House selling price prediction

Anonymous

Contents

1. Introduction

Online services such as zillow zestimates [1] provide accuarate information on how much houses sell for using gathered data, providing useful information for the realtor and the person selling the house. This notebook explores the possibilities on using stan to build regression models to predict housing prices on an zipcode level and explores how the different variables effect the price.

For this purpose we conduct an small data exploration on an real world dataset. Fit two competing models an an linear and an nonlinear varying intercept model. Compare their performance on a test set and through model diagnostics. Finally we conclude with a discussion of the results.

[1] https://www.zillow.com/zestimate/

```
library(rstan)
options(mc.cores = 4)#parallel::detectCores())
library(loo)
library(bayesplot)
library(ggplot2)
library(matrixStats)
library(dplyr)
library(GGally)
library(corrplot)
library(reshape2)
library(ElemStatLearn)
library(glmnet)
library(plotmo)
library(Metrics)
source('stan_utility.R')
set.seed(42)
```

2. Dataset description

For the prediction task we have chosen House Sales in King County, USA dataset [2], which provides data for the houses sold between May 2014 / May 2015 in the area in an regression friendly form. We transform the price dependant variable on a log-scale and the independent variables to 0 mean and 1 variance (mean center and unit variance scale) to guarantee better numerical accuracy and faster convergence. We also remove variables sqft_above and sqft_basement from the dataset as they are colinear wth sqft_living.

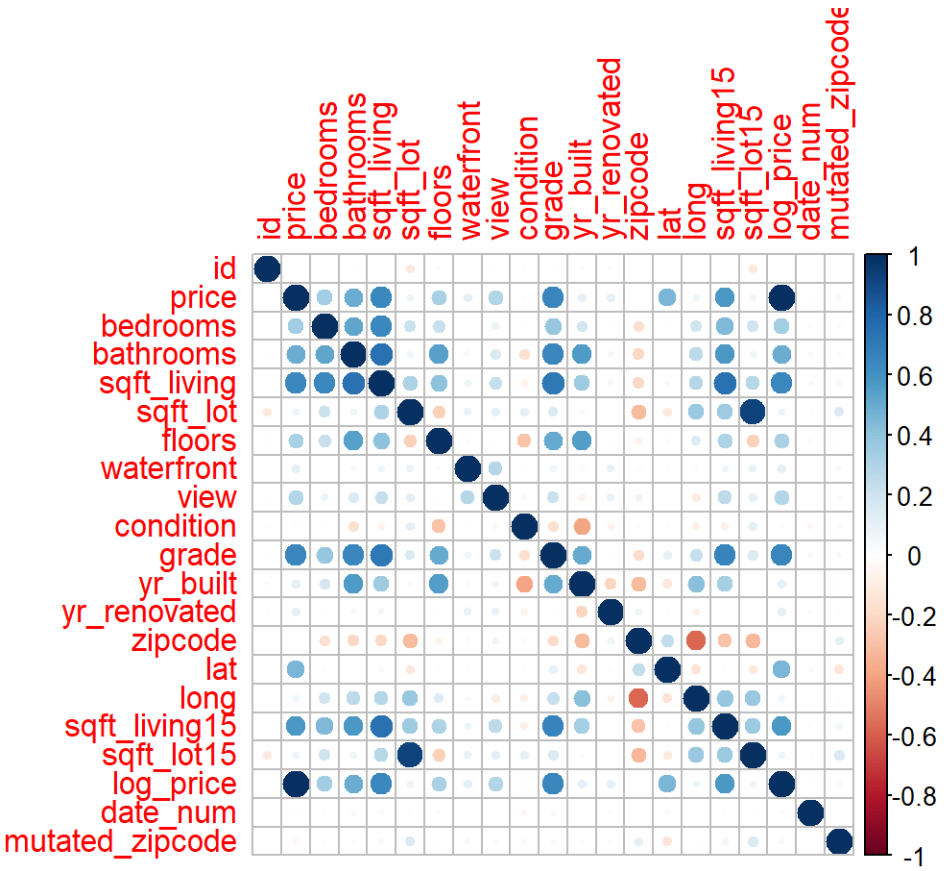
[2] https://www.kaggle.com/harlfoxem/housesalesprediction

```
houseprice = read.csv("data/kc_house_data.csv", header = TRUE)

#suffle rows to guarantee no row depencies
houseprice = houseprice[sample(nrow(houseprice)),]

#drop colinear columns sqft_living = sqft_above + sqft_basement
houseprice = subset(houseprice, select = -c(sqft_above, sqft_basement))

#transform the depended variable to log scale to  ensure better numerical accuracy
houseprice$log_price = log(houseprice$price)


datecol <- as.POSIXct(houseprice$date, format="%Y%m%dT%H%M%S")
houseprice$date_num = as.numeric(datecol)
unique_zips = unique(houseprice$zipcode)
houseprice$mutated_zipcode = match(houseprice$zipcode,  unique_zips)

head(houseprice)
```

```
##               id            date   price bedrooms bathrooms sqft_living
## 19772 9523100731 20140930T000000  580000        3      2.50        1620
## 20253 8563010130 20140725T000000 1300000        3      2.50        3350
## 6184  9284801435 20141203T000000  471000        4      1.75        1760
## 17946 6672920150 20150406T000000  330000        3      2.00        1500
## 13868 7524950210 20150401T000000  910000        4      2.50        2770
## 11217 5592900105 20150213T000000  435000        4      1.75        2520
##       sqft_lot floors waterfront view condition grade yr_built
## 19772     1171      3          0    4         3     8     2008
## 20253     7752      1          0    0         3     9     2009
## 6184      5750      1          0    2         5     7     1962
## 17946    11233      1          0    0         3     7     1987
## 13868     9798      2          0    0         4     9     1986
## 11217     7200      1          0    2         5     7     1955
##       yr_renovated zipcode     lat     long sqft_living15 sqft_lot15
## 19772            0   98103 47.6681 -122.355          1620       1505
## 20253            0   98008 47.6263 -122.099          2570       7988
## 6184             0   98126 47.5521 -122.373          1860       5750
## 17946            0   98019 47.7279 -121.967          1580      14013
## 13868            0   98027 47.5620 -122.081          3040      11100
## 11217            0   98056 47.4835 -122.192          2360       7300
##       log_price   date_num mutated_zipcode
## 19772  13.27078 1412024400               1
## 20253  14.07787 1406235600               2
## 6184   13.06261 1417557600               3
```

```
## 17946  12.70685 1428267600              4
## 13868  13.72120 1427835600              5
## 11217  12.98310 1423778400              6
```

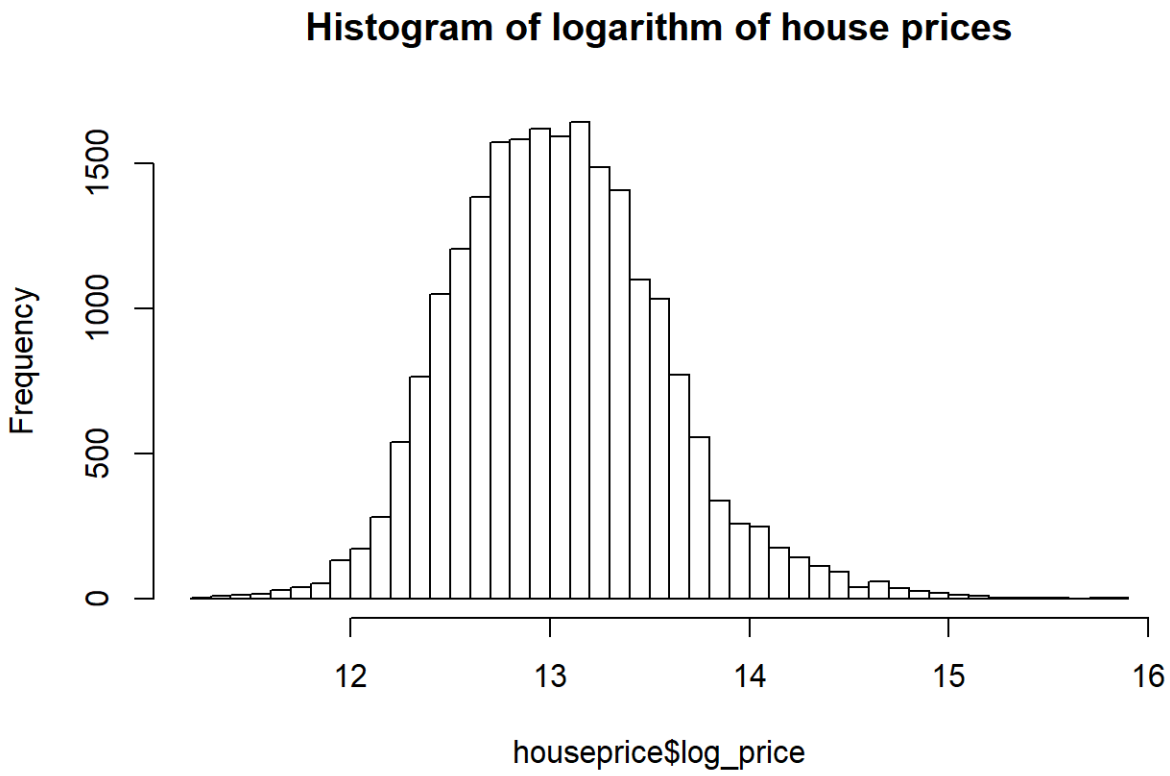
```r
M <- cor(houseprice[-2], method="spearman")
corrplot(M, method = "circle")
```



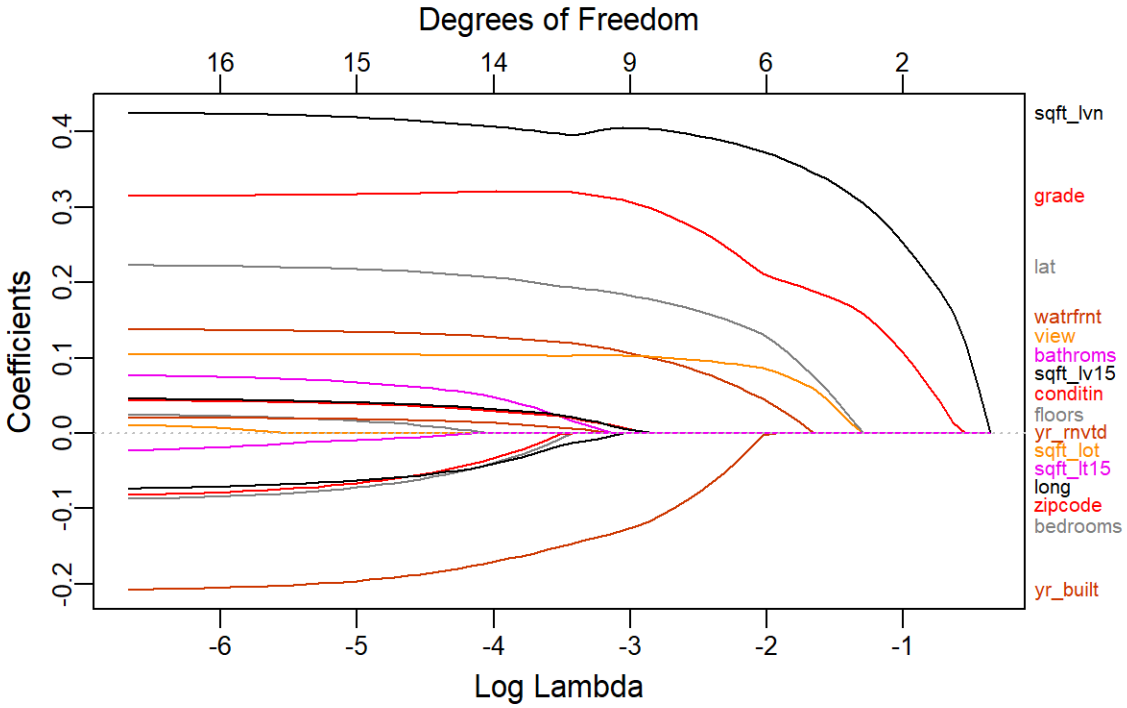
As the used dataset contains multiple predictors with linear and non-linear depencies, we use lasso regression to perform variable selection on the dataset to find an smaller subset of predictor variables to use in our model. This is neccessary for the purposes of the notebook to speed up the calculations and to better guarantee convergence. Also, the lasso regression plot beneath suggests that most of the variance in price can be explain by a much smaller subset of features.

