Bayesian approaches of Housing price analysis

Frank Jiang

5/14/2020

1 Data preprocessing

This dataset we used in this project is the Ames Housing dataset compiled by Dean De Cock for use in data science. It is a more expanded version of the often-cited Boston Housing dataset. Our version was a split version from Kaggle, which was originally used for machine learning and advanced regression techniques training. The link are presented below https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data.

However, we decide to use the Bayesian approach to analyze this dataset and potentially create a bayesian model that could predict the housing prices base on some of the predictors available in this dataset.

To create priors and draws using the dataset, we preprocess the dataset(including creating a dummy variable for the year built and the second floor). The new predictors are generated as below. We define old in year built to be before 2008. Also, the area of the second floor is recoded to whether the original housing has a second floor or not.

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
housing_data<- read.csv('train.csv')
#create binary variable for yearbuilt
housing_data$yearbuilt_bi<- ifelse(
  housing_data$YearBuilt<2008,
  yes='old',
  no ='new'
)
housing_data$yearbuilt_bi<- as.factor(housing_data$yearbuilt_bi)

#recode whether the house has central air or not
housing_data$CentralAir<- ifelse(
```

```r
  housing_data$CentralAir=='Y',
  yes=1,
  no=0
)
housing_data$CentralAir<- as.factor(housing_data$CentralAir)

#filter sale condition to only two conditions, normal and abnormal
housing_data<- housing_data %>% filter(housing_data$SaleCondition=='Normal'| housing_data$SaleCondition=
housing_data$SaleCondition<- as.factor(as.character(housing_data$SaleCondition))

#recode area of the second floor to whether if the house has a second floor
housing_data$secondfloor_bi<- ifelse(
  housing_data$X2ndFlrSF==0,
  yes=0,
  no=1
)
housing_data$secondfloor_bi<- as.factor(housing_data$secondfloor_bi)
housing_data$Street<- as.factor(housing_data$Street)

#filter the general zone classification of the house area to only Residential High density, Low denisty
housing_data<- housing_data %>% filter(MSZoning=='RH'|MSZoning=='RL'
                                       |MSZoning=='RM')
housing_data$MSZoning<- as.character(housing_data$MSZoning)
housing_data$MSZoning<- as.factor(housing_data$MSZoning)
summary(housing_data$SalePrice)
```

1.1 Gaussian Model

In this section, we decide to draw priors from the predictors we select and use the gaussian family to draw from priors. First, we use get_prior to get a quick look at what all parameters are referred to.Then, we use brm to fit the priors predictive distribution by setting the parameter sample_prior to 'only'.

We decide to use a normal distribution for all the parameters and gradually add more predictors to the model to see how it performs. The model selection and evaluation are completed in the later section.

1.2 Priors and Priors predictive distribution with brm

```r
library(brms)
```

```
## Loading required package: Rcpp

## 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
```

```r
get_prior(SalePrice~LotArea+GrLivArea, family='Poisson',data=housing_data)
prior1<- c(set_prior('normal(3,5)',class='b',coef='GrLivArea'),
           set_prior('normal(1,5)',class='b',coef='LotArea'),
           set_prior('normal(10000,1000)',class='Intercept'))
ppd1<- brm(SalePrice~LotArea+GrLivArea,family='gaussian',data=housing_data,prior=prior1,sample_prior='o
```

```
## Compiling the C++ model

## Start sampling
```

```r
get_prior(SalePrice~log(LotArea)+log(GrLivArea),family='gaussian',data=housing_data)
prior2<- c(set_prior('normal(7,10)',class='b',coef='logGrLivArea'),
           set_prior('normal(2,10)',class='b',coef='logLotArea'),
           set_prior('normal(10000,1000)',class='Intercept'))
ppd2<- brm(SalePrice~log(LotArea)+log(GrLivArea),family='gaussian',data=housing_data,prior=prior2,sampl
```

```
## Compiling the C++ model
## Start sampling
```

```r
get_prior(SalePrice~yearbuilt_bi+log(LotArea)+MSZoning+Street+secondfloor_bi+CentralAir+secondfloor_bi+
prior3<- c(set_prior('normal(0.4,1)',class='b',coef='CentralAir1'),
           set_prior('normal(5,10)',class='b',coef='logGrLivArea'),
           set_prior('normal(2,10)',class='b',coef='logLotArea'),
           set_prior('normal(0.2,1)',class='b',coef='MSZoningRL'),
           set_prior('normal(0.1,1)',class='b',coef='MSZoningRM'),
           set_prior('normal(0,0.5)',class='b',coef='SaleConditionNormal'),
           set_prior('normal(0.8,1)',class='b',coef='secondfloor_bi1'),
           set_prior('normal(0.1,0.8)',class='b',coef='StreetPave'),
           set_prior('normal(-0.1,0.3)',class='b',coef='yearbuilt_biold'),
           set_prior('normal(10000,1000)',class='Intercept'))
ppd3<- brm(SalePrice~yearbuilt_bi+log(LotArea)+MSZoning+Street+secondfloor_bi+CentralAir+secondfloor_bi
```

```
## Compiling the C++ model
## Start sampling
```

1.3 Basic Pipeline Model Selection

In this section, we did some basic overview of the fitted prior predictive distribution, and see if the fitted prior predictive distributions are all plausible.

```r
summary(ppd1)
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: SalePrice ~ LotArea + GrLivArea
##    Data: housing_data (Number of observations: 1247)
## 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 -3999.26  53207.61 -108604.33 97950.76 1.00     4632     2841
## LotArea       0.91      4.95      -8.71    10.79 1.00     4572     2924
## GrLivArea     2.94      4.97      -6.85    12.84 1.00     3829     2768
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI  u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma 55318.59  58406.26  2068.11 211075.75 1.00     3361     2113
##
## 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).
```

```r
summary(ppd2)
```

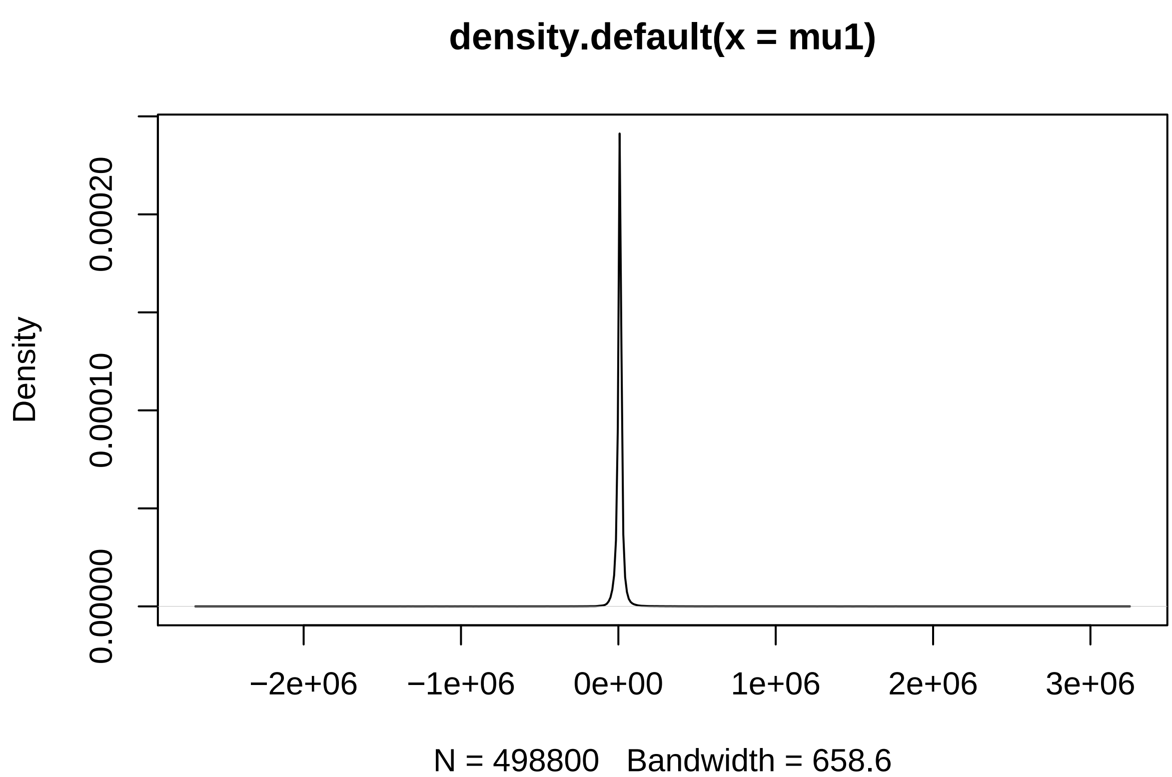
```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: SalePrice ~ log(LotArea) + log(GrLivArea)
##    Data: housing_data (Number of observations: 1247)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Population-Level Effects:
##              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     9934.68   1010.71  7963.68 11911.88 1.00     4579     2591
## logLotArea       1.84     10.07   -17.55    21.60 1.00     4089     2762
## logGrLivArea     7.20     10.23   -12.82    26.78 1.00     4078     2754
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI  u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma 53540.95  54522.29  1949.04 194161.71 1.00     3863     1756
##
## 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).
```

