <- pp_expect(draws, nsamples = 2000)
```

These draws from the prior distribution of the conditional expectation are pretty uniform across people, which is often a reasonable place to start from.

```
summary(apply(mu, MARGIN = 2, FUN = min))
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.003946 0.007942 0.008310 0.007551 0.008310 0.008310
```

```
summary(apply(mu, MARGIN = 2, FUN = max))

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.9976 0.9976 0.9976 0.9977 0.9976 0.9992

summary(apply(mu, MARGIN = 2, FUN = median))

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.4941 0.4945 0.4945 0.4962 0.4947 0.5018

summary(apply(mu, MARGIN = 2, FUN = mean))

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.5001 0.5001 0.5001 0.5008 0.5006 0.5025

summary(apply(mu, MARGIN = 2, FUN = IQR))

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.4826 0.4826 0.4826 0.4876 0.4826 0.5227
```

1.2 Posterior Distribution


Then, we can condition on the data and draw from the posterior distribution:

```
post <- update(draws, sample_prior = "no")

post

## Family: bernoulli
## Links: mu = logit
## Formula: vote_presidential_2008_1 ~ numhh_list + treatment
## Data: oregon (Number of observations: 72452)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##           total post-warmup samples = 4000
##
## Population-Level Effects:
##               Estimate Est.Error 1-95% CI u-95% CI
## Intercept          -0.67      0.01  -0.69  -0.65
## numhh_listsignedselfupP1additionalperson    0.01      0.02  -0.03   0.04
## numhh_listsignedselfupP2additionalpeople    -0.21     0.18  -0.56   0.14
## treatment           0.03      0.02  -0.00   0.06
##
##               Rhat Bulk_ESS Tail_ESS
## Intercept      1.00    4506    2724
## numhh_listsignedselfupP1additionalperson 1.00    4177    2997
## numhh_listsignedselfupP2additionalpeople 1.00    4034    2843
## treatment      1.00    3878    2613
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).

hypothesis(post, hypothesis = "treatment > 0")
```



```
## Hypothesis Tests for class b:
##           Hypothesis Estimate Est.Error CI.Lower CI.Upper Evid.Ratio Post.Prob
## 1 (treatment) > 0      0.03      0.02      0      0.05      24.97      0.96
##      Star
## 1      *
```

```
## ---
## '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.
```

Despite our skeptical priors, given the data, the posterior probability that the treatment increases voter turnout is 0.96. Note that in the original analysis, the authors failed to reject the null hypothesis that the treatment effect is zero, although they used a Gaussian rather than a Bernoulli likelihood and maximized it rather than obtaining a posterior distribution.

For what it is worth, the posterior distribution of the conditional expectation for each person is now not close to uniform

```
mu <- pp_expect(post, nsamples = 2000)
summary(apply(mu, MARGIN = 2, FUN = min))

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.1857 0.3295 0.3295 0.3304 0.3337 0.3337

summary(apply(mu, MARGIN = 2, FUN = max))

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.3453 0.3453 0.3453 0.3499 0.3527 0.4149

summary(apply(mu, MARGIN = 2, FUN = median))

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.2935 0.3378 0.3378 0.3406 0.3443 0.3457

summary(apply(mu, MARGIN = 2, FUN = mean))

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.2940 0.3378 0.3378 0.3406 0.3442 0.3458

summary(apply(mu, MARGIN = 2, FUN = IQR))

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.003190 0.003190 0.003190 0.004160 0.004329 0.050805
```

which is not unusual when conditioning on so many observations in a simple model.

1.3 Alternative Model

Nevertheless, the previous model includes none of the predictors that political scientists usually utilize to predict voter turnout, such as age and income (here of the zip code that the person lives in). I put these in as splines because there is little reason to assume that the log-odds are linear functions of them. In addition, I include dummy variables for whether the person voted in a previous election, speaks English, and is female.