!!! Notice that Lasso regression estimates are calculated using an linear model so they might not be the best predictors for an non-linear model.

```r
hist(houseprice$log_price, breaks=50, main="Histogram of logarithm of house prices")
```



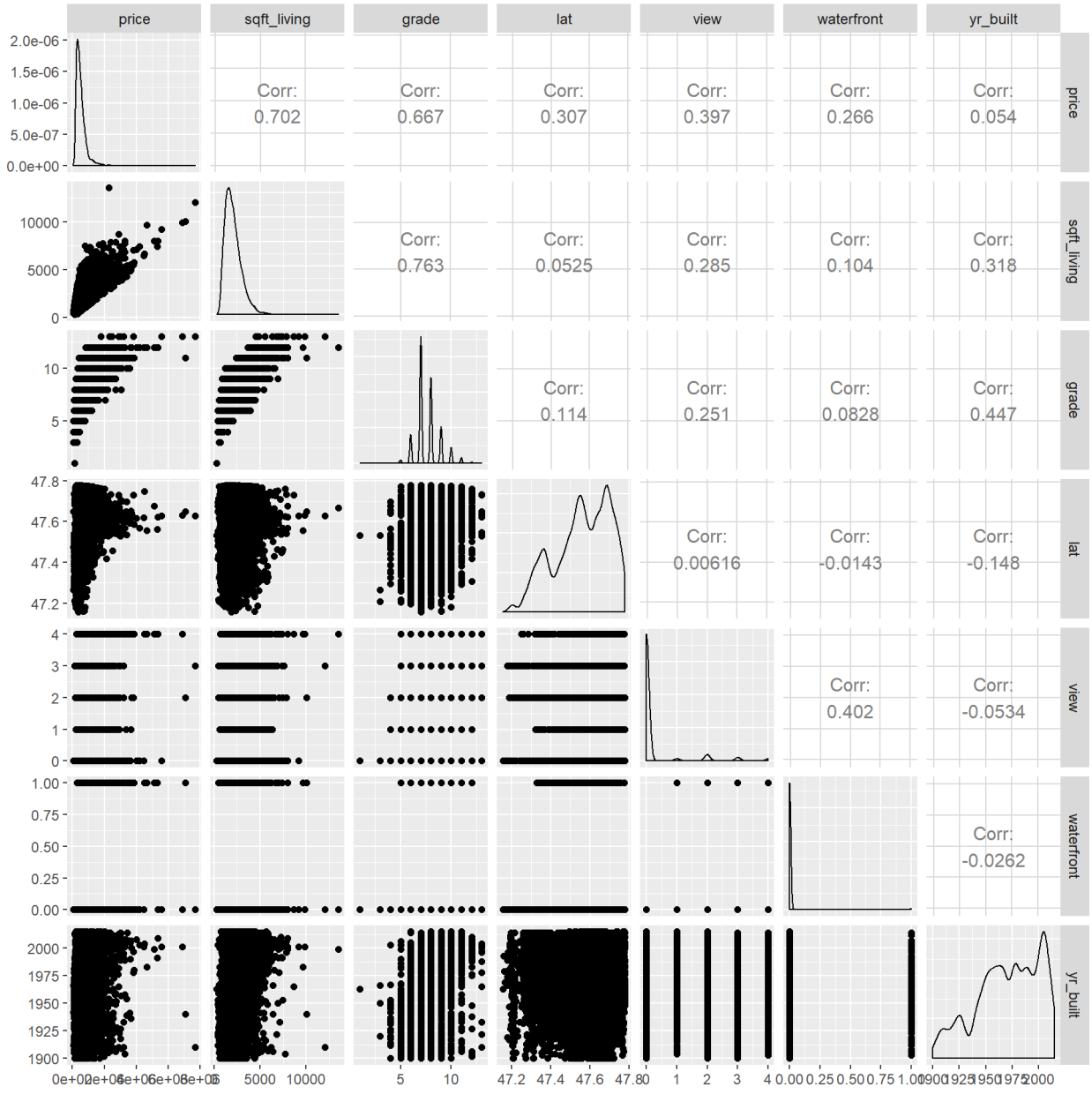

```
houseprice_scaled <- mutate_if(houseprice, is.numeric, list(~scale(.) %>% as.vector))

response = houseprice_scaled[3]
obs = houseprice_scaled[4:19]
ridge_regression <- glmnet(y=data.matrix(response), x=data.matrix(obs), alpha = 1)
plot_glmnet(ridge_regression, xvar = "lambda", label = TRUE)
```



From the lasso regression plot we can see that: sqft_living, grade, lat, view, waterfront and yr_built are the 6 best variables for the model. When they are plotted in the matrix plot underneath, we can see that they chosen variables exhibit various linear and nonlinear effects on price.

We choose 5 of these variables for the stan model except for the waterfront variable as its binary nature causes problems in convergence in the case of polynomial model used.

```
ggpairs(houseprice  %>% select(price, sqft_living, grade, lat, view,waterfront,yr_built))
```



3. Model description

We fit two varying intercept regression models: an multiple linear and an multiple polynomial model. Varying intercept models [3] are multilevel models, which let the intercept parameter vary between the groups while sharing the slope parameters. This is useful as the different zipcodes likely have different price levels that we can take into account in the intercept. We also could consider varying slope parameters by the category, but getting the model to convergece with this little data would likely be difficult.

[3] https://psmits.github.io/paleo_book/varying-intercept-models.html

3.1 Prior choices

In [4] it is recommended to scale the parameters to unit scale and to use student-t distribution $t_\nu(0, 1)$, where $3 < \nu < 7$, as a prior for linear regression coefficients. Student-t distribution has heavier tails than a normal distribution, but less heavy tails than a cauchy distribution, making it able to predict further away values while still keeping most of the mass near the mean.

$$t_{\nu_{pdf}} = \frac{\Gamma\frac{\nu+1}{2}}{\sqrt{\nu\pi}\Gamma\frac{\nu}{2}}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

[4] https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations#prior-for-linear-regression

3.2 Stan models

We have built the models using stan radon case study [6] as a starting point to build our regression models. We have expanded on the varying intercept model of the example by adding multiple linear and polynomial terms into the model. In our model intercept parameters vary by the zipcode while the slope parameters are shared across zipcodes.

[5] https://mc-stan.org/users/documentation/case-studies/radon.html

Grouped multiple linear

$$y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i$$

where $j = 1, \ldots, 70$ denotes the group of the observation. and $\epsilon_i \sim N(0, \sigma)$. The model can also be written as $y_i \sim N(\alpha_{j[i]} + \beta x_i, \sigma)$.

```
cat(readLines('models/grouped_multiple_linear.stan'), sep='\n')
```

```
## data {
##   int<lower=1> N;
##   int<lower=1> N_pred;
##   int<lower=1> N_groups;
##   int<lower=1> K;
##   vector[N] y;
##   matrix[N, K] X;
##   matrix[N_pred, K] X_pred;
##   int<lower=1> groups[N];
##   int<lower=1> groups_pred[N_pred];
## }
## parameters {
##   vector[N_groups] alpha;
##   vector[K] beta;
##   real<lower=0> sigma;
## }
## //transformed parameters {
## //  vector[N] mu;
## //  vector[N_pred] mu_pred;
## //  mu = alpha + X * beta;
## //  mu_pred = alpha + X_pred * beta;
## //}
```

```
## model {
##   real nu = 3;
##   alpha ~ student_t(nu,0,1);
##   beta ~ student_t(nu,0,1);
##   sigma ~ student_t(nu,0,1);
##   for (i in 1:N){
##     y[i] ~ normal(alpha[groups[i]] + X[i] * beta, sigma);
##   }
## }
## generated quantities {
##   vector[N_pred] y_pred;
##   vector[N] log_lik;
##   vector[N] y_rep ; // replicated data
##
##   for (i in 1:N_pred) {
##     y_pred[i] = normal_rng(alpha[groups_pred[i]] + X_pred[i] * beta, sigma);
##   }
##
##   for (i in 1:N) {
##     log_lik[i] = normal_lpdf(y[i] | alpha[groups[i]] + X[i] * beta, sigma);
##     y_rep[i] = normal_rng(alpha[groups[i]] + X[i] * beta, sigma);
##   }
## }
```

Grouped multiple polynomial

$$y_i = \alpha_{j[i]} + \beta x_i + \gamma x_i^2 + \epsilon_i$$

```
cat(readLines('models/grouped_multiple_polynomial.stan'), sep='\n')
```

```
## data {
##   int<lower=1> N;
##   int<lower=1> N_pred;
##   int<lower=1> N_groups;
##   int<lower=1> K;
##   vector[N] y;
##   matrix[N, K] X;
##   matrix[N, K] X_second;
##   matrix[N_pred, K] X_pred;
##   matrix[N_pred, K] X_pred_second;
##   int<lower=1> groups[N];
##   int<lower=1> groups_pred[N_pred];
## }
## parameters {
##   vector[N_groups] alpha;
##   vector[K] beta;
##   vector[K] beta_second;
##   real<lower=0> sigma;
## }
## //transformed parameters {
## //  vector[N] mu;
## //  vector[N_pred] mu_pred;
## //  mu = alpha + X * beta;
```

```
## //  mu_pred = alpha + X_pred * beta;
## //}
## model {
##   real nu = 3;
##   alpha ~ student_t(nu,0,1);
##   beta ~ student_t(nu,0,1);
##   beta_second ~ student_t(nu,0,1);
##   sigma ~ student_t(nu,0,1);
##   for (i in 1:N){
##     y[i] ~ normal(alpha[groups[i]] + X[i] * beta + X_second[i] * beta_second, sigma);
##   }
## }
## generated quantities {
##   vector[N_pred] y_pred;
##   vector[N] log_lik;
##   vector[N] y_rep;
##
##   for (i in 1:N_pred) {
##     y_pred[i] = normal_rng(alpha[groups_pred[i]] + X_pred[i] * beta+ X_pred_second[i] * beta_second,
##   }
##
##   for (i in 1:N) {
##     log_lik[i] = normal_lpdf(y[i] | alpha[groups[i]] + X[i] * beta + X_second[i] * beta_second, sigma
##     y_rep[i] = normal_rng( alpha[groups[i]] + X[i] * beta + X_second[i] * beta_second, sigma);
##   }
## }
##
```

3.3 Running the models

We train the models on 80%/20%-test split using 10000 first datapoints. Models are fitted for all the 71 zipcodes. Full data is not used as loo fails with large datasets.

```
usable_numeric_columns = c("sqft_living", "grade","view", "lat", "yr_built")
```

```
training_indices = 0:8000
testing_indices = 8001:10000

used_columns = usable_numeric_columns
target_column = c("log_price")
group_column =  c("mutated_zipcode")
original_target = houseprice[,target_column]
training_data = houseprice_scaled[training_indices,used_columns]
testing_data =  houseprice_scaled[testing_indices, used_columns]
training_target = houseprice_scaled[training_indices,target_column]
testing_target_scaled =  houseprice_scaled[testing_indices, target_column]
testing_target =  houseprice[testing_indices, target_column]

X_var = training_data
X_var_pred = testing_data
y_var = training_target
group_var = houseprice[training_indices,group_column]
group_var_pred = houseprice[testing_indices,group_column]
```

```
data_list = list(
  X = X_var,
  X_pred = X_var_pred,
  K = ncol(X_var),
  N = nrow(X_var),
  N_pred = nrow(X_var_pred),
  N_groups = length(unique_zips),
  y = y_var,
  groups = group_var,
  groups_pred  = group_var_pred
)
head(X_var)
```

```
##   sqft_living      grade       view         lat   yr_built
## 1  -0.5007396  0.2919089  4.9140157  0.77976752  1.2594678
## 2   1.3828873  1.1426405 -0.3057524  0.47810123  1.2935122
## 3  -0.3483074 -0.5588228  2.3041317 -0.05739251 -0.3065744
## 4  -0.6313958 -0.5588228 -0.3057524  1.21133795  0.5445355
## 5   0.7513823  1.1426405 -0.3057524  0.01405477  0.5104911
## 6   0.4791819 -0.5588228  2.3041317 -0.55247163 -0.5448852
```

```
denormalize_results <- function(new_values, sd, mean){
  return (new_values * sd + mean)
}
orig_sd = sd(original_target)
orig_mean = mean(original_target)
```