```r
summary(ppd3)
```

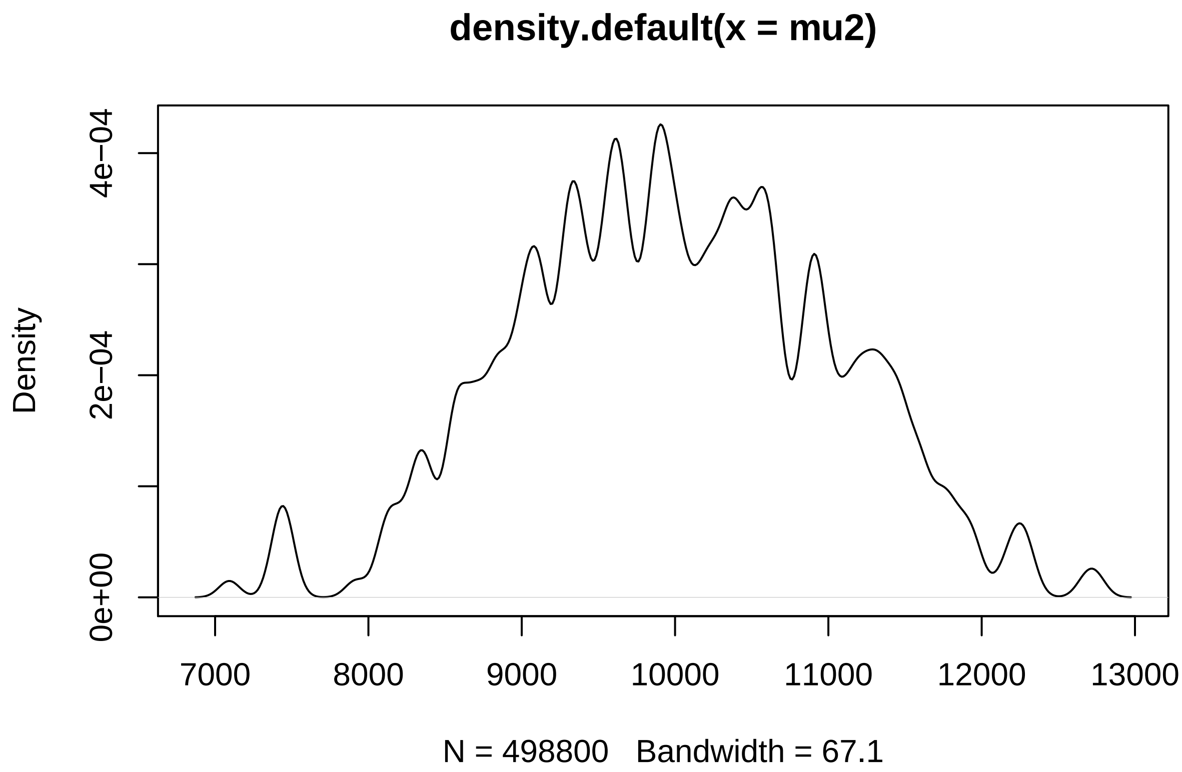
```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: SalePrice ~ yearbuilt_bi + log(LotArea) + MSZoning + Street + secondfloor_bi + CentralAir +
##    Data: housing_data (Number of observations: 1247)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Population-Level Effects:
##                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept            9942.16   1013.47  7951.86 11917.56 1.00     9363     2805
## yearbuilt_biold        -0.10      0.30    -0.69     0.49 1.00     9386     2850
## logLotArea              2.13      9.88   -17.16    21.00 1.00     8967     3037
## MSZoningRL              0.20      0.99    -1.71     2.11 1.00     9678     2795
## MSZoningRM              0.10      1.00    -1.90     2.02 1.00     8216     3088
## StreetPave              0.10      0.78    -1.41     1.62 1.00     9105     3300
```

```
## secondfloor_bi1          0.81      1.03     -1.23     2.90 1.00     8923     2604
## CentralAir1              0.40      1.02     -1.60     2.37 1.00     9411     2949
## logGrLivArea             4.98     10.29    -15.23    24.57 1.00     9511     2586
## SaleConditionNormal     -0.00      0.50     -0.99     0.97 1.00     8690     2644
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI  u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma 55624.57  64050.56  1908.51 211342.79 1.00     5967     2515
##
## 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).
```

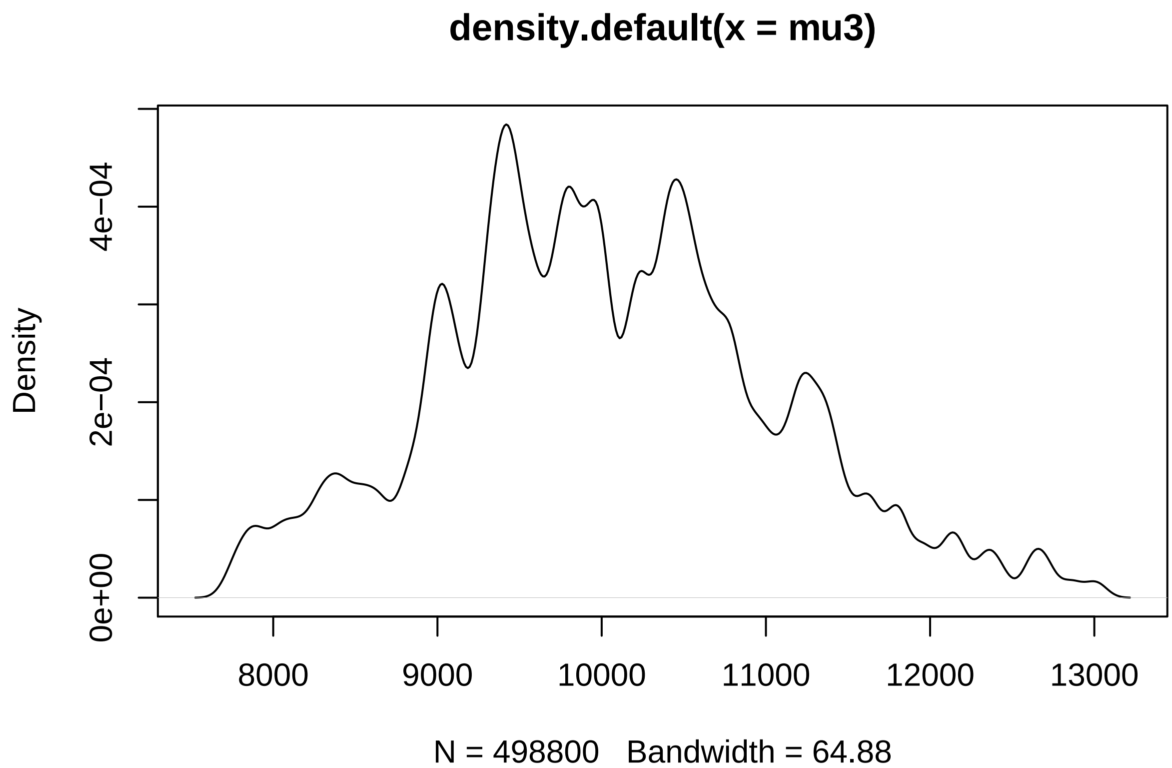
```r
mu1<- pp_expect(ppd1,nsamples=400)
plot(density(mu1))
```




```r
mu2<- pp_expect(ppd2,nsamples=400)
plot(density(mu2))
```




```
mu3<- pp_expect(ppd3,nsamples=400)
plot(density(mu3))
```



From the density graph, we can conclude that though it is not a super smooth normal distribution, all three models which include the prior predictive distribution are plausible. However, model 2 and 3 which uses the log transformation of the living area and lot area predictors performs better than only including the original variables.

1.4 Negative Binomial Model

In this section, we decide to fit the prior predictive distribution using the negative binomial distribution. In order to compare prior predictive distribution using negative binomial distribution and gaussian distribution, we decide to use the same predictors as the model above.

1.5 Priors and Priors predictive distribution with brm