```
oregon$age_list <- oregon$age_list / 10
oregon$zip_hh_inc_list <- oregon$zip_hh_inc_list / 1000
alternate <- update(post, formula. = . ~ . + prevote + english_list + female_list +
  (age_list) + s(zip_hh_inc_list), newdata = oregon)
```

```
## Warning: Rows containing NAs were excluded from the model.
```

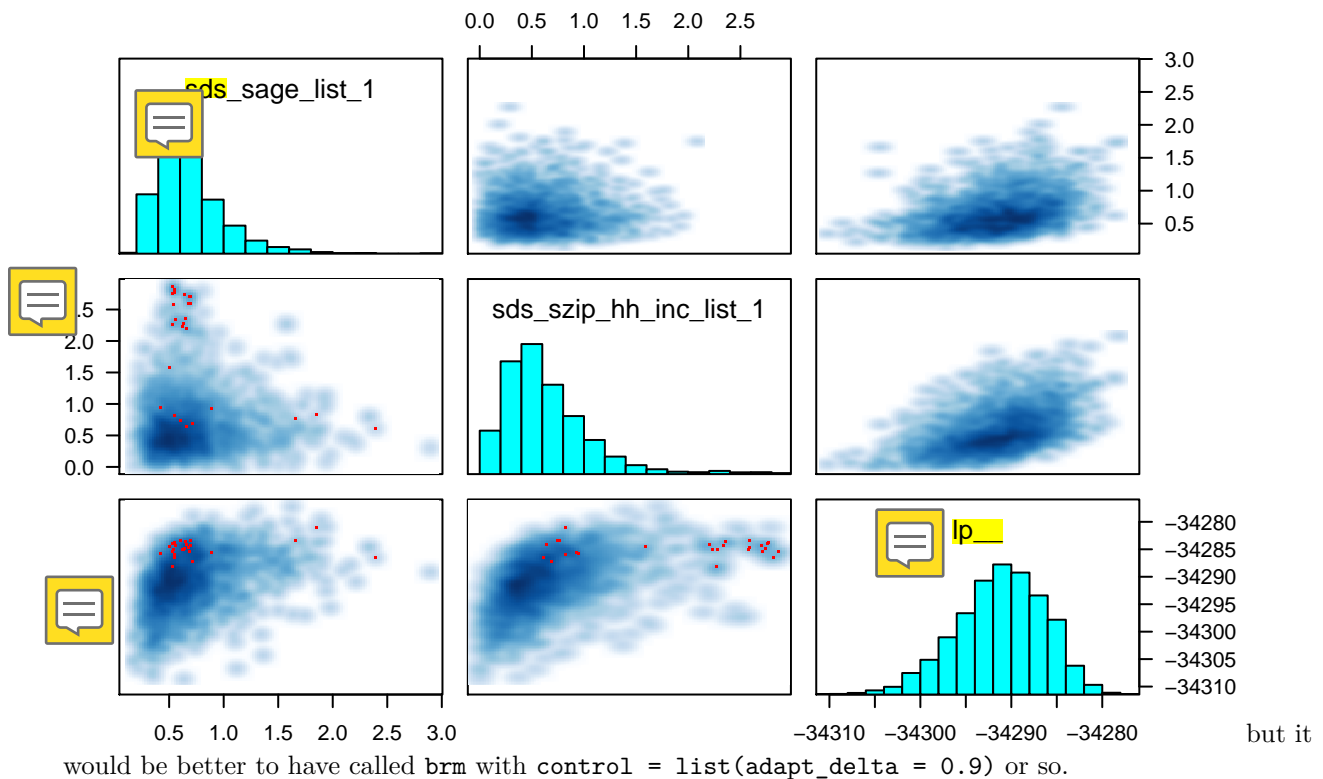
```
## Warning: Rows containing NAs were excluded from the model.
```

```
## Warning: There were 40 divergent transitions after warmup. Increasing adapt_delta above 0.8 may help
## http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
```

Warning: Examine the `pairs()` plot to diagnose sampling problems

I did not do anything to remove the divergent transitions in order to show the pairs plot when that happens

```
pairs(alternate$fit, pars = c("sds_sage_list_1", "sds_szip_hh_inc_list_1", "lp__"), las = 1)
```



would be better to have called `brm` with `control = list(adapt_delta = 0.9)` or so.

As might be anticipated, this model is expected to predict future data much better