Grouped multiple linear

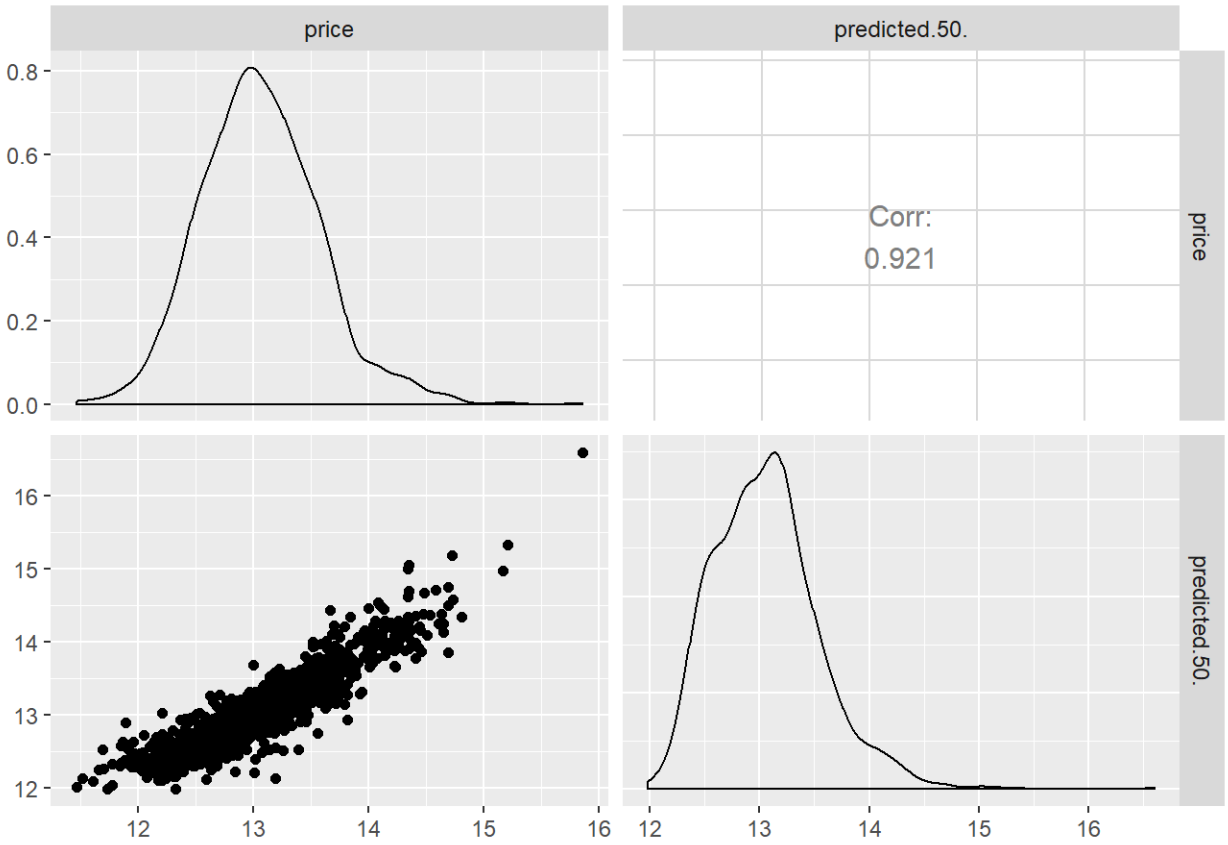
```
multiple_linear_fit <- stan(file = 'models/grouped_multiple_linear.stan', data = data_list)
```

```
## Warning: Bulk Effective Samples Size (ESS) is too low, indicating posterior means and medians may be
## Running the chains for more iterations may help. See
## http://mc-stan.org/misc/warnings.html#bulk-ess
```

```
## Warning: Tail Effective Samples Size (ESS) is too low, indicating posterior variances and tail quanti
## Running the chains for more iterations may help. See
## http://mc-stan.org/misc/warnings.html#tail-ess
```

```
predicted_draws = extract(multiple_linear_fit)$y_pred
predicted_raws = colQuantiles(predicted_draws, probs = c(0.05, 0.5, 0.95))
predicted_prices = denormalize_results(predicted_raws, orig_sd, orig_mean)
rep_prices = extract(multiple_linear_fit)$y_rep %>% colQuantiles(probs = c(0.05, 0.5, 0.95)) %>% denorma
```

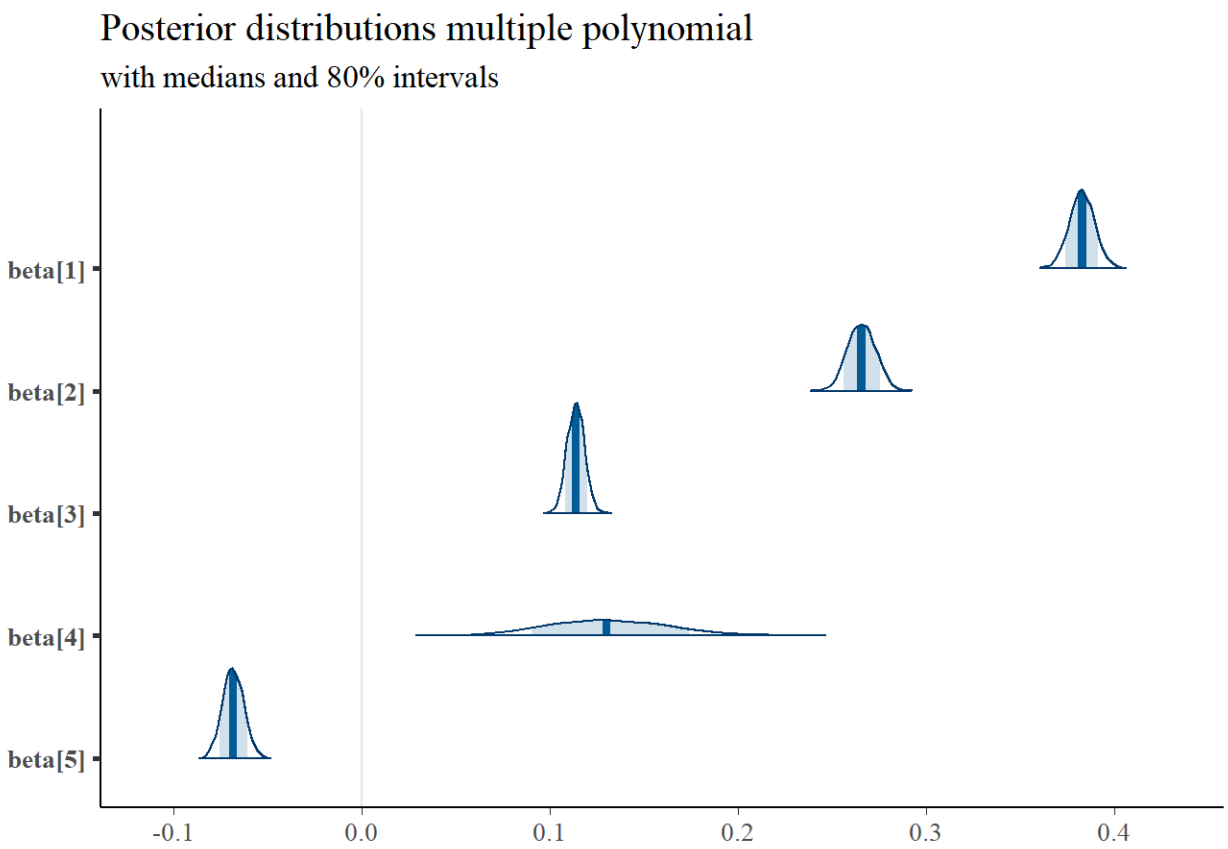
```
result_testing = data.frame(price = (testing_target), predicted = (predicted_prices))
ggpairs(result_testing, columns = c("price", "predicted.50."))
```




```
#result_rep = data.frame(price = exp(original_target[training_indices]), predicted = exp(rep_prices))
#ggpairs(result_rep, columns = c("price", "predicted.50."))
#mae(exp(original_target[training_indices]),exp(rep_prices))
#rep_order = order(original_target[training_indices])
#rep_error = abs(exp(original_target[training_indices]) - exp(rep_prices))
#plot(original_target[training_indices][testing_rising_order], rep_error[testing_rising_order], col="#0
```

We can see that the best parameters correspond to lasso regression best parameters. Fourth parameter corresponding to view quality has a very wide posterior distribution.

```
posterior_linear <- as.matrix(multiple_linear_fit)
posterior_linear <- posterior_linear[,c("beta[1]", "beta[2]","beta[3]","beta[4]","beta[5]")]


plot_title <- ggtitle("Posterior distributions multiple polynomial",
                      "with medians and 80% intervals")

mcmc_areas(posterior_linear,
           pars = c("beta[1]", "beta[2]","beta[3]","beta[4]","beta[5]"),
           prob = 0.8) + plot_title
```



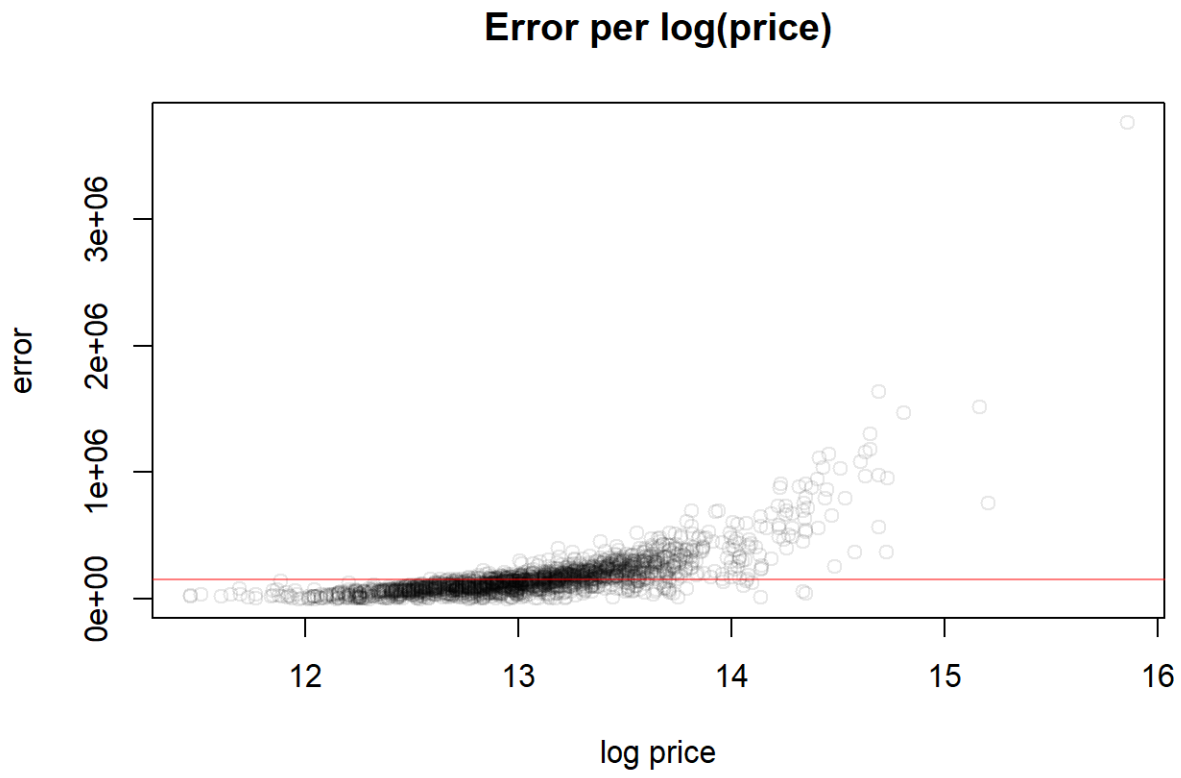
Mean absolute error is quite large

```
mae_lin = mae(exp(testing_target),exp(predicted_prices))
mae_lin
```

```
## [1] 155366.6
```