```
get_prior(SalePrice~yearbuilt_bi+log(LotArea)+MSZoning+Street+secondfloor_bi+CentralAir+secondfloor_bi+
prior4<- c(set_prior('normal(0.4,1)',class='b',coef='CentralAir1'),
           set_prior('normal(5,10)',class='b',coef='logGrLivArea'),
           set_prior('normal(2,10)',class='b',coef='logLotArea'),
           set_prior('normal(0.2,1)',class='b',coef='MSZoningRL'),
           set_prior('normal(0.1,1)',class='b',coef='MSZoningRM'),
           set_prior('normal(0,0.5)',class='b',coef='SaleConditionNormal'),
           set_prior('normal(0.8,1)',class='b',coef='secondfloor_bi1'),
           set_prior('normal(0.1,0.8)',class='b',coef='StreetPave'),
           set_prior('normal(-0.1,0.3)',class='b',coef='yearbuilt_biold'),
```

```r
          set_prior('normal(10000,1000)',class='Intercept'))
ppd4<- brm(SalePrice~yearbuilt_bi+log(LotArea)+MSZoning+Street+secondfloor_bi+CentralAir+secondfloor_bi
```

```
## Compiling the C++ model
```

```
## Start sampling
```

1.6 Model Plausibility check

```r
summary(ppd4)
```

```
## Warning: There were 2788 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: negbinomial
##   Links: mu = log; shape = identity
## Formula: SalePrice ~ yearbuilt_bi + log(LotArea) + MSZoning + Street + secondfloor_bi + CentralAir +
##    Data: housing_data (Number of observations: 1247)
## 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            9935.03   1032.86  7891.96 11953.49 1.00      839     1108
## yearbuilt_biold        -0.11      0.31    -0.74     0.49 1.00      879      994
## logLotArea              1.75     10.08   -17.63    21.51 1.01      868     1250
## MSZoningRL              0.16      1.02    -1.81     2.18 1.00      872     1311
## MSZoningRM              0.10      0.97    -1.88     1.96 1.00      966     1214
## StreetPave              0.10      0.81    -1.47     1.73 1.01      745      930
## secondfloor_bi1         0.78      1.03    -1.27     2.83 1.01      906     1006
## CentralAir1             0.49      1.00    -1.47     2.40 1.00      804     1295
## logGrLivArea            5.31     10.10   -14.23    24.15 1.00     1047     1454
## SaleConditionNormal    -0.01      0.51    -1.00     1.04 1.00      881      987
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## shape     0.96      8.89     0.00     2.33 1.01      487      823
##
## 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).
```

```r
mu4<- pp_expect(ppd4,nsamples=400)
```

Here, we did a similar plausibility check of the prior predictive distribution by plotting the density plot. It can be concluded that the prior draw did not violate any prior assumption.

2 Posterior Distribution

After confirming that all three priors draw are plausible, we fit the posterior distribution using the three priors we choose as below.

```
post_gaussian_1<- brm(SalePrice~log(LotArea)+log(GrLivArea),family='gaussian',data=housing_data,prior=pr
```

```
## Compiling the C++ model

## recompiling to avoid crashing R session

## Start sampling
```

```
post_gaussian_2<- brm(SalePrice~yearbuilt_bi+log(LotArea)+MSZoning+Street+secondfloor_bi+CentralAir+seco
```

```
## Compiling the C++ model

## recompiling to avoid crashing R session

## Start sampling
```

```
post_negbinom<- brm(SalePrice~yearbuilt_bi+log(LotArea)+MSZoning+Street+secondfloor_bi+CentralAir+secon
```

```
## Compiling the C++ model

## recompiling to avoid crashing R session

## Start sampling
```

```
summary(post_gaussian_1)
summary(post_gaussian_2)
summary(post_negbinom)
```

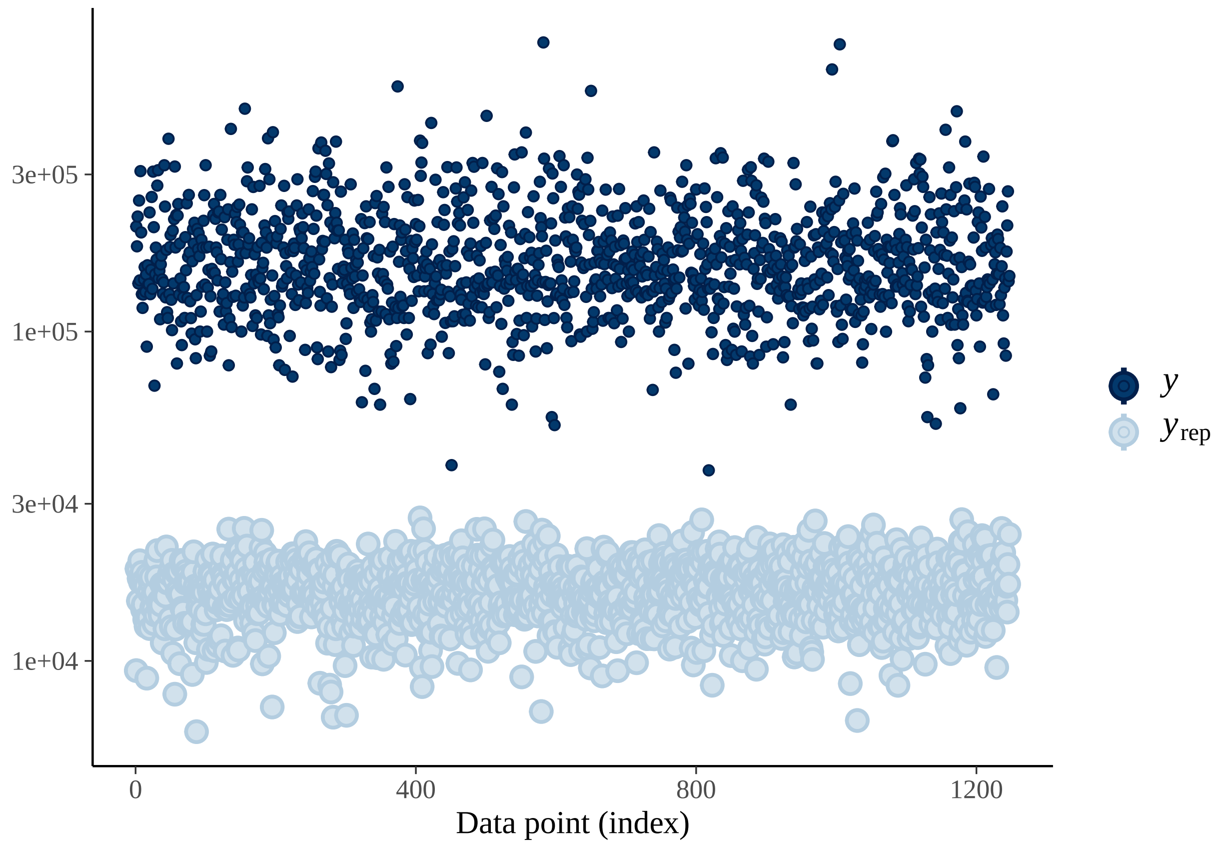
Despite our skeptical error, We can tell that from our results, we did not observe any data failed to convergence from all three posterior predictive distribution. Neither do we find an ineffective sample size and non-potential Rhat value.

Here we plot the posterior predictive distributions (in light blue) plotted against the true distribution of the outcome variable. A separate dot plot and density plot are both presented as below.

```
pp_check(post_gaussian_1,type='loo_intervals')+ggplot2::scale_y_continuous(trans='log10')
```

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

## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous y-axis
```