```
loo_subsample(post, alternate)
```

Warning: Different subsamples in 'alternate' and 'post'. Naive diff SE is used.

Output of model 'post':

##

Computed from 4000 by 400 subsampled log-likelihood

values from 72452 total observations.

##

	Estimate	SE subsampling	SE
elpd_loo	-46478.1	84.3	0.9
p_loo	3.5	0.0	0.5
looic	92956.2	168.5	1.7

elpd_loo -46478.1 84.3 0.9

p_loo 3.5 0.0 0.5

looic 92956.2 168.5 1.7

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 'alternate':

##

Computed from 4000 by 400 subsampled log-likelihood

values from 72452 total observations.

##

```
##           Estimate      SE subsampling SE
## elpd_loo -34261.3 152.1           8.4
## p_loo      24.0   0.5           6.6
## looic      68522.7 304.1          16.9
## -----
## 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
## alternate      0.0      0.0      0.0
## post          12216.7    173.8    8.5
```

and the estimate of the treatment effect is both still positive.

alternate

```
## Warning: There were 40 divergent transitions after warmup. Increasing
## adapt_delta above 0.8 may help. See http://mc-stan.org/misc/
## warnings.html#divergent-transitions-after-warmup
```

```
## Family: bernoulli
## Links: mu = logit
## Formula: vote_presidential_2008_1 ~ numhh_list + treatment + prevote + english_list + female_list +
## Data: oregon (Number of observations: 72452)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##           total post-warmup samples = 4000
##
```