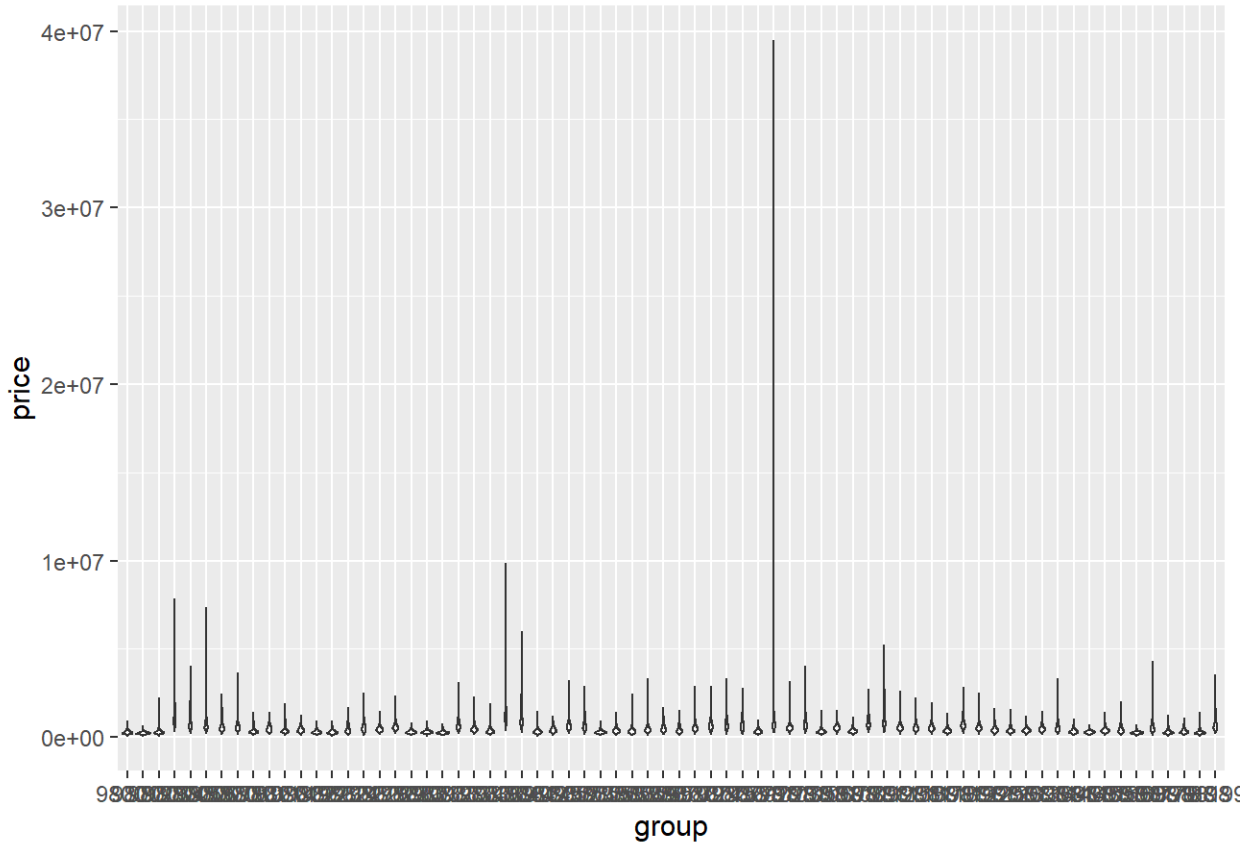
The error increases as a function of price

```
testing_rising_order = order(testing_target)
error = abs(exp(testing_target) - exp(predicted_prices))
plot(testing_target[testing_rising_order], error[testing_rising_order], col="#00000018", ylab = "error"
lines(c(0, 20000), c(150000, 150000), col = "#ff000080")
```




```
#hist(error, breaks = 500, xlim = c(0, 2e6))
#lines(c(150000, 150000), c(0, 20000), col = "red")
```

```
#TODO: Redo the zipcode plot
violin_predicted = extract(multiple_linear_fit)$y_pred
violin_predicted = exp(denormalize_results(violin_predicted, orig_sd, orig_mean))
#violin_predicted = exp(predicted_prices)
violin_groups = outer(1:nrow(violin_predicted), 1:ncol(violin_predicted),
                      FUN=function(r,c) unique_zips[group_var_pred[c]] )
violin_predicted = c(t(violin_predicted))
violin_groups = as.factor(c(t(violin_groups)))
violin_data_list_thing = data.frame(price=violin_predicted, group=violin_groups)

p <- ggplot(violin_data_list_thing, aes(x=group, y=price)) +
  geom_violin()
p
```



Grouped multiple polynomial

```
X_var_second = X_var^2
X_var_pred_second = X_var_pred^2

#not used (third degree polynomial model data)
X_var_third = X_var^3
X_var_pred_third = X_var_pred^3

data_list = list(
  X = X_var,
  X_second = X_var_second,
  X_third = X_var_third,
  X_pred = X_var_pred,
  X_pred_second = X_var_pred_second,
  X_pred_third = X_var_pred_third,
  K = ncol(X_var),
  N = nrow(X_var),
  N_pred = nrow(X_var_pred),
  N_groups = length(unique_zips),
  y = y_var,
  groups = group_var,
  groups_pred  = group_var_pred
)
```

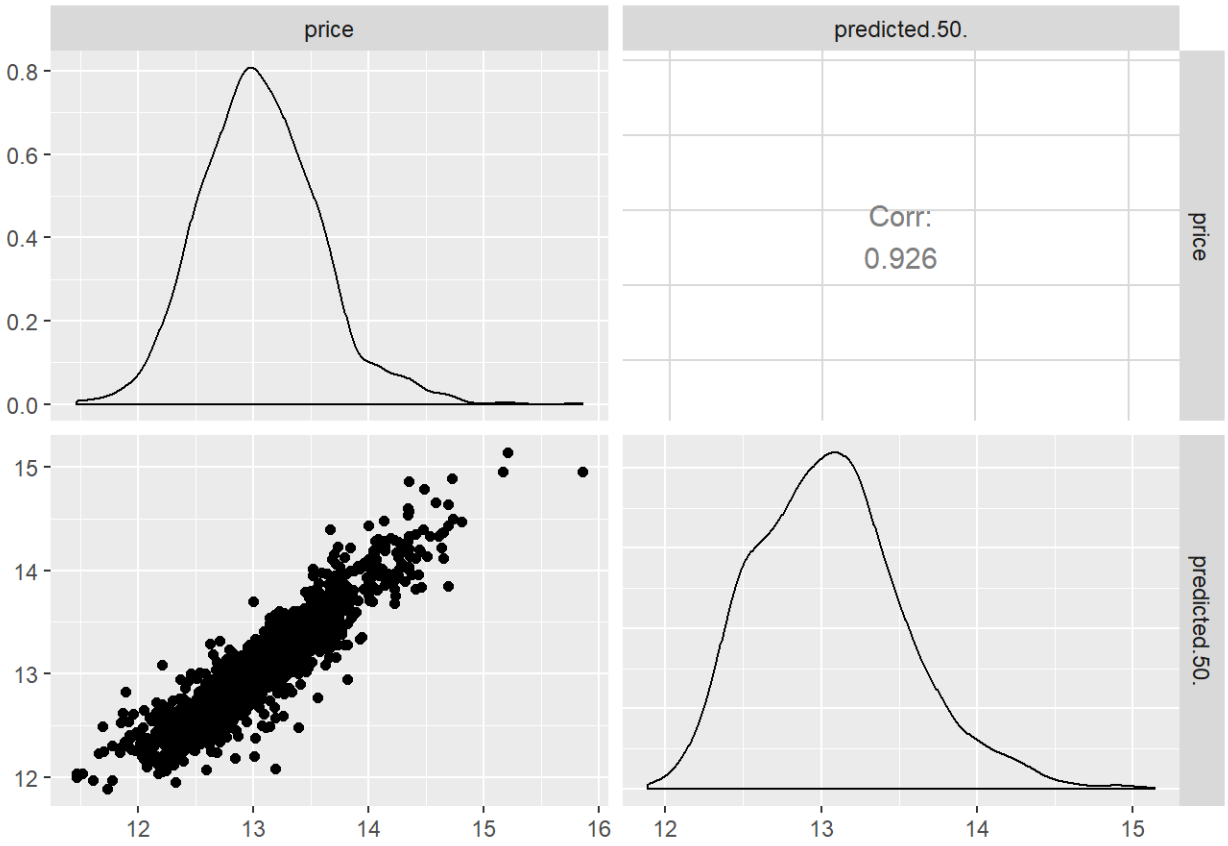
```
multiple_polynomial_fit <- stan(file = 'models/grouped_multiple_polynomial.stan',
                                data = data_list)
```

```
## Warning: Bulk Effective Samples Size (ESS) is too low, indicating posterior means and medians may be
## Running the chains for more iterations may help. See
## http://mc-stan.org/misc/warnings.html#bulk-ess
```

```
## Warning: Tail Effective Samples Size (ESS) is too low, indicating posterior variances and tail quanti
## Running the chains for more iterations may help. See
## http://mc-stan.org/misc/warnings.html#tail-ess
```

```
predicted_draws = extract(multiple_polynomial_fit)$y_pred
predicted_raws = colQuantiles(predicted_draws, probs = c(0.05, 0.5, 0.95))
predicted_prices = denormalize_results(predicted_raws, orig_sd, orig_mean)
```

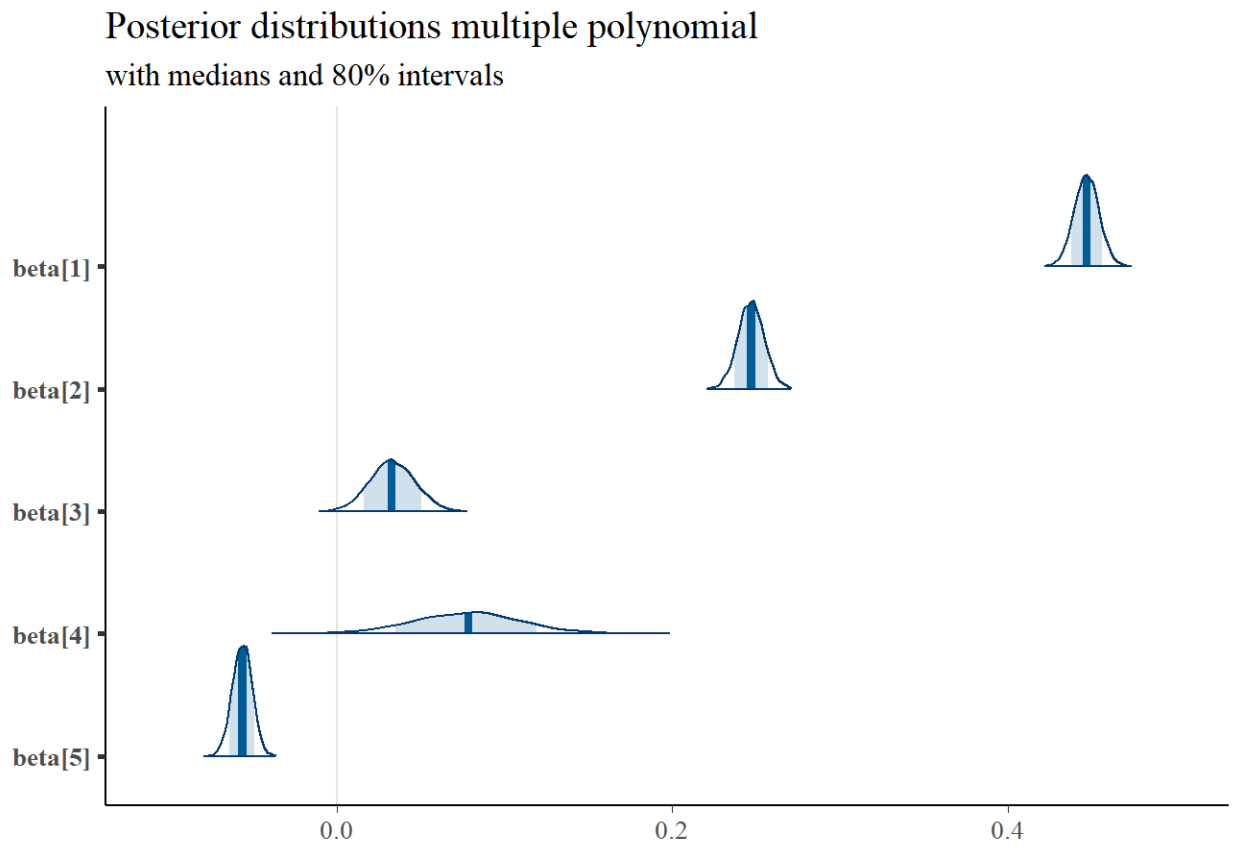
```
result_testing = data.frame(price = testing_target, predicted = predicted_prices)
ggpairs(result_testing, columns = c("price", "predicted.50."))
```