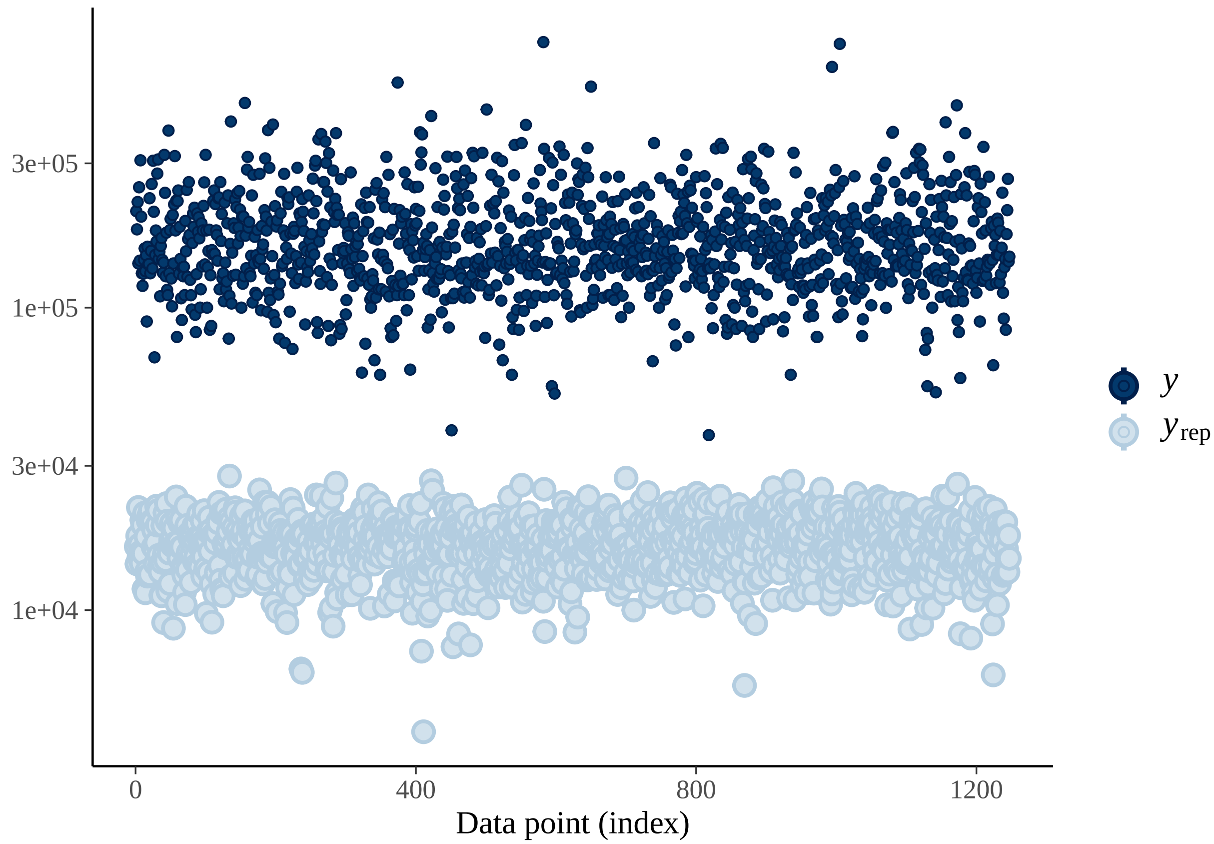
```
## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 1247 rows containing missing values (geom_segment).

## Warning: Removed 1247 rows containing missing values (geom_segment).
```




```
pp_check(post_gaussian_1,type='loo_intervals')+ggplot2::scale_y_continuous(trans='log10')
```

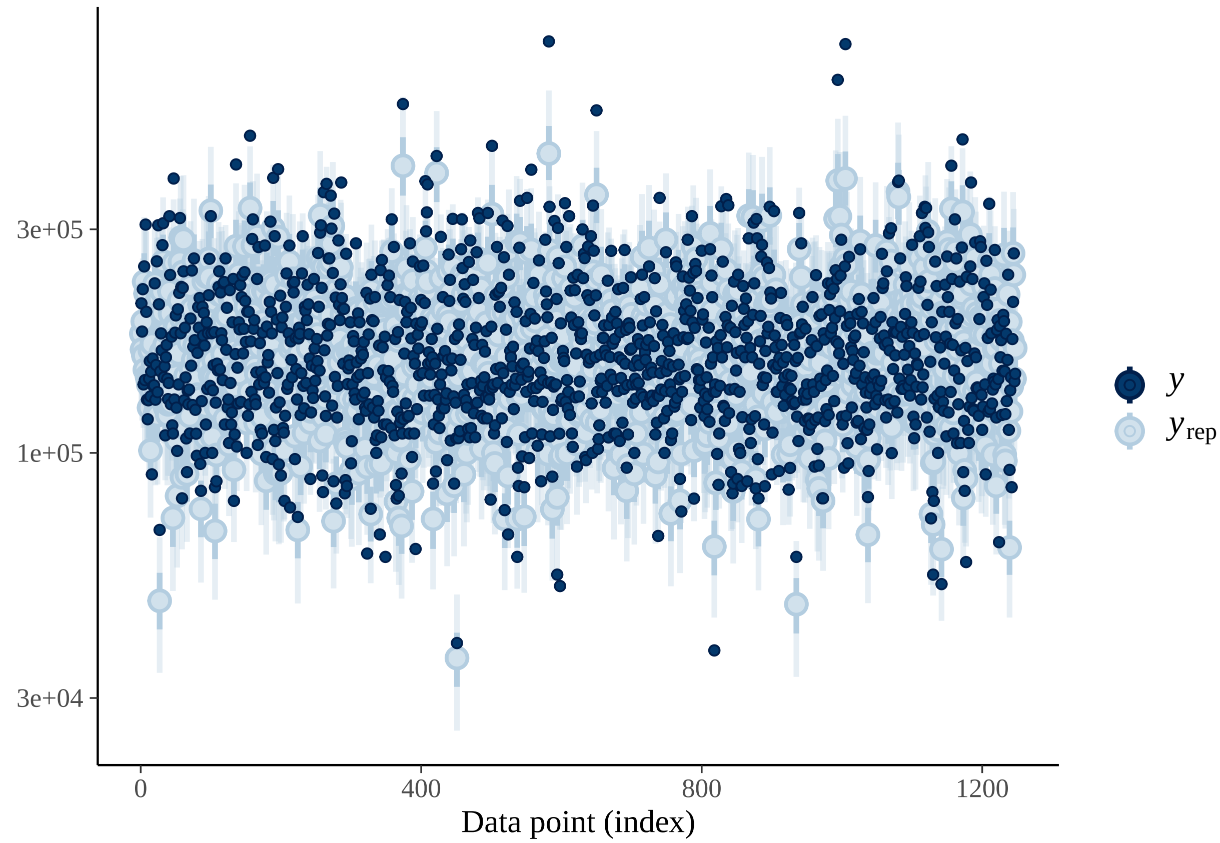
```
## Using all posterior samples for ppc type 'loo_intervals' by default.

## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 1247 rows containing missing values (geom_segment).

## Warning: Removed 1247 rows containing missing values (geom_segment).
```



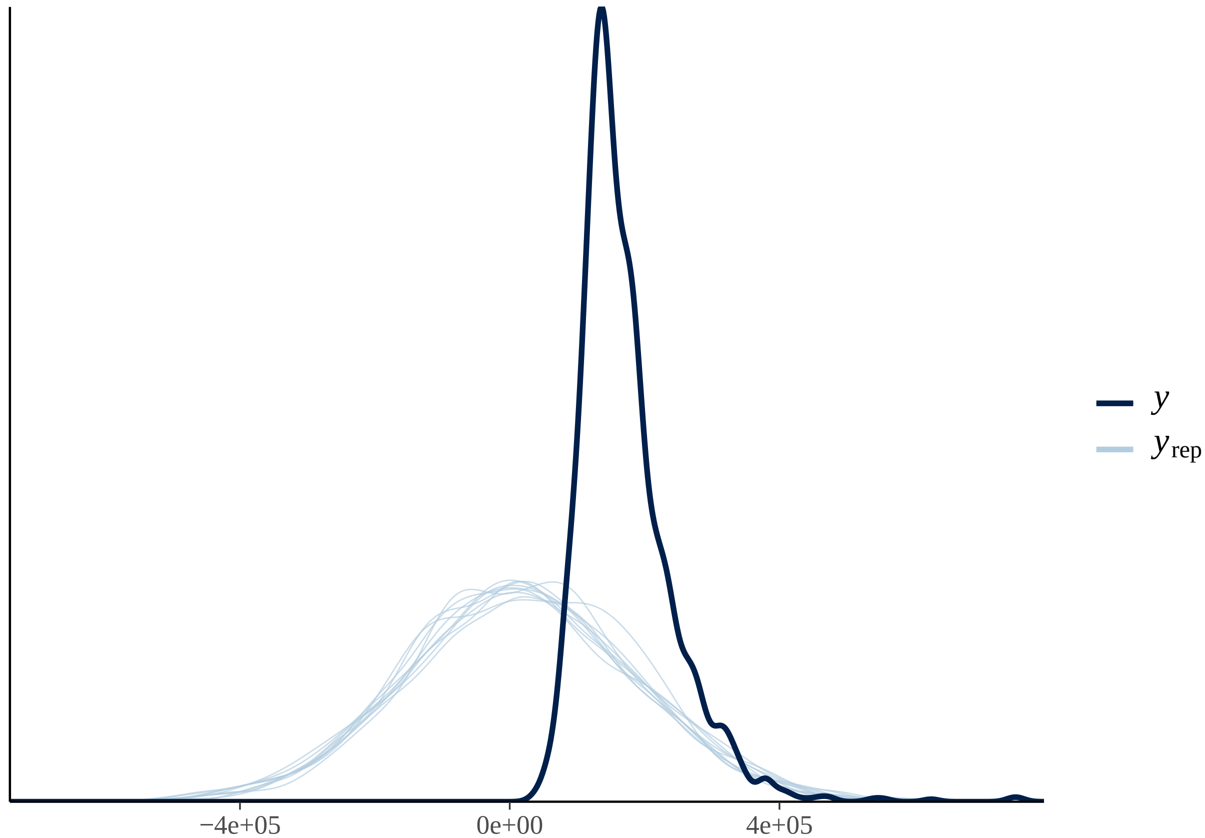

```r
pp_check(post_negbinom,type='loo_intervals')+ggplot2::scale_y_continuous(trans='log10')
```