Smooth Terms:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS
sds(sage_list_1)	0.68	0.31	0.28	1.49	1.00	1270
sds(szip_hh_inc_list_1)	0.64	0.43	0.09	1.77	1.00	654

Tail_ESS

sds(sage_list_1)	1787
sds(szip_hh_inc_list_1)	637

##

Population-Level Effects:

	Estimate	Est.Error	1-95% CI	u-95% CI
Intercept	-3.14	0.05	-3.25	-3.03
numhh_listsignedselfupP1additionalperson	0.13	0.02	0.08	0.18
numhh_listsignedselfupP2additionalpeople	-0.11	0.22	-0.57	0.31
treatment	0.04	0.02	0.00	0.08
prevote	2.91	0.02	2.87	2.96
english_listRequestedEnglishmaterials	1.62	0.05	1.52	1.72
female_list1:Female	0.33	0.02	0.29	0.36
sage_list_1	1.05	0.59	-0.11	2.19
szip_hh_inc_list_1	0.13	0.68	-1.21	1.54

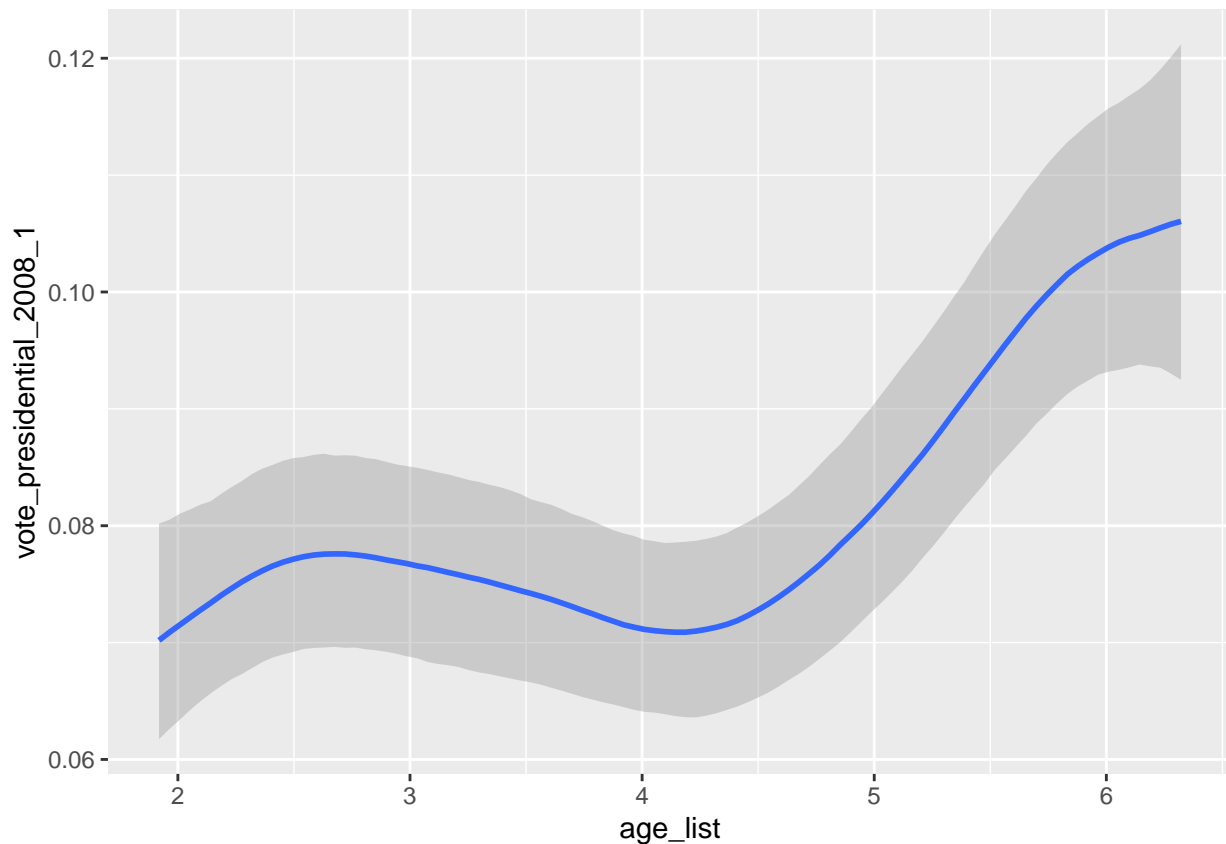
Rhat Bulk_ESS Tail_ESS

Intercept	1.00	4009	3074
numhh_listsignedselfupP1additionalperson	1.00	4874	2721
numhh_listsignedselfupP2additionalpeople	1.00	4244	2960
treatment	1.00	4017	2929
prevote	1.00	4889	2935
english_listRequestedEnglishmaterials	1.00	4374	3240

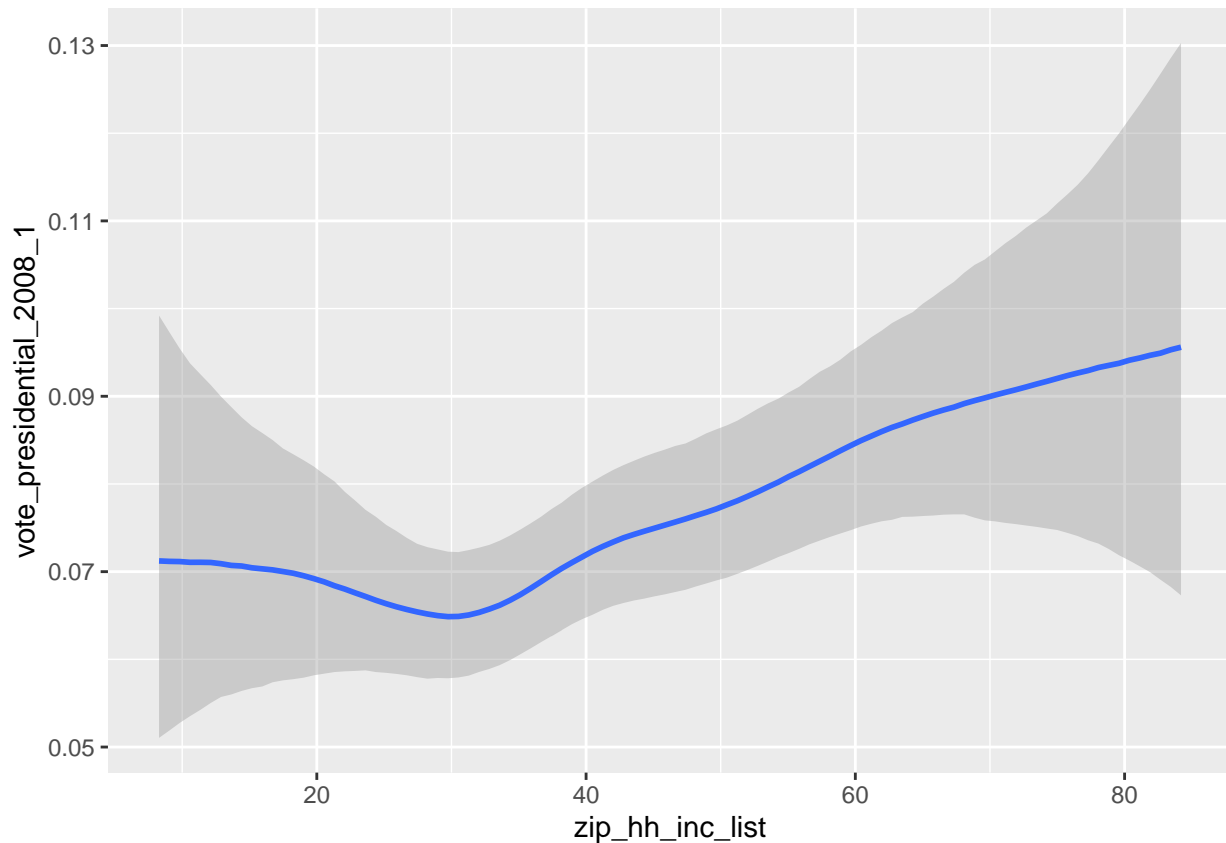
```
## female_list1:Female          1.00    3460    2651
## sage_list_1                 1.00    1996    1939
## szip_hh_inc_list_1          1.01    1385     679
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