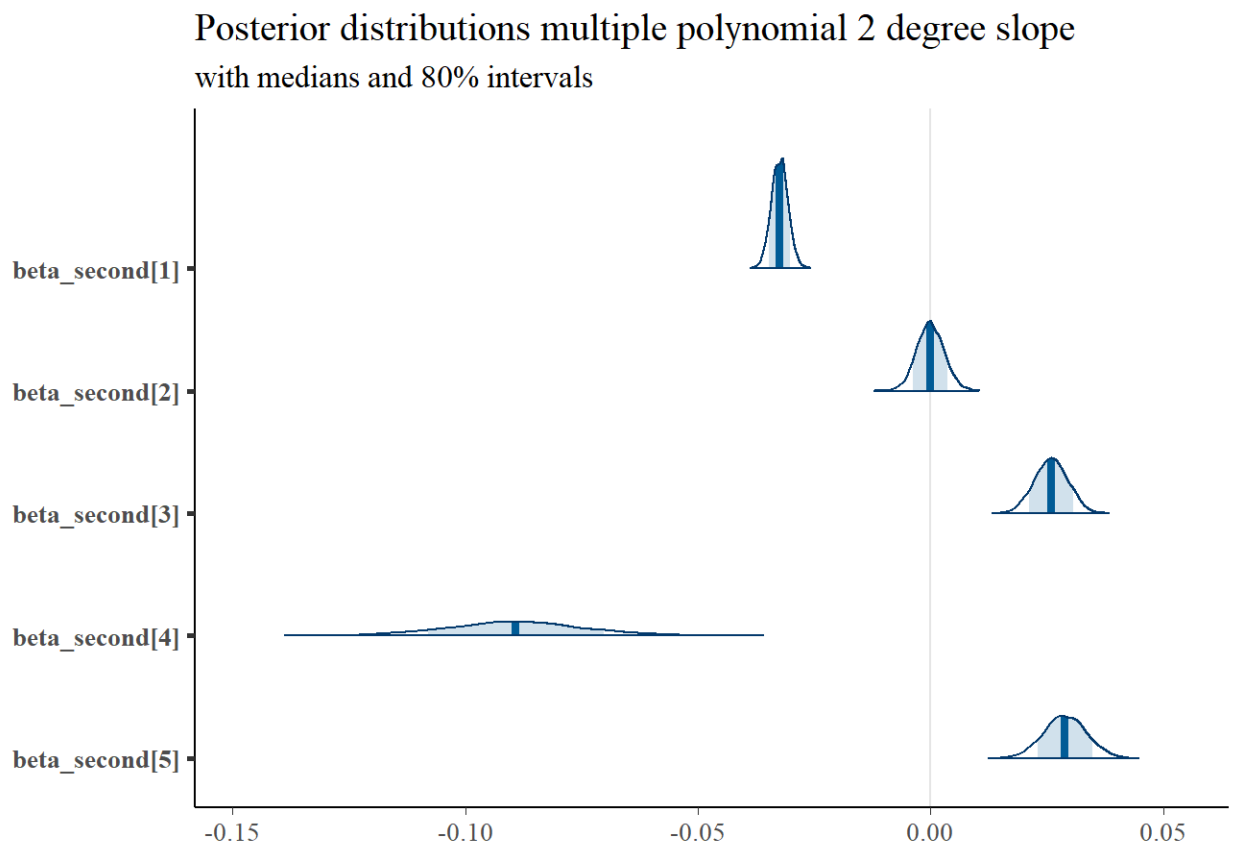
From the first order polynomial part the parameters, which had the most effect correspond to lasso regression best parameters.

```
posterior_polynomial <- as.matrix(multiple_polynomial_fit)
posterior_polynomial <- posterior_polynomial[,c("beta[1]", "beta[2]","beta[3]","beta[4]","beta[5]",
                                "beta_second[1]", "beta_second[2]","beta_second[3]","beta_second[4]",
```

```
plot_title <- ggtitle("Posterior distributions multiple polynomial",
                      "with medians and 80% intervals")

mcmc_areas(posterior_polynomial,
           pars = c("beta[1]", "beta[2]","beta[3]","beta[4]","beta[5]"),
           prob = 0.8) + plot_title
```




```
plot_title <- ggtitle("Posterior distributions multiple polynomial 2 degree slope",
                      "with medians and 80% intervals")

mcmc_areas(posterior_polynomial,
           pars = c("beta_second[1]", "beta_second[2]","beta_second[3]","beta_second[4]","beta_second[5]
           prob = 0.8) + plot_title
```

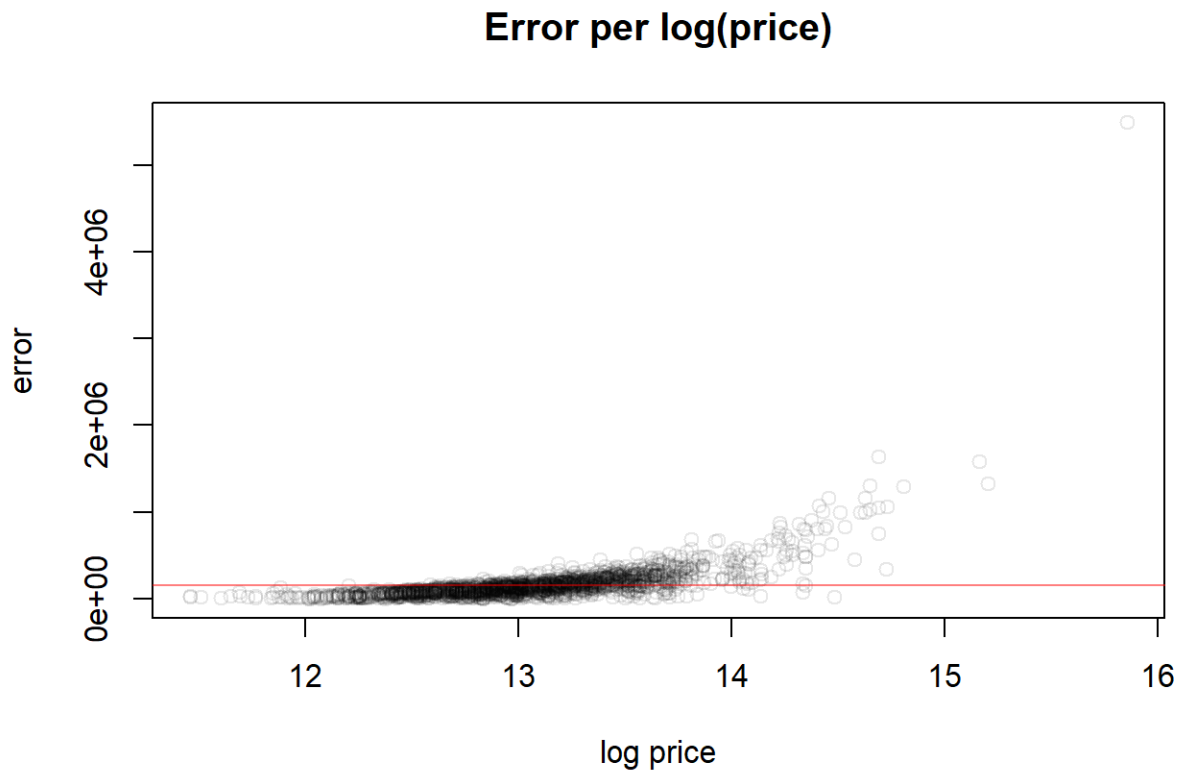


Mean absolute error still is quite large, but there is a small improved compared to linear model.

```
mae_pol = mae(exp(testing_target),exp(predicted_prices))
mae_pol
```

```
## [1] 147179
```

```
testing_rising_order = order(testing_target)
error = abs(exp(testing_target) - exp(predicted_prices))
plot(testing_target[testing_rising_order], error[testing_rising_order], col="#00000018", ylab = "error"
lines(c(0, 20000), c(150000, 150000), col = "#ff000080")
```




```
#hist(error, breaks = 500, xlim = c(0, 2e6))
#lines(c(150000, 150000), c(0, 20000), col = "red")
```

4. Convergence diagnostics

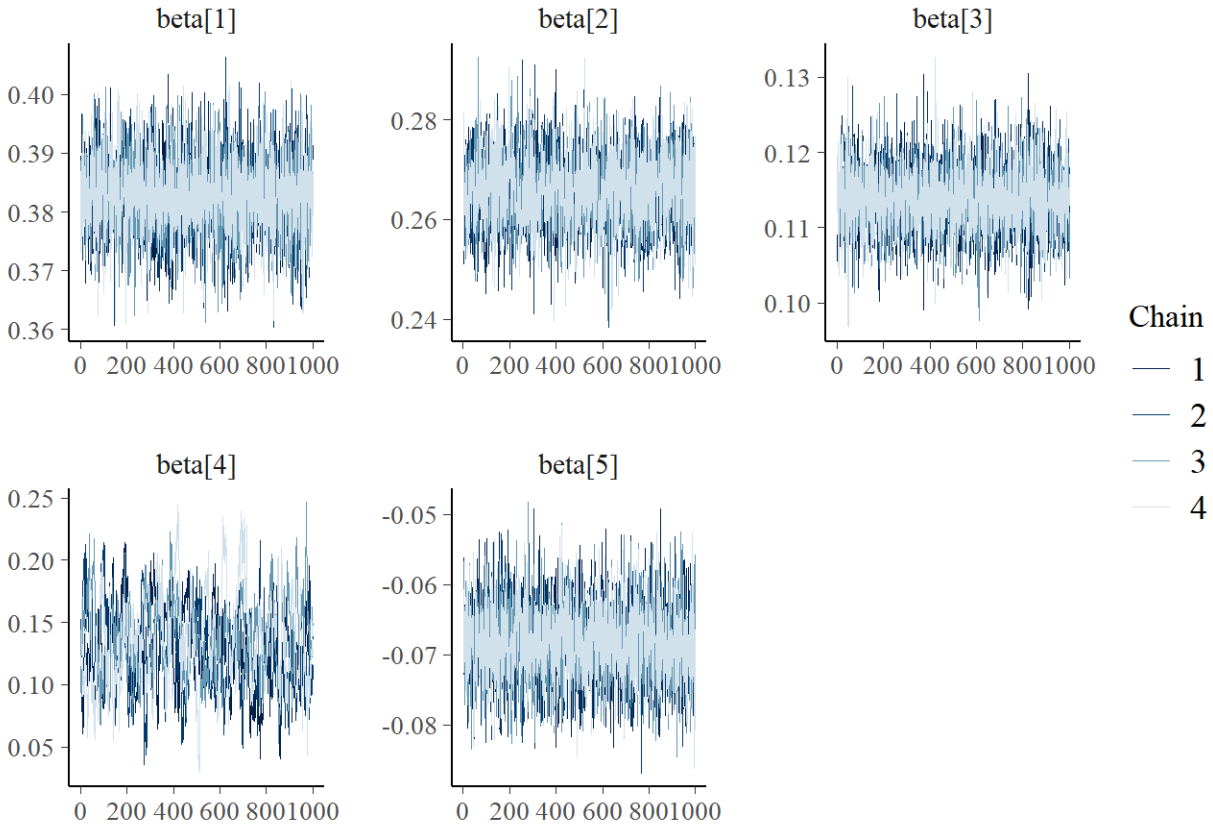
Grouped multiple linear

From the model diagnostics we can see that that the model has converged.

```
check_all_diagnostics(multiple_linear_fit)
```

```
## [1] "n_eff / iter looks reasonable for all parameters"
## [1] "Rhat looks reasonable for all parameters"
## [1] "0 of 4000 iterations ended with a divergence (0%)"
## [1] "0 of 4000 iterations saturated the maximum tree depth of 10 (0%)"
## [1] "E-BFMI indicated no pathological behavior"
```

```
posterior_divergences <- as.array(multiple_linear_fit)
mcmc_trace(multiple_linear_fit, regex_pars = "beta")
```