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



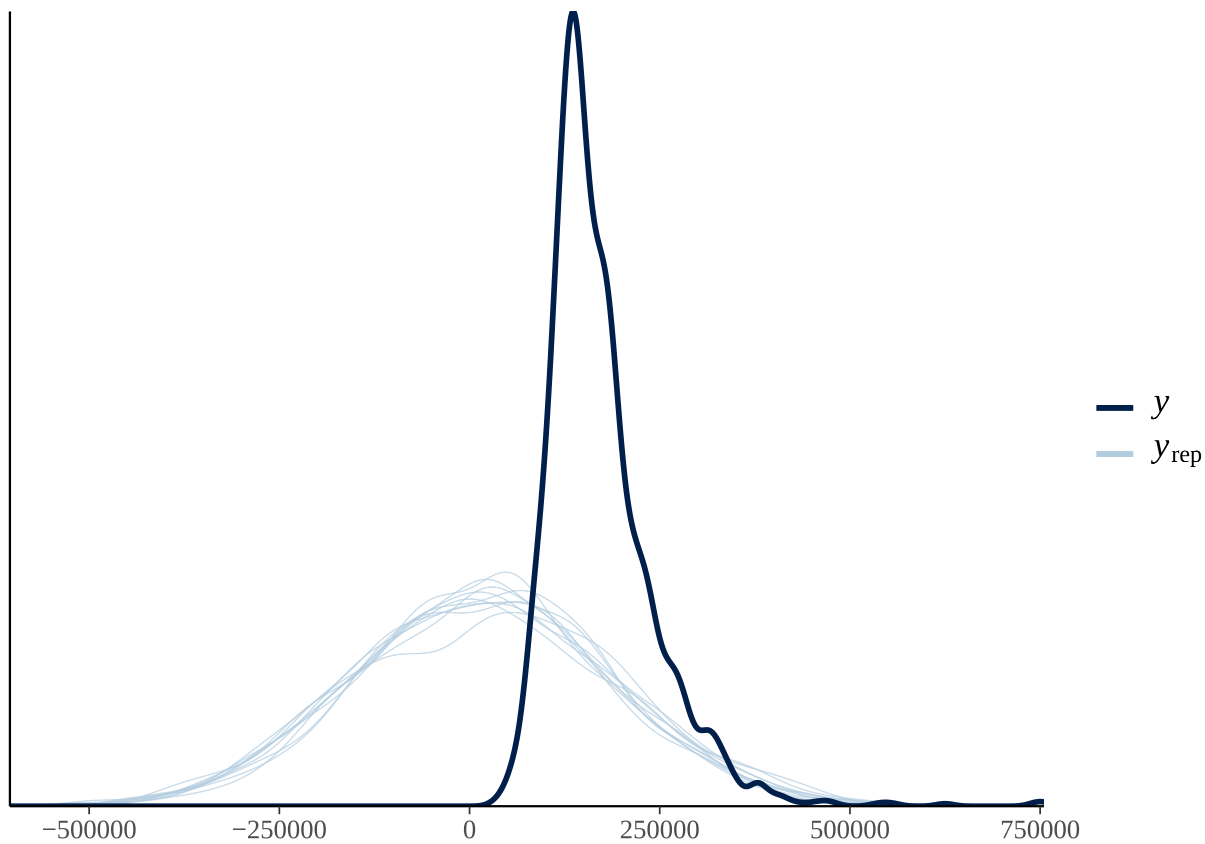

```r
pp_check(post_gaussian_1,plotfun='dens_overlay',nreps=15)
```

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

## Warning: The following arguments were unrecognized and ignored: plotfun, nreps
```




```r
pp_check(post_gaussian_2,plotfun='dens_overlay',nreps=15)
```

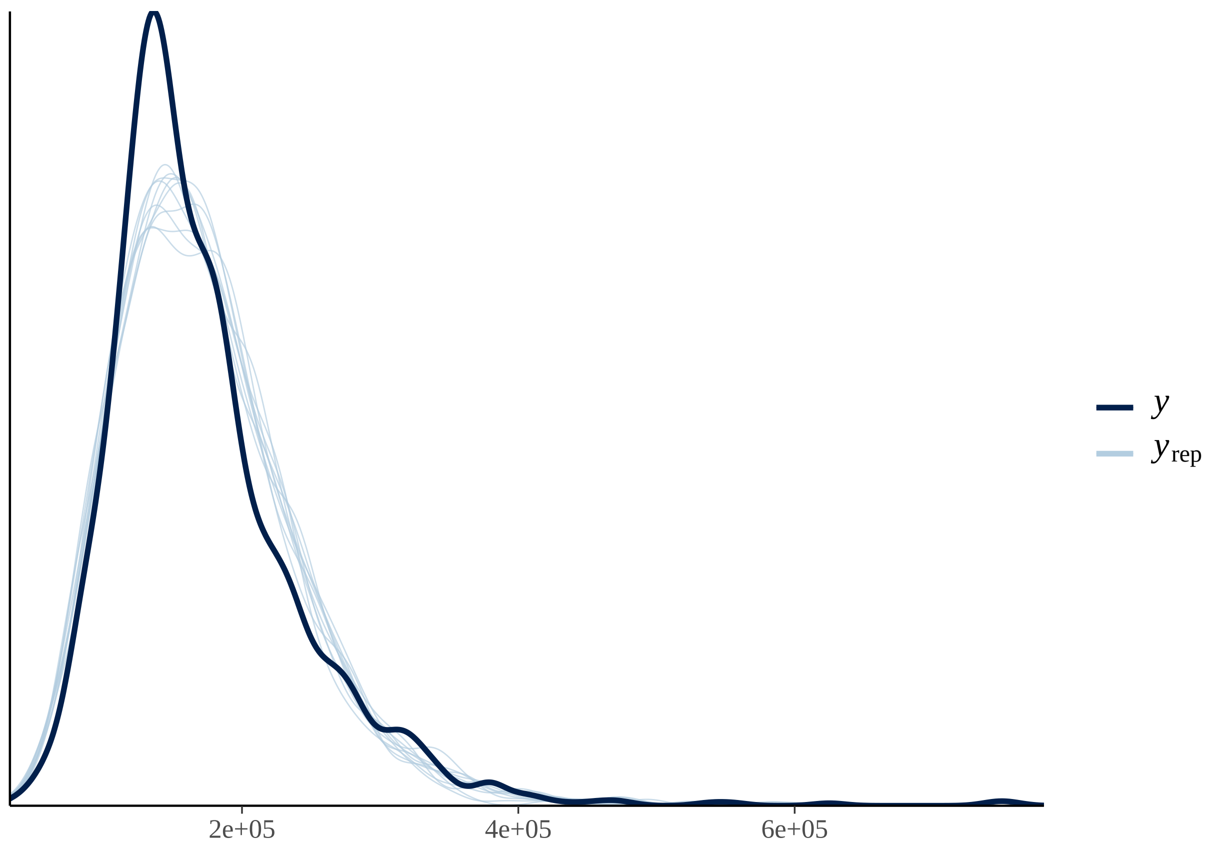
```
## Using 10 posterior samples for ppc type 'dens_overlay' by default.

## Warning: The following arguments were unrecognized and ignored: plotfun, nreps
```




```r
pp_check(post_negbinom,plotfun='dens_overlay',nreps=15)
```

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

```
## Warning: The following arguments were unrecognized and ignored: plotfun, nreps
```



To our surprise, we can carefully conclude that the previous two models did not represent the original data well. However, we still cannot make a decision on which model should we choose yet.

3 Model Decision and interpretation

```
loo_compare(loo(post_gaussian_1),loo(post_gaussian_2),loo(post_negbinom))
```

```
##                 elpd_diff se_diff
## post_negbinom       0.0       0.0
## post_gaussian_1 -2089.9      27.3
## post_gaussian_2 -2090.0      27.3
```