However, in substantive terms the effect is rather small. For what it is worth, which is not much considering these are not the variables of interest, we can plot the posterior distribution of the non-linear functions relating age and income (in the person's zip code) to the probability of voting.

```
conditional_effects(alternate, effects = "age_list")
```



```
conditional_effects(alternate, effects = "zip_hh_inc_list")
```



2 Coronavirus in NYC

```
ROOT <- "https://raw.githubusercontent.com/nychealth"
NYC <- readr::read_csv(paste0(ROOT, "/coronavirus-data/master/case-hosp-death.csv"))
NYC$day <- 1:nrow(NYC)
```

2.1 Negative Binomial Model

```
nb <- brm(CASE_COUNT ~ poly(day, degree = 2), data = NYC, family = negbinomial,
  prior = prior(normal( 1, 0.50), class = "b", coef = "polydaydegreeEQ21") +
  prior(normal(-0.5, 0.25), class = "b", coef = "polydaydegreeEQ22") +
  prior(normal(5, 3), class = "Intercept") +
  prior(exponential(1), class = "shape"))
```

2.2 Poisson Model

```
po <- update(nb, family = poisson)
```

2.3 Model Comparison

First, the Poisson model is a special case of the negative binomial model as the overdispersion (shape) parameter goes to infinity. Clearly, its posterior distribution is small, indicating considerable overdispersion relative to a Poisson model

```
nb
```

```
## Family: negbinomial
## Links: mu = log; shape = identity
## Formula: CASE_COUNT ~ poly(day, degree = 2)
## Data: NYC (Number of observations: 61)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##           total post-warmup samples = 4000
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept           7.86      0.15    7.57    8.16 1.00    3898    2720
## polydaydegreeEQ21     1.19      0.48    0.24    2.13 1.00    4396    3117
## polydaydegreeEQ22    -0.71      0.25   -1.20   -0.21 1.00    4256    3092
##
## Family Specific Parameters:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## shape           0.76      0.12    0.55    1.01 1.00    3805    2958
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Second, the Pareto k estimates for the negative binomial model are all fine