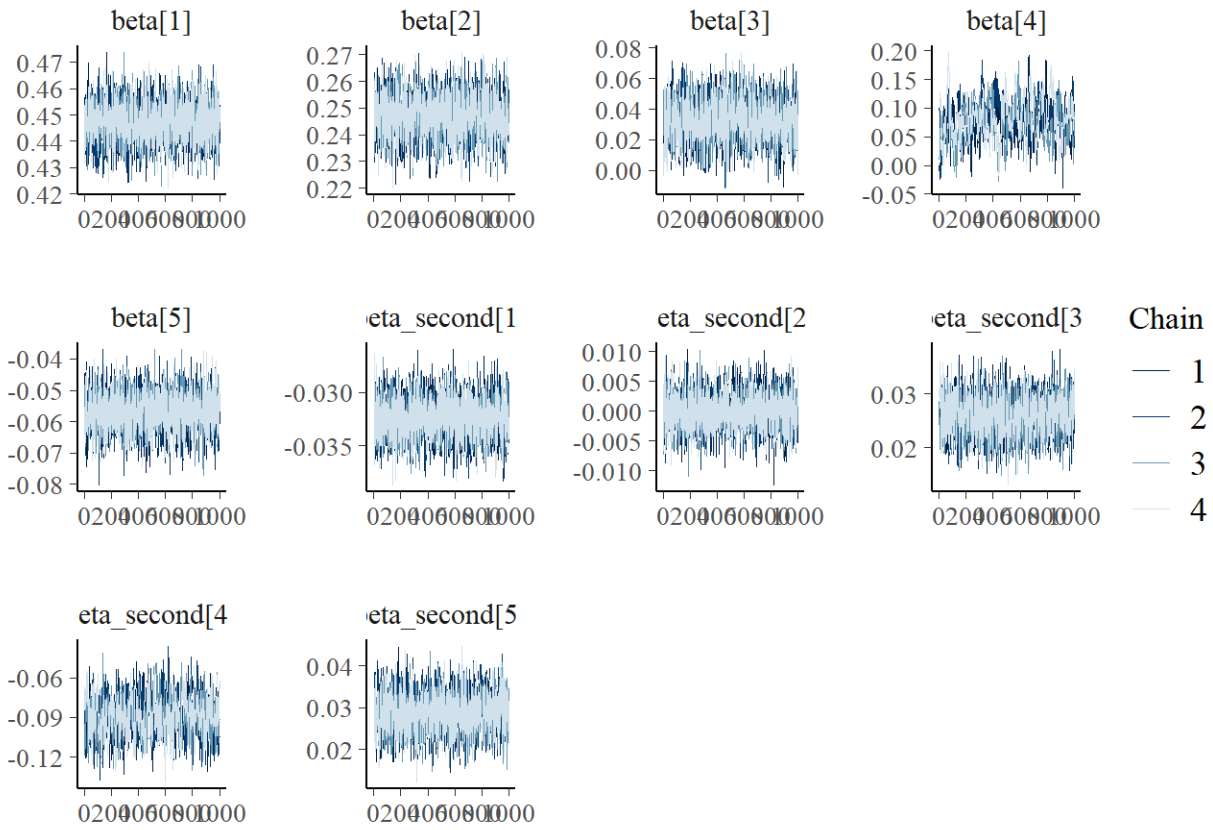
Grouped multiple polynomial

From the model diagnostics we can see that that the model has converged.

```
#print(multiple_polynomial_fit,  pars = c("alpha", "beta", "beta_second"))
check_all_diagnostics(multiple_linear_fit)
```

```
## [1] "n_eff / iter looks reasonable for all parameters"
## [1] "Rhat looks reasonable for all parameters"
## [1] "0 of 4000 iterations ended with a divergence (0%)"
## [1] "0 of 4000 iterations saturated the maximum tree depth of 10 (0%)"
## [1] "E-BFMI indicated no pathological behavior"
```

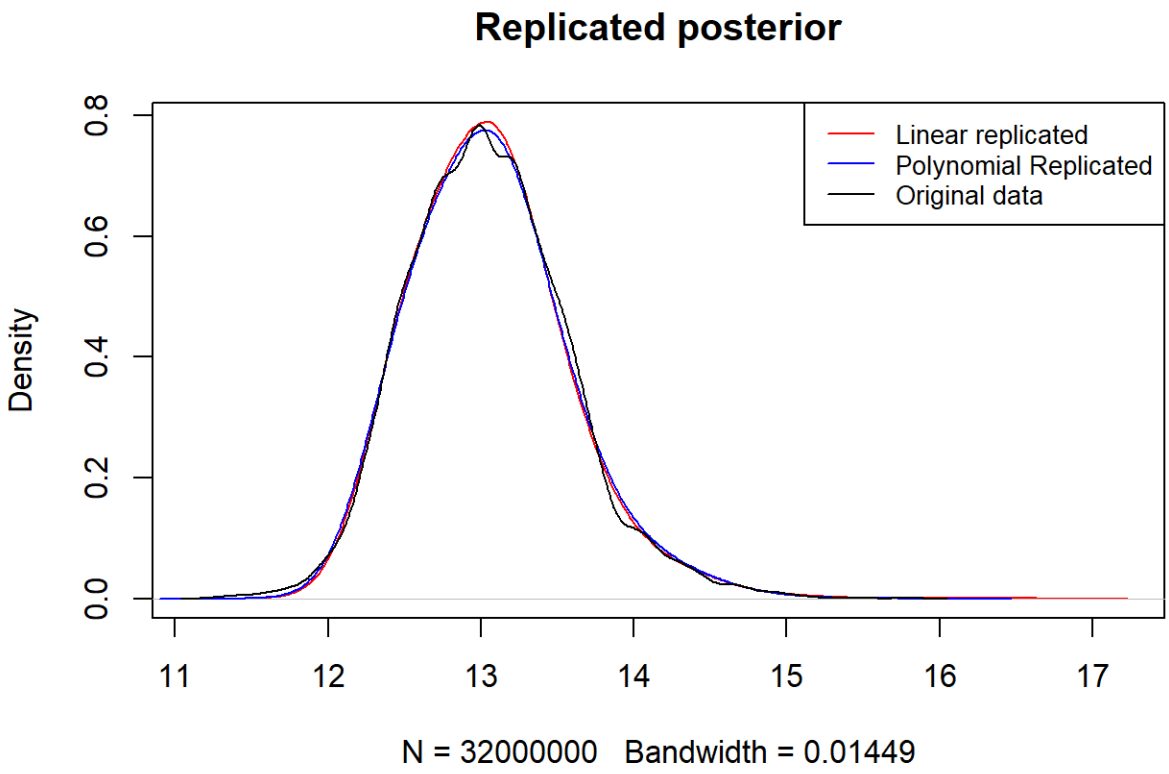
```
mcmc_trace(multiple_polynomial_fit, regex_pars = "beta")
```



5. Posterior predictive checking

From the plot beneath, we can see that replicated data is nearly indistinguishable from the target

```
replicated_data_lin = denormalize_results(extract(multiple_linear_fit)$y_rep, orig_sd, orig_mean)
replicated_data_pol = denormalize_results(extract(multiple_polynomial_fit)$y_rep, orig_sd, orig_mean)

plot(density(replicated_data_lin), col="red", main="Replicated posterior" )
lines(density(replicated_data_pol), col="blue")
lines(density(original_target), col="black")
legend(x="topright",
       legend=c("Linear replicated", "Polynomial Replicated", "Original data"),
       col=c("red", "blue", "black"), lty=1:1, cex=0.8)
```



Multiple linear model Leave-One-Out (LOO) predictive checks

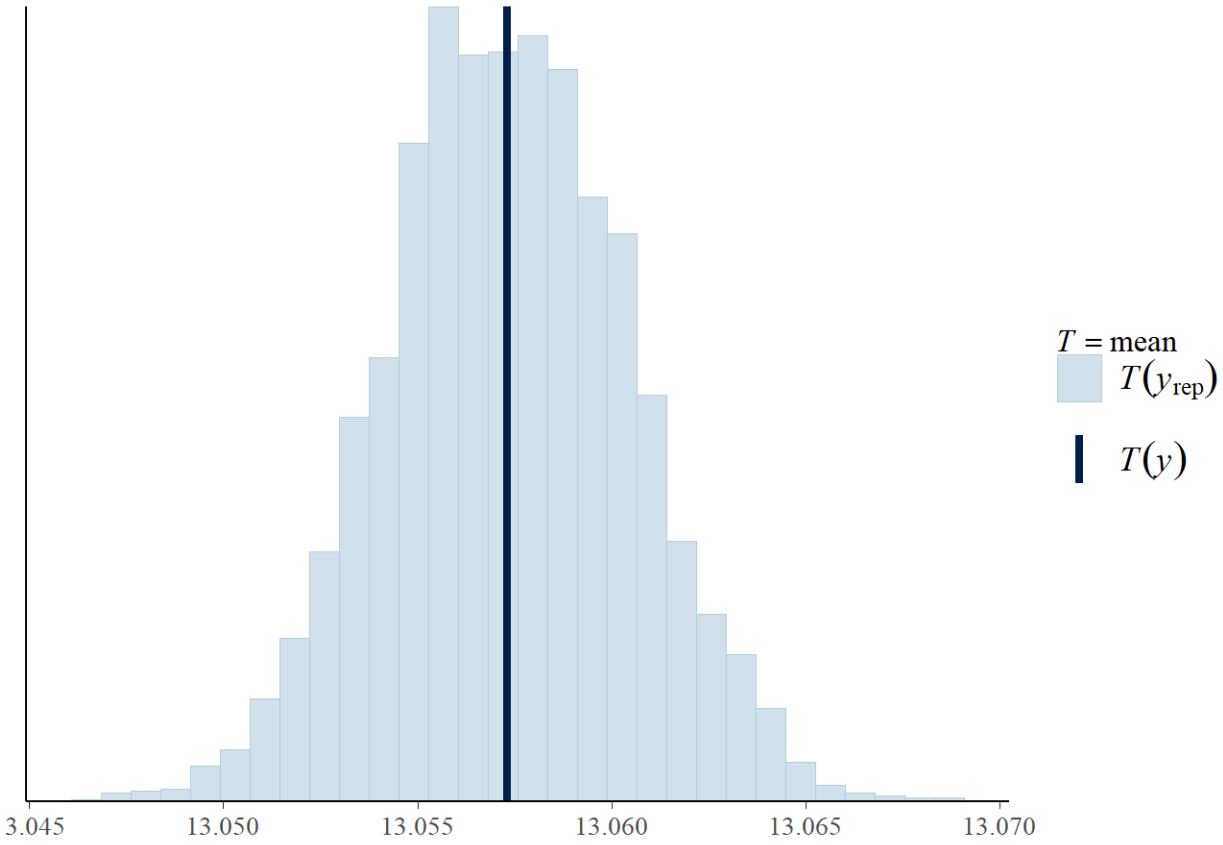
Loo pit plot suggests that our model is not able to predict all the variance in the data.

```
original_training_order = order(original_target[training_indices])
loo_lin <- loo(multiple_linear_fit, save_psis = TRUE,cores = getOption("mc.cores", 4))
```