```
loo(post_negbinom)
```

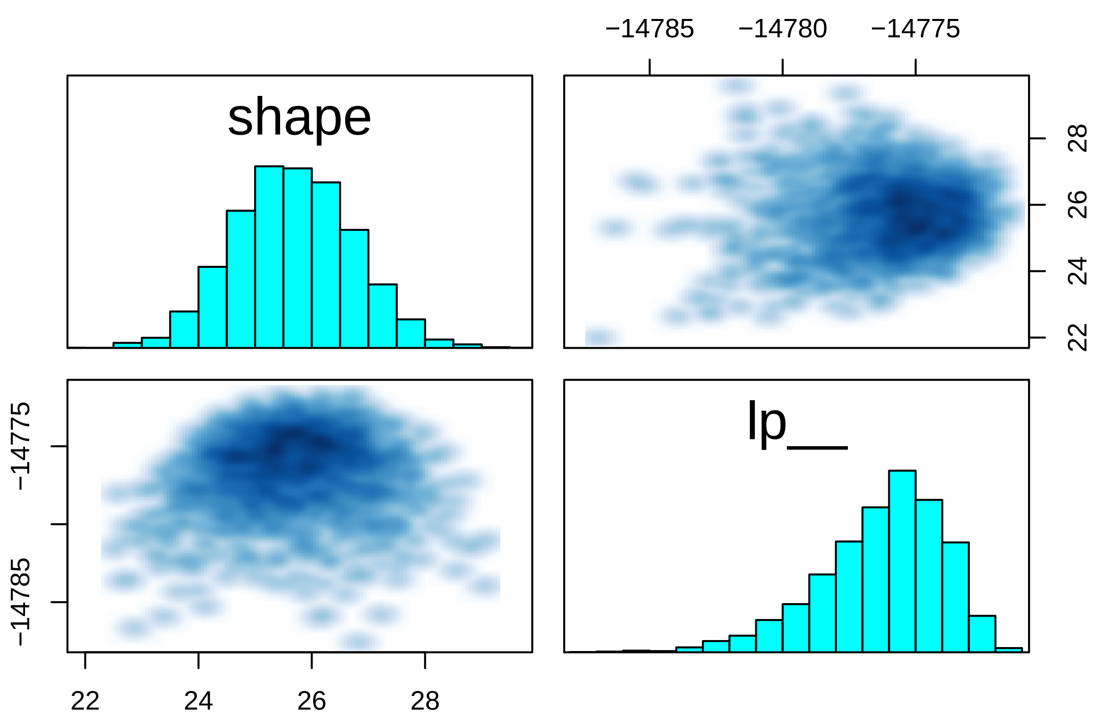
```
##
## Computed from 4000 by 1247 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo -14708.8 31.8
## p_loo        13.0  1.4
## looic     29417.5 63.6
## ------
## Monte Carlo SE of elpd_loo is 0.1.
```

```
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

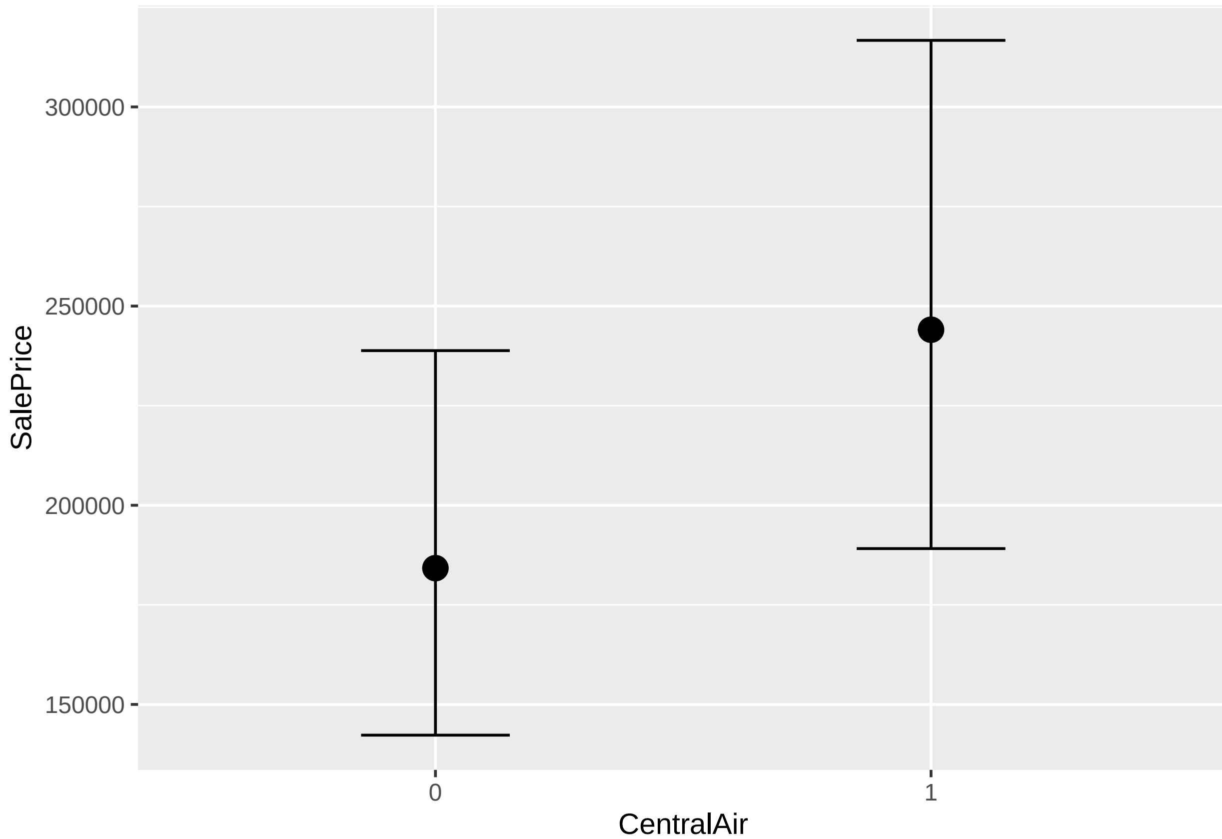
After comparing three models using the Expected Log Predictive Density(ELPD) as the criterion, we can find out that the difference between the model complicated model using Gaussian distribution and simpler model is not significant. Thus, we should choose the simpler model between those two. However, the model using negative binomial distribution is significantly ahead. Thus, we decide to choose the negative binomial model to pursue further analysis.

Here, we focus on the conditional effect of different predictors, which have a relatively small effect on the house pricing comparing to the living area and lot area. We can plot the posterior distribution of the effect of those variables.

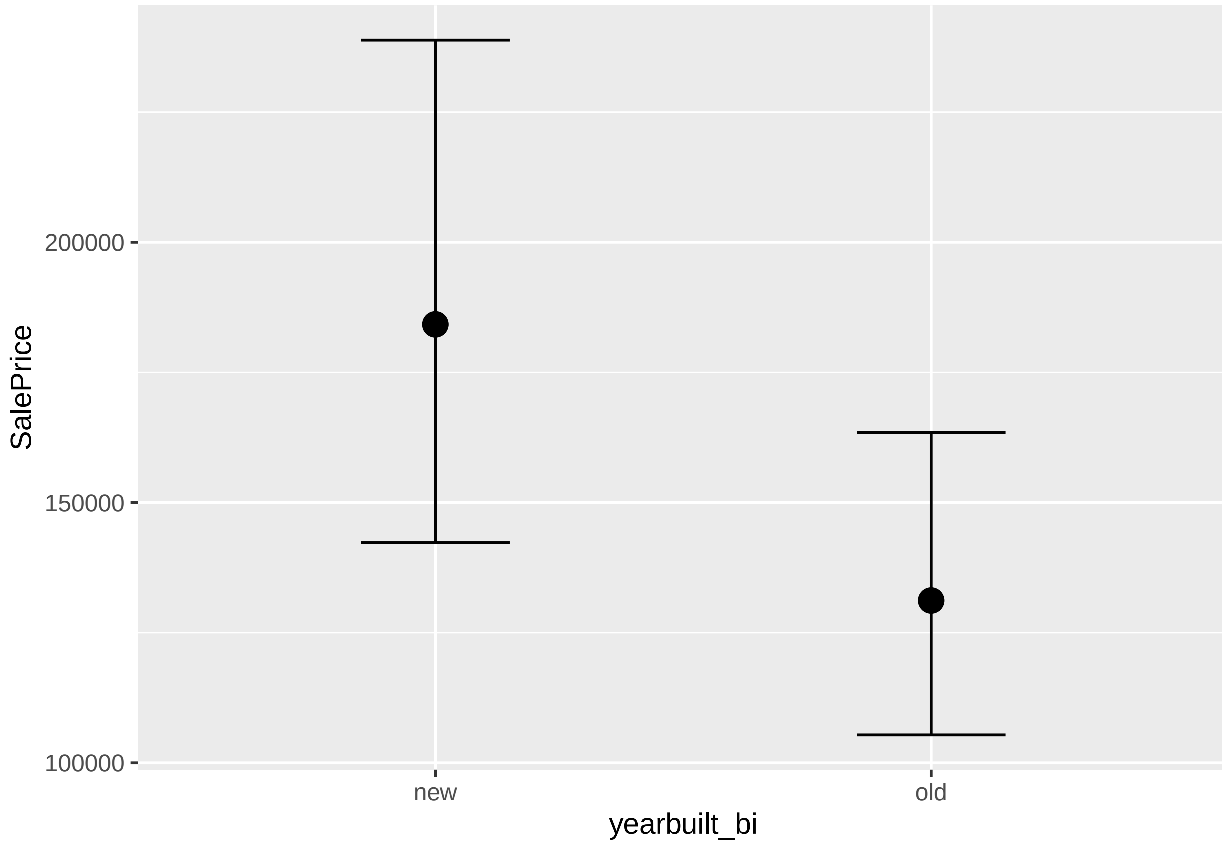
```r
pairs(post_negbinom$fit,pars = c('shape','lp__'))
```