```
loo(nb)
```

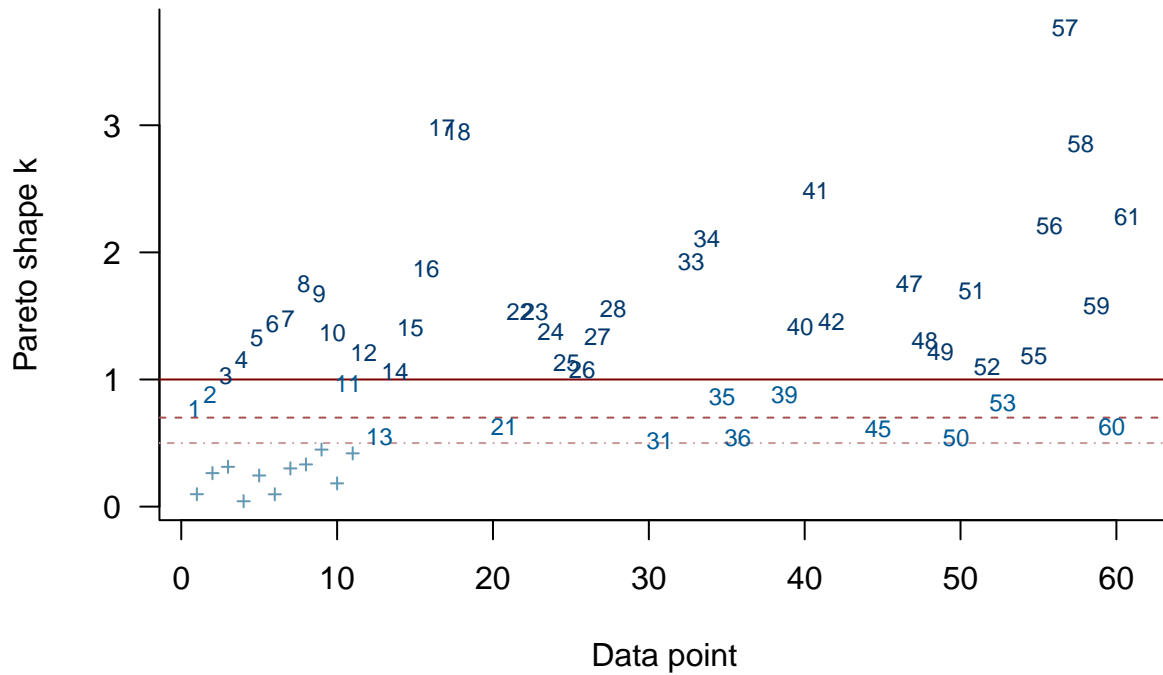
```
##
## Computed from 4000 by 61 log-likelihood matrix
##
##           Estimate   SE
## elpd_loo   -540.7  8.1
## p_loo        2.4  0.6
## looic       1081.3 16.2
## -----
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

whereas that is not true for the Poisson model, indicating that its posterior distribution is sensitive to particular observations

```
plot(loo(po), label_points = TRUE)
```

```
## Warning: Found 43 observations with a pareto_k > 0.7 in model 'po'. With this
## many problematic observations, it may be more appropriate to use 'kfold' with
## argument 'K = 10' to perform 10-fold cross-validation rather than L00.
```