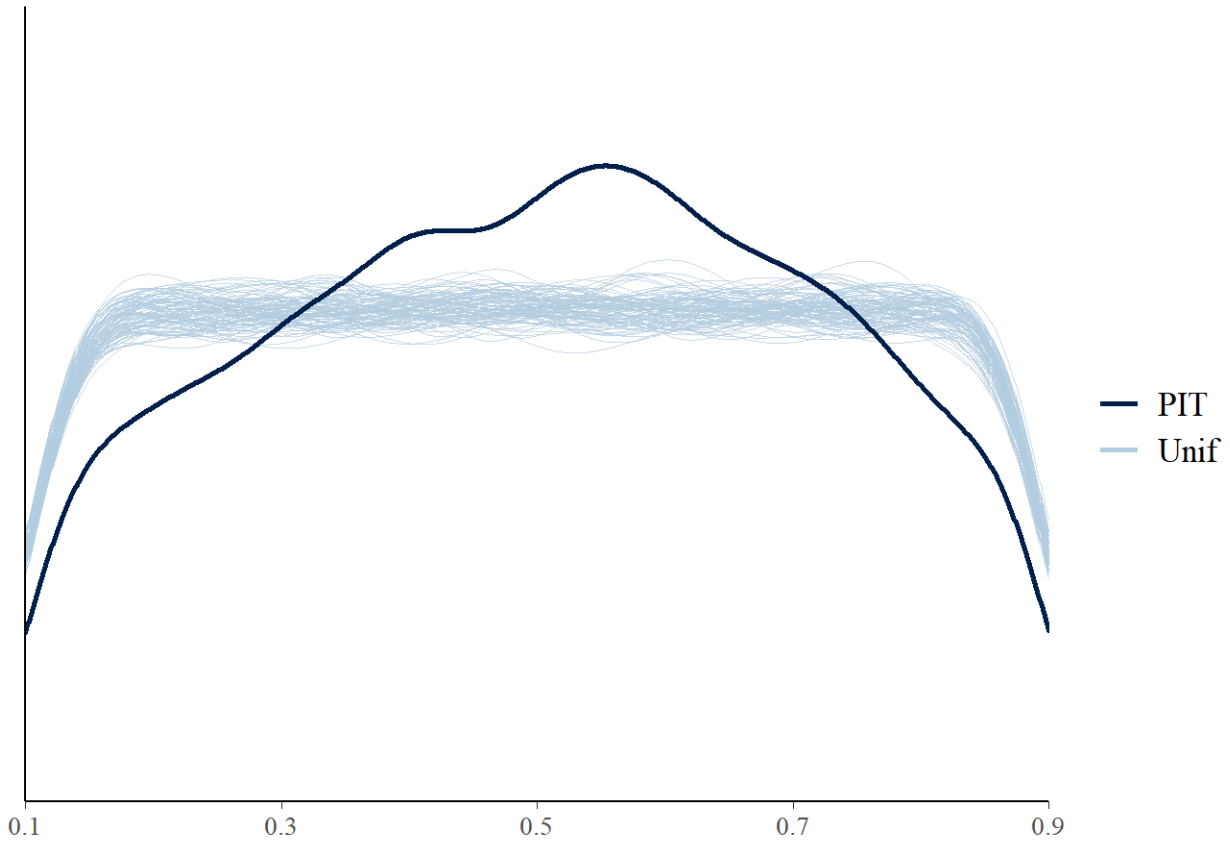
```
## Warning: Some Pareto k diagnostic values are slightly high. See help('pareto-k-diagnostic') for deta
```

```
psis_lin <- loo_lin$psis_object
lw_lin <- weights(psis_lin)
pp_check(c(original_target[training_indices]), yrep = replicated_data_lin, fun = "stat")
```

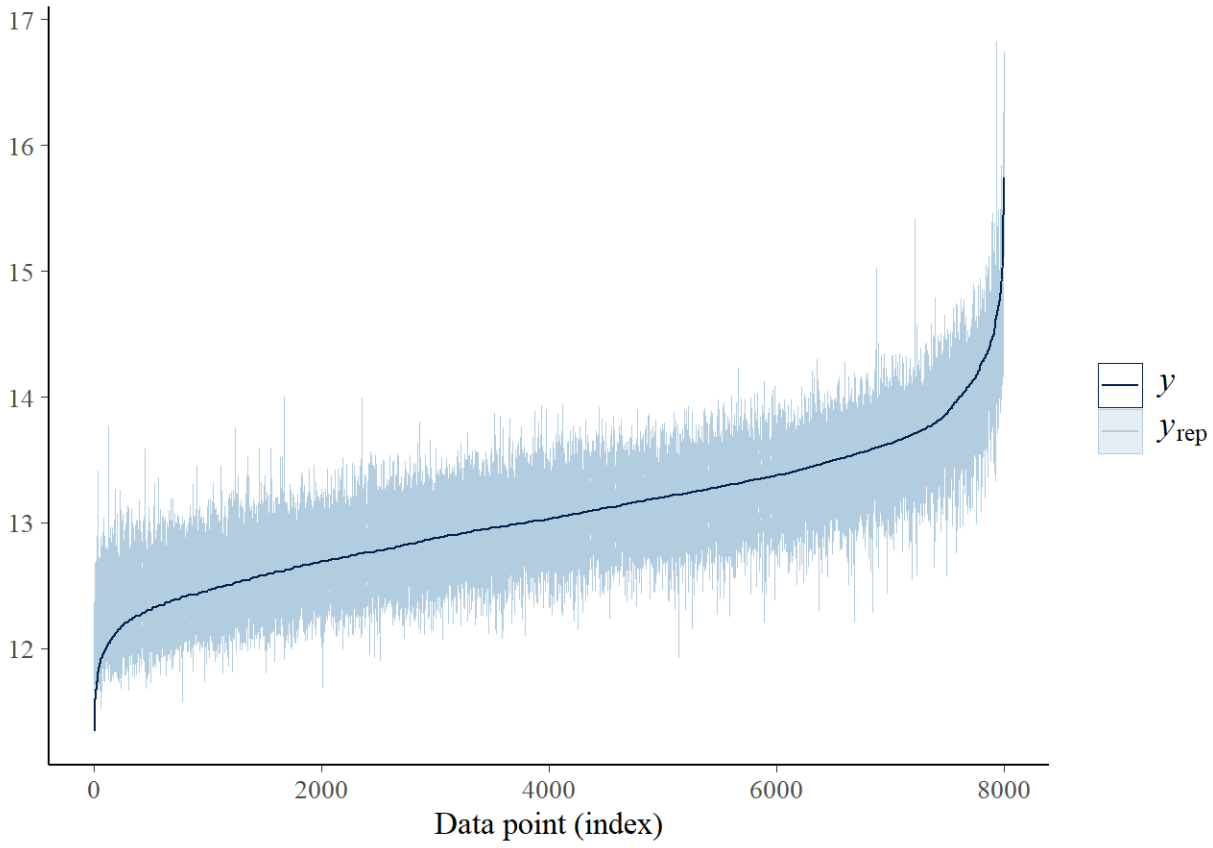
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



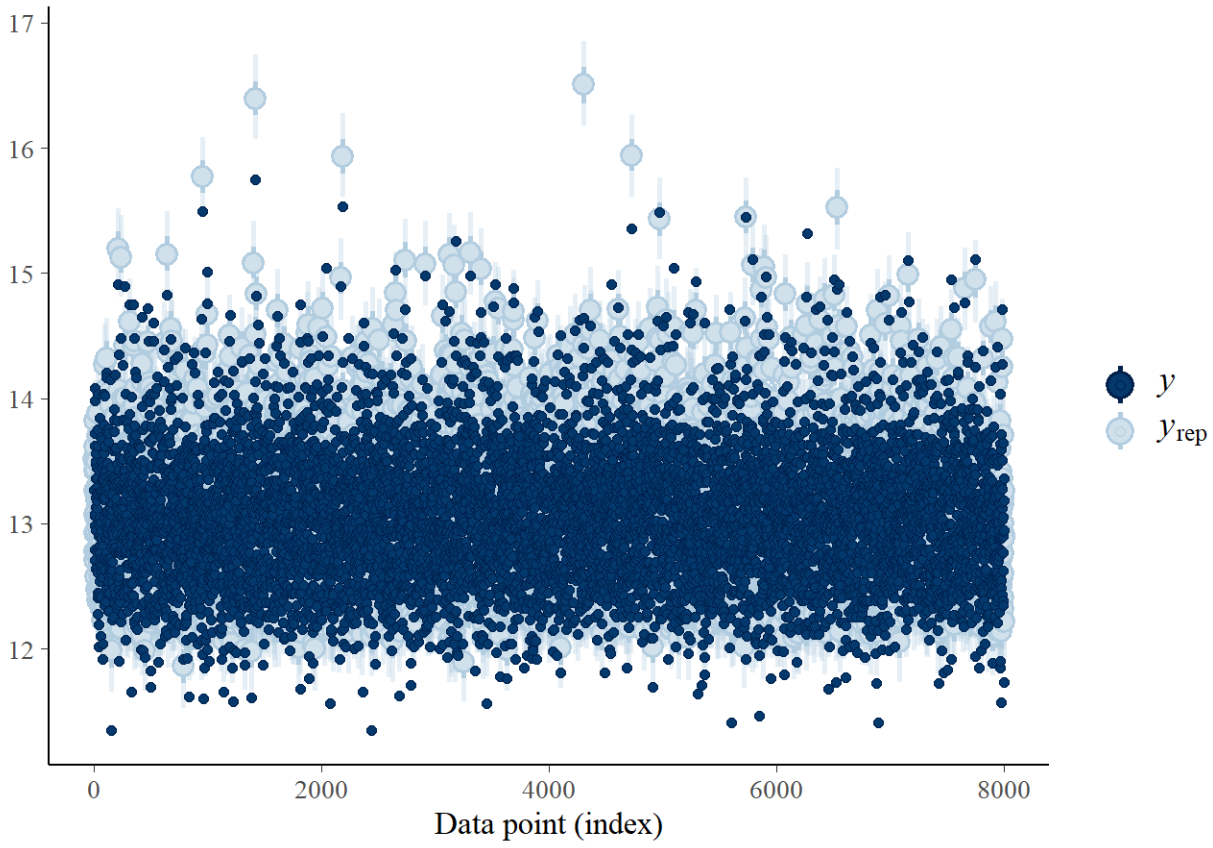

```
ppc_loo_pit_overlay(c(original_target[training_indices]), yrep = replicated_data_lin,
                    lw = lw_lin)
```




```
ppc_loo_ribbon(c(original_target[training_indices][original_training_order]),
               yrep = replicated_data_lin[,original_training_order],
               lw = lw_lin, psis_object = psis_lin)
```




```
ppc_loo_intervals(c(original_target[training_indices]),
                  yrep = replicated_data_lin, psis_object = psis_lin)
```



Multiple polynomial model Leave-One-Out (LOO) predictive checks

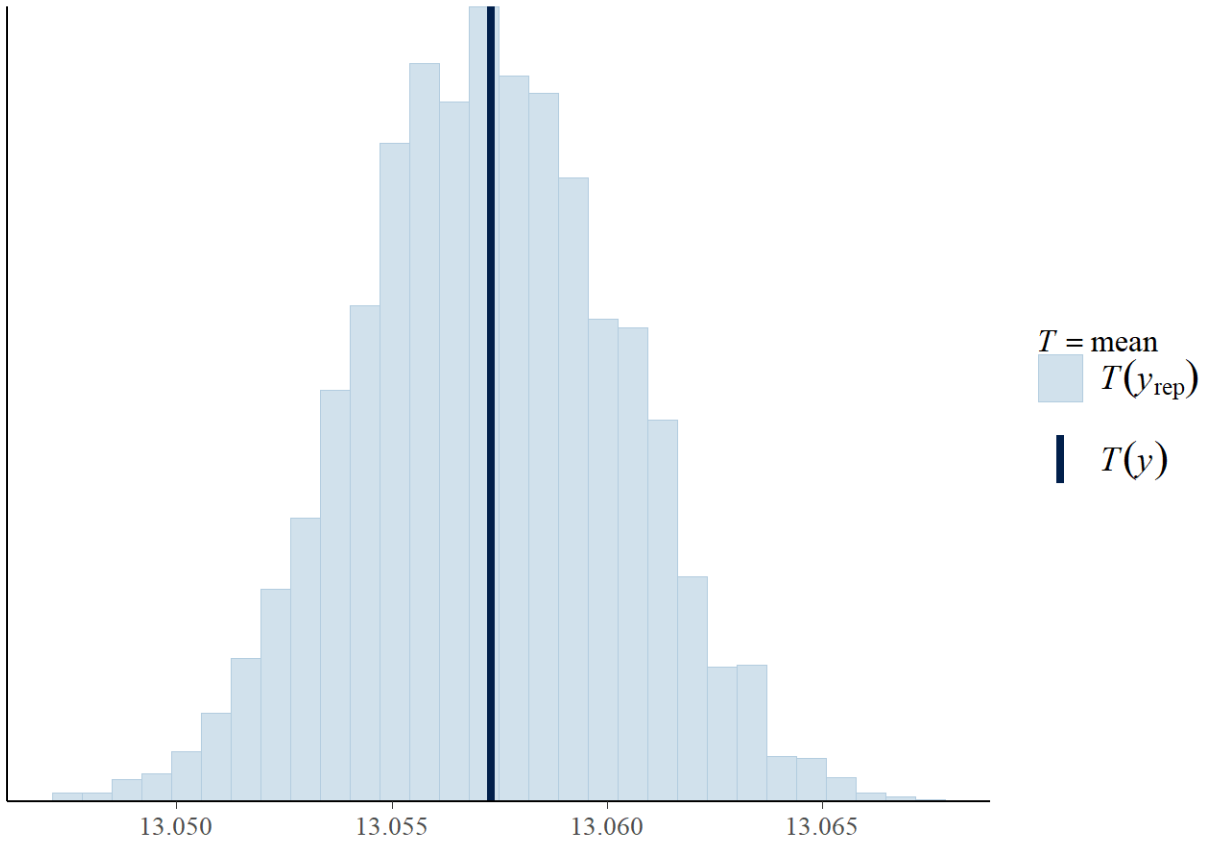
Loo pit plot suggests that our model is not able to predict all the variance in the data.

```
loo_pol <- loo(multiple_polynomial_fit, save_psis = TRUE)
```