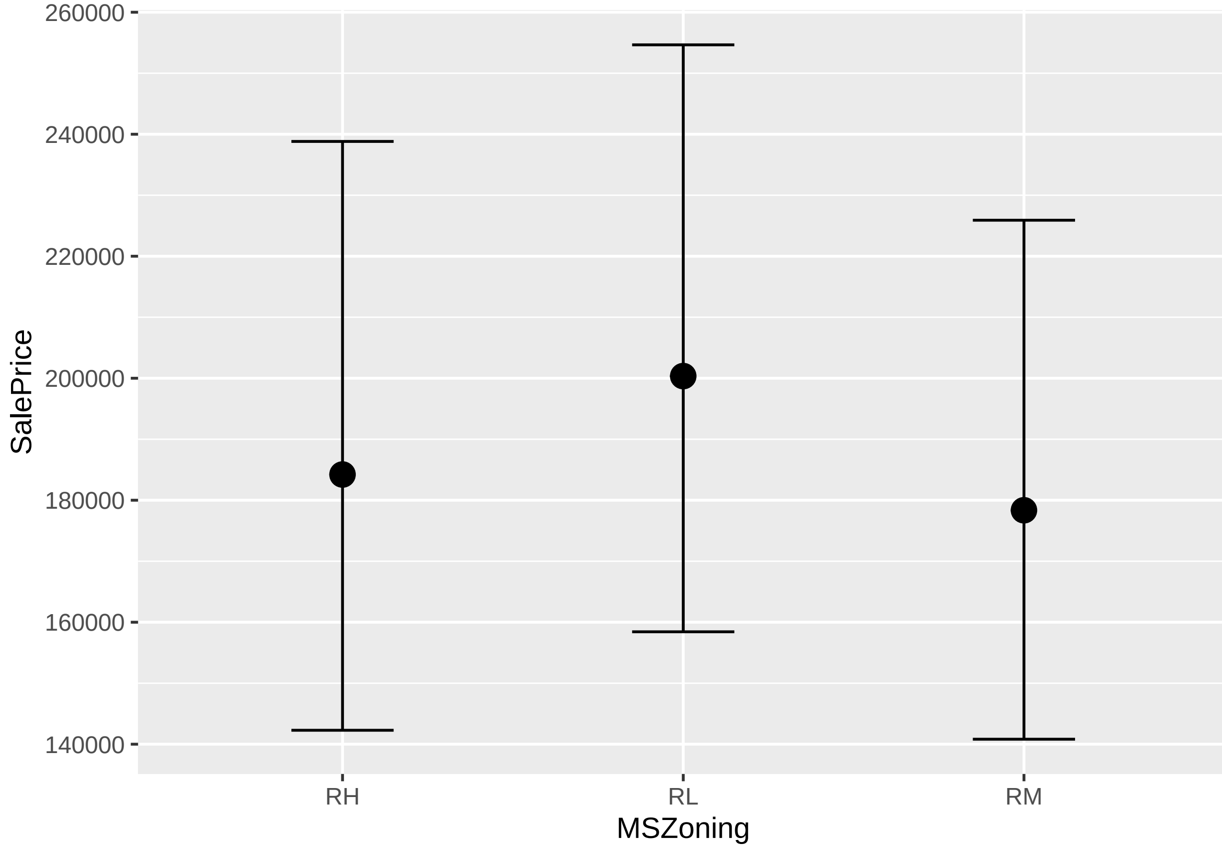

```r
conditional_effects(post_negbinom,effects='CentralAir')
```



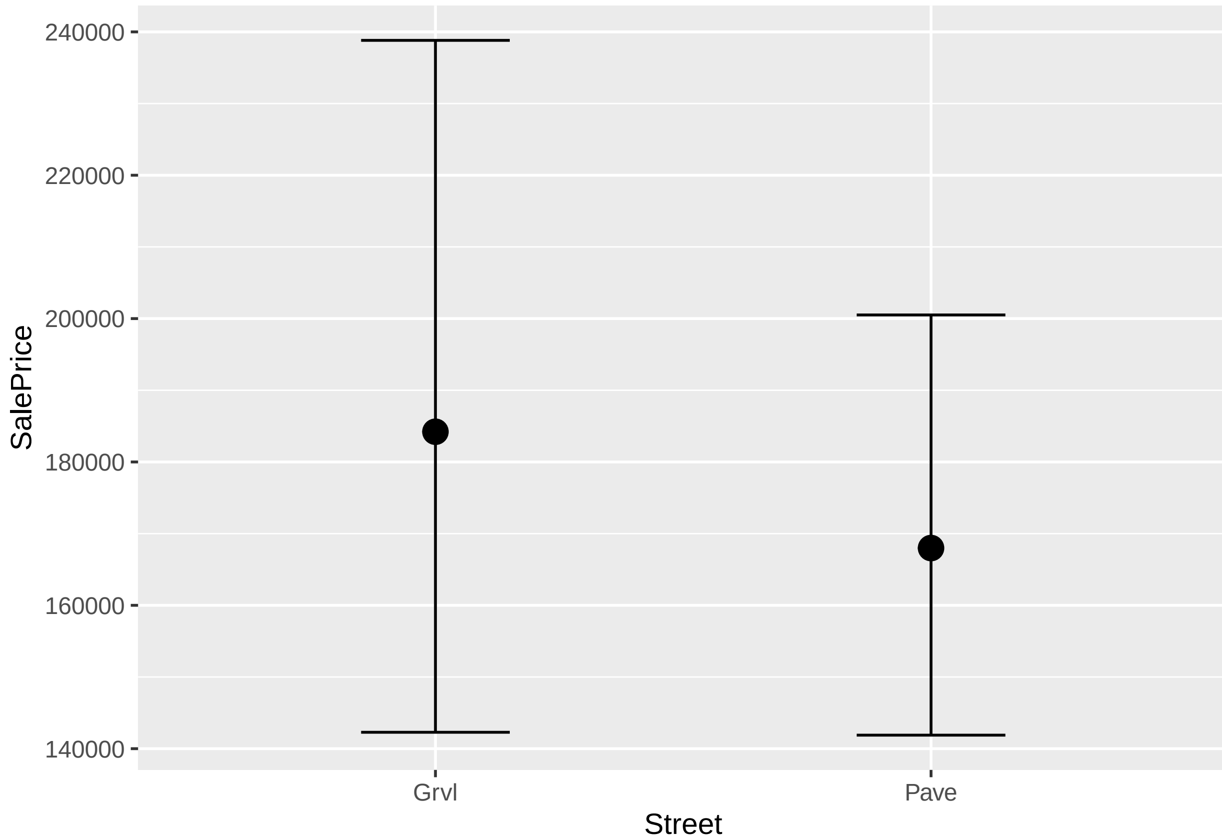

```r
conditional_effects(post_negbinom,effects='yearbuilt_bi')
```