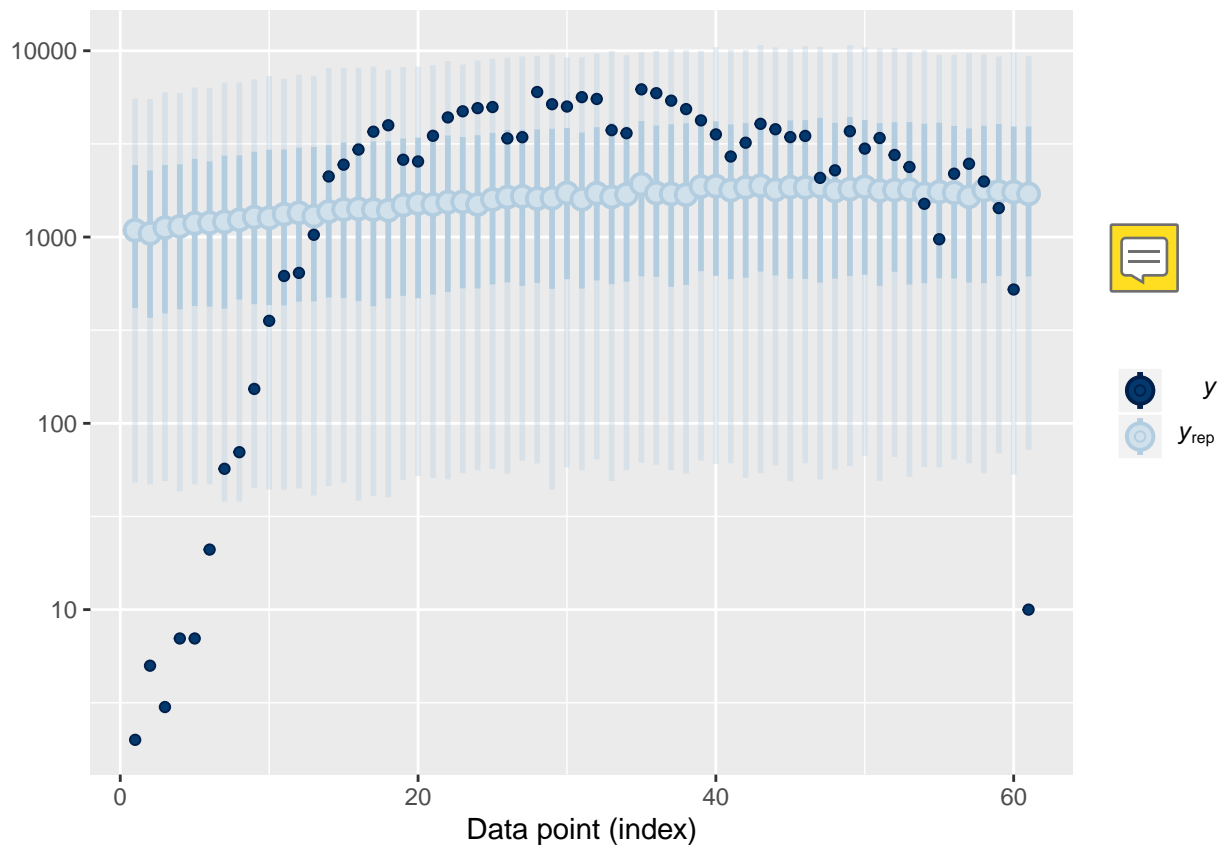

PSIS diagnostic plot



Third, although the negative binomial model seems too simplistic

```
pp_check(nb, type = "loo_intervals") + ggplot2::scale_y_continuous(trans = 'log10')

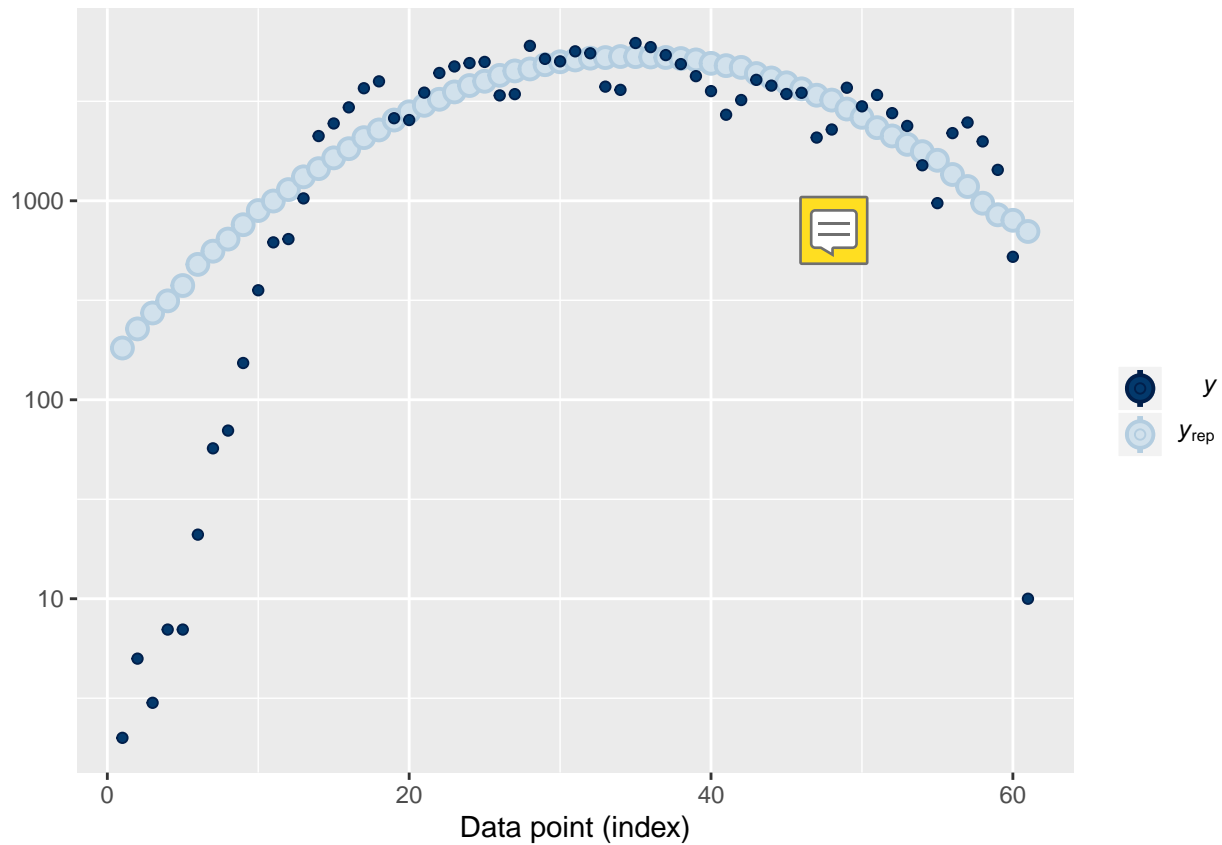
## Using all posterior samples for ppc type 'loo_intervals' by default.
```



Poisson model is way overconfident in its predictions

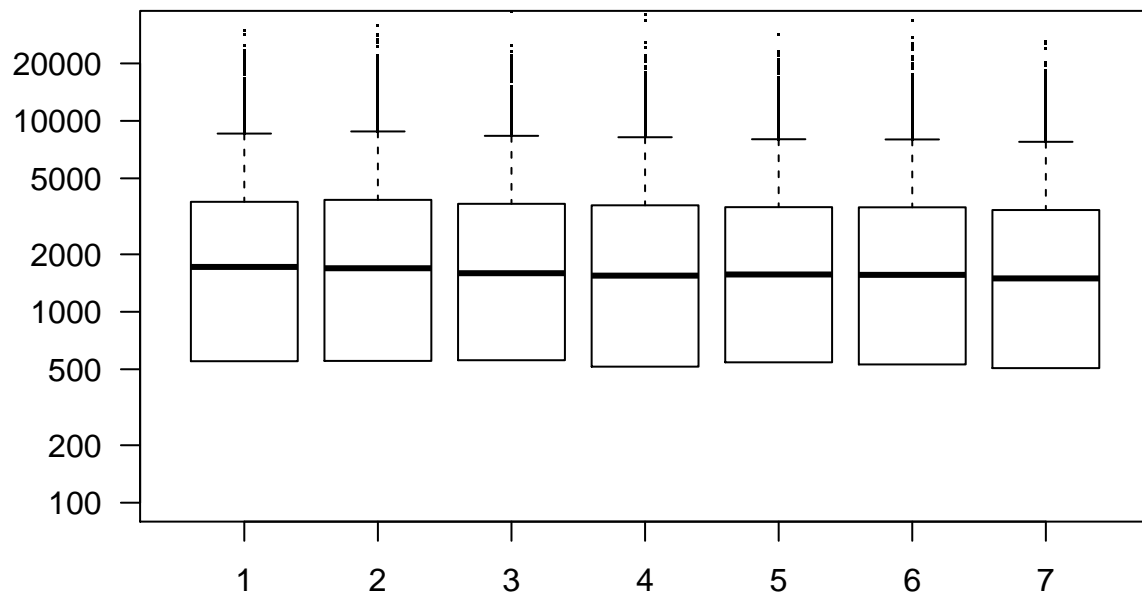
```
pp_check(po, type = "loo_intervals") + ggplot2::scale_y_continuous(trans = 'log10')
```

```
## Using all posterior samples for ppc type 'loo_intervals' by default.
```



2.4 Posterior Prediction

```
nd <- data.frame(day = (nrow(NYC) + 1):(nrow(NYC) + 7))
PPD <- posterior_predict(nb, newdata = nd)
boxplot(PPD, log = "y", pch = ".", ylim = c(100, 30000), las = 1)
```



The uncertainty is considerable since there is about a 50-50 chance that the number of new cases will be

between 500 and 3000 each day. However, this simple model seems to be reacting too slowly to better data coming out of NYC in recent days, since the posterior medians are a bit below 2000 and hopefully the number of new cases will drop below 200.

It should be pointed out that these “curve fitting” models of the coronavirus have not been very good compared to models that incorporate epidemiological theory.