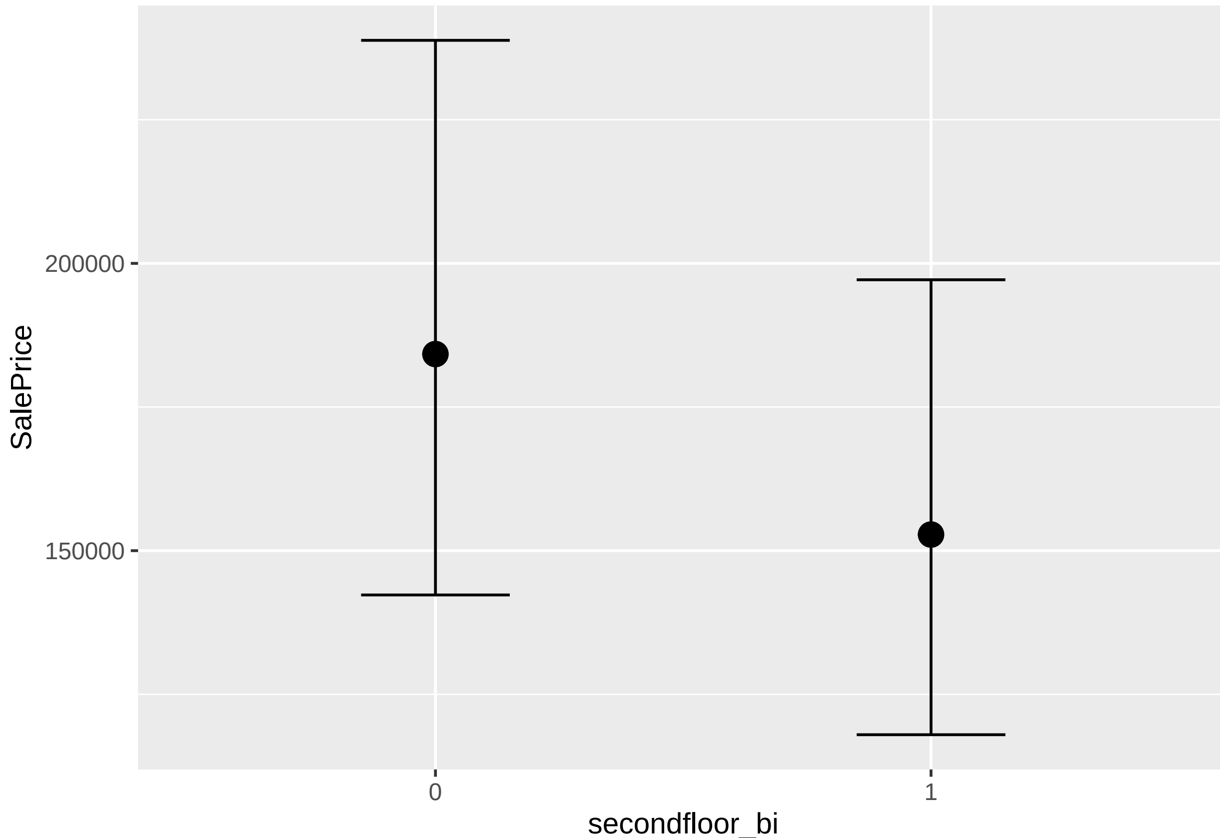

```
conditional_effects(post_negbinom,effects='MSZoning')
```




```
conditional_effects(post_negbinom,effects='Street')
```




```
conditional_effects(post_negbinom,effects='secondfloor_bi')
```



There are a few things that are interesting to mention. First of all, as expected, houses with a central air system worth more in value. Also, the house that is built prior to 2008 has a noticeably decrease in its sale price than those who built after 2008. Contrary to what we previously thought, the house with a second floor has a lower sale price comparing to those with only one floor. What's more, the house with a front street of gravel rather than paved end up in higher sale price. Last but not least, houses in the residential low-density area have a higher sale price, while there seems to not be a difference between houses in residential medium-density and high-density areas.

Here, we test the posterior probability that certain predictors increase a house sale price. We find out that the Living area, lot area, and central air has a probability of 1, while street pave has a value of 0.17 and second floor with a value of 0. However, these do not show a causal effect between these predictors and the outcome variable which is the sale price of the house.

```
hypothesis(post_negbinom,hypothesis='logGrLivArea > 0')
```

```
## Hypothesis Tests for class b:
##            Hypothesis Estimate Est.Error CI.Lower CI.Upper Evid.Ratio Post.Prob
## 1 (logGrLivArea) > 0     0.91      0.02     0.87     0.95        Inf         1
##   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.
```

```r
hypothesis(post_negbinom,hypothesis='logLotArea>0')
```

```
## Hypothesis Tests for class b:
##         Hypothesis Estimate Est.Error CI.Lower CI.Upper Evid.Ratio Post.Prob
## 1 (logLotArea) > 0     0.05      0.01     0.03     0.07        Inf         1
##   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.
```

```r
hypothesis(post_negbinom,hypothesis='CentralAir1>0')
```

```
## Hypothesis Tests for class b:
##          Hypothesis Estimate Est.Error CI.Lower CI.Upper Evid.Ratio Post.Prob
## 1 (CentralAir1) > 0     0.28      0.02     0.24     0.32        Inf         1
##   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.
```

```r
hypothesis(post_negbinom,hypothesis='StreetPave>0')
```

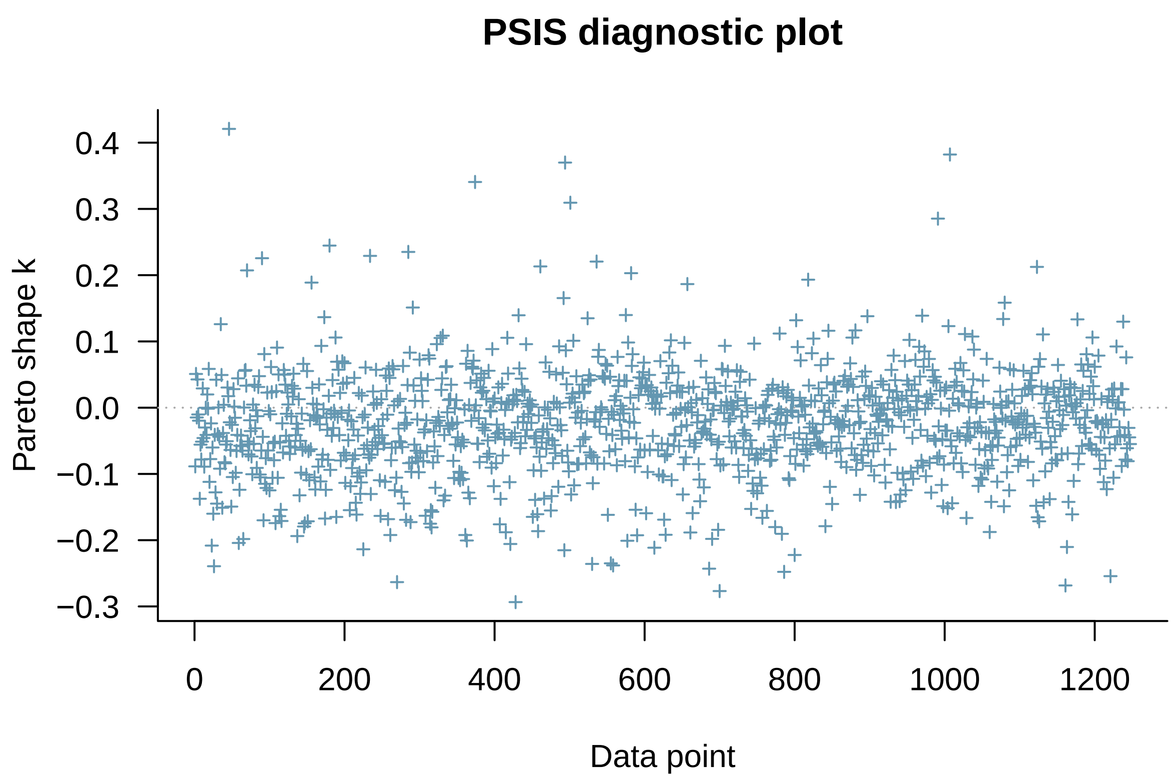
```
## Hypothesis Tests for class b:
##         Hypothesis Estimate Est.Error CI.Lower CI.Upper Evid.Ratio Post.Prob
## 1 (StreetPave) > 0    -0.09       0.1    -0.25     0.07       0.21      0.18
##   Star
## 1
## ---
## 'CI': 90%-CI for one-sided and 95%-CI for two-sided hypotheses.
## '*': For one-sided 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.
```

```r
hypothesis(post_negbinom,hypothesis='secondfloor_bi1>0')
```

```
## Hypothesis Tests for class b:
##              Hypothesis Estimate Est.Error CI.Lower CI.Upper Evid.Ratio
## 1 (secondfloor_bi1) > 0    -0.19      0.01    -0.21    -0.16          0
##   Post.Prob Star
## 1         0
## ---
## '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.
```

Lastly, we check for the outliers in our selected model with the negative binomial distribution. We can find out that data entry 46 seems to be significantly off. More investigation needs to be done to find out why this is the case.

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



4 Drawbacks and improvement

Despite using the negative binomial distribution seems to be a reasonable approach to apply bayesian statistics to predict future housing sale price. Several things could be improved. First, more distribution can be used in our priors draw such as student t distribution. Also, we could potentially use Poisson distribution fit posterior predictive distribution with brm. Unfortunately, there are also some drawbacks to our analysis. We did not include some of the predictors available in the dataset according to our previous assumptions and expectations. Also, we did not measure the overfitting issue in our selected model, which could be done in the future.

5 Final Suggestions

Overall, we believe that our fitted distribution is relatively efficient in measuring house sale price. We could raise some suggestions for property construction and development as well as the general public looking for house purchase. As mentioned before, excluding larger living areas and lot area, small factors that could increase housing prices including constructed without a second floor, in the residential medium density area, built after 2008, with a central air system, and with the front street with gravel.