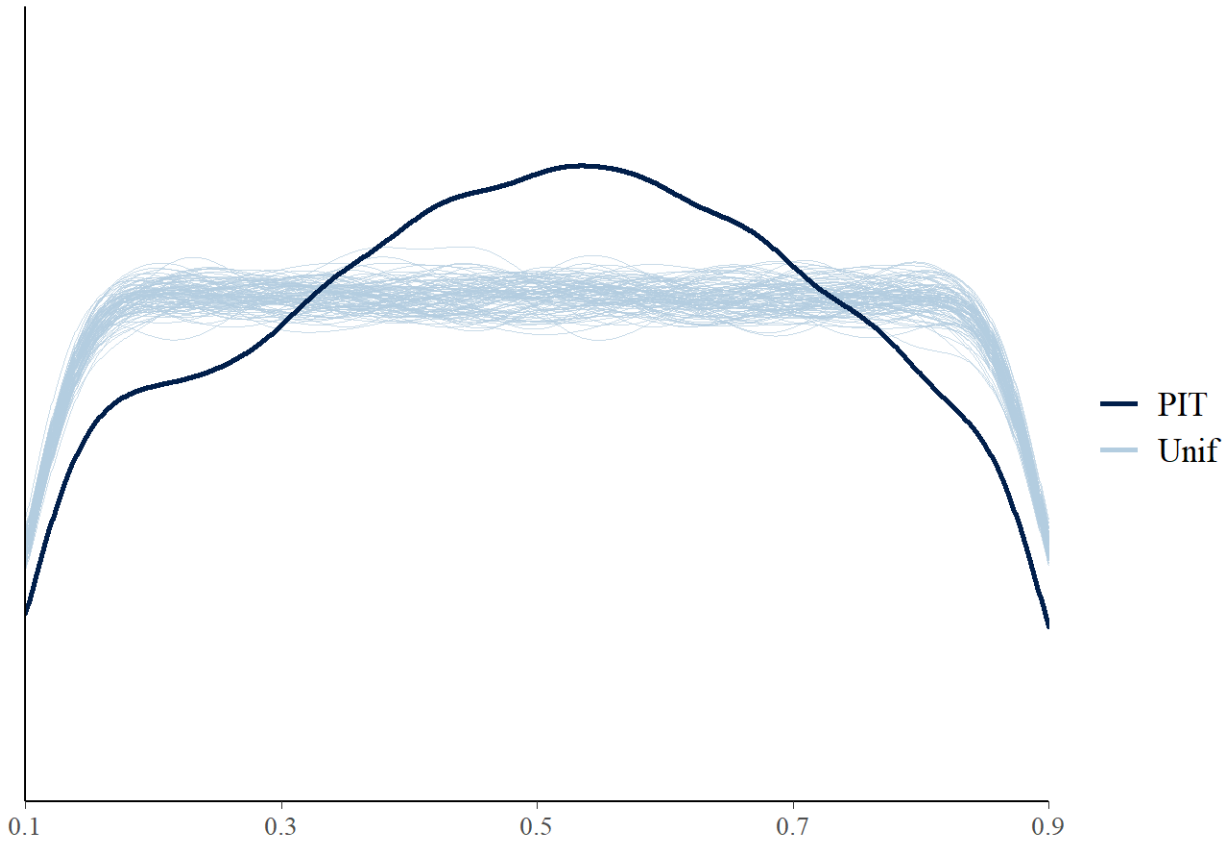
```
## Warning: Some Pareto k diagnostic values are too high. See help('pareto-k-diagnostic') for details.
```

```
psis_pol <- loo_pol$psis_object
lw_pol <- weights(psis_pol)
pp_check(c(original_target[training_indices]), yrep = replicated_data_pol, fun = "stat")
```

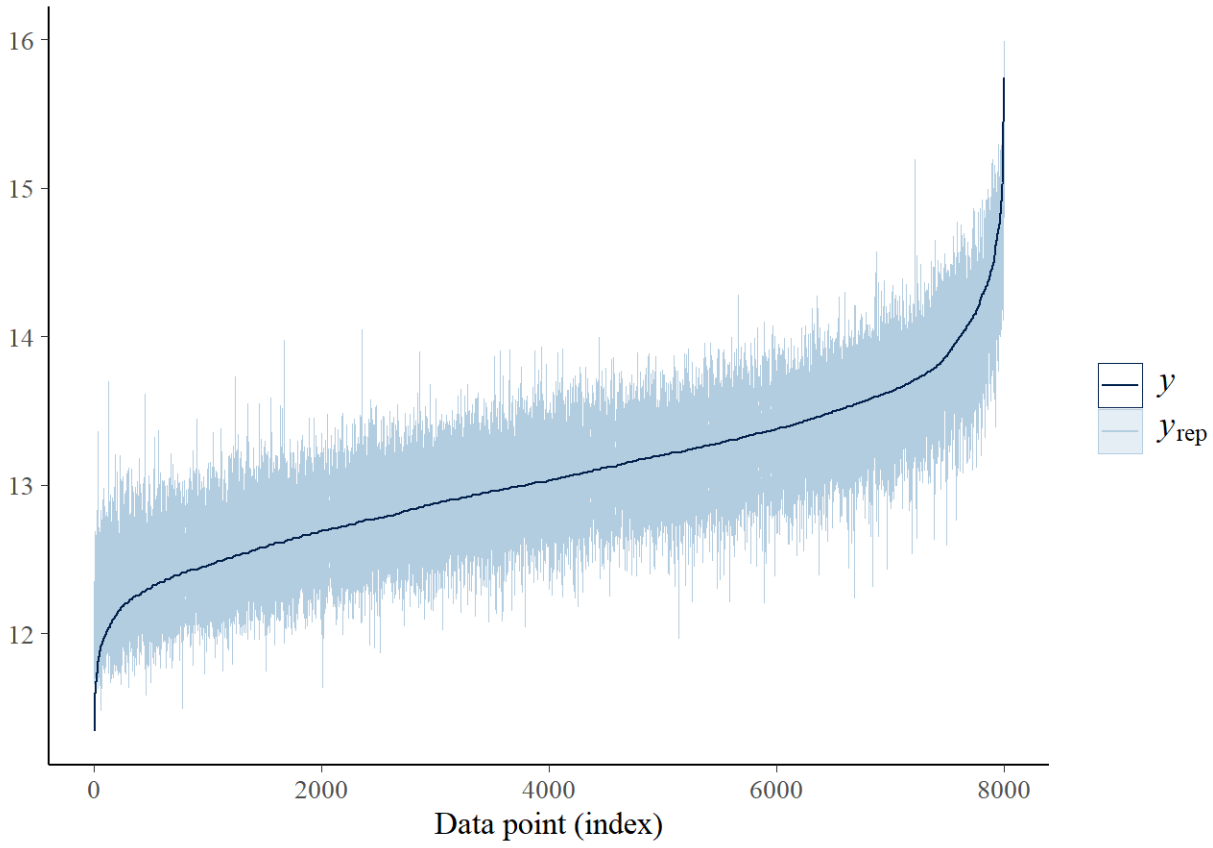
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



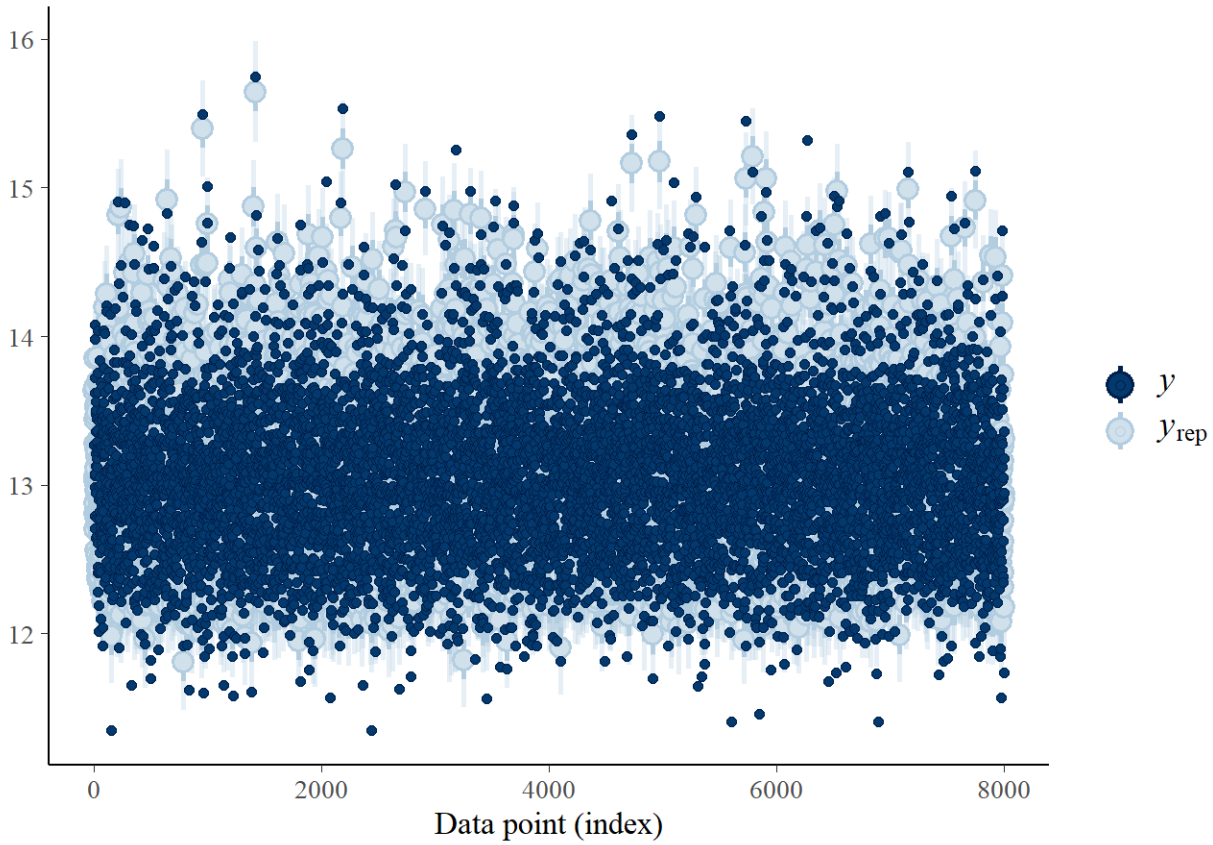

```
ppc_loo_pit_overlay(c(original_target[training_indices]), yrep = replicated_data_pol,
                    lw = lw_pol)
```




```
ppc_loo_ribbon(c(original_target[training_indices][original_training_order]),
               yrep = replicated_data_pol[,original_training_order],
               lw = lw_pol, psis_object = psis_pol)
```




```
ppc_loo_intervals(c(original_target[training_indices]),
                  yrep = replicated_data_pol, psis_object = psis_pol)
```



6. Predictive performance assesment

From the mean squared errors we can see that the polynomial model performed better on the test set

```
# compare errors
data.frame(linear = mae_lin, polynomial = mae_pol)
```

```
##     linear polynomial
## 1 155366.6     147179
```

PSIS-lOO

Obtained elpd information criteria values of the two models are largely the same with the polynomial model having an larger value, suggesting it is better of the two models. The k-values of the models are small expect for one observations for both models suggesting the models fit the data well. The bad k-value is likely caused by the number of observations being too small as I was not able to run the model with the whole dataset. Still the model might be misspesified as p_loo < number of parameters, which corresponds to model being too flexible or having too weak of a population prior [6].

[6] https://mc-stan.org/loo/reference/loo-glossary.html

Multiple linear regression

```
# Extract log-likelihood
multiple_linear_log_lik <- extract_log_lik(multiple_linear_fit, merge_chains = FALSE)

# PSIS-LOO elpd values
r_eff <- relative_eff(exp(multiple_linear_log_lik))
multiple_linear_loo_lin <- loo(multiple_linear_log_lik, r_eff = r_eff)
```

```
## Warning: Some Pareto k diagnostic values are slightly high. See help('pareto-k-diagnostic') for deta
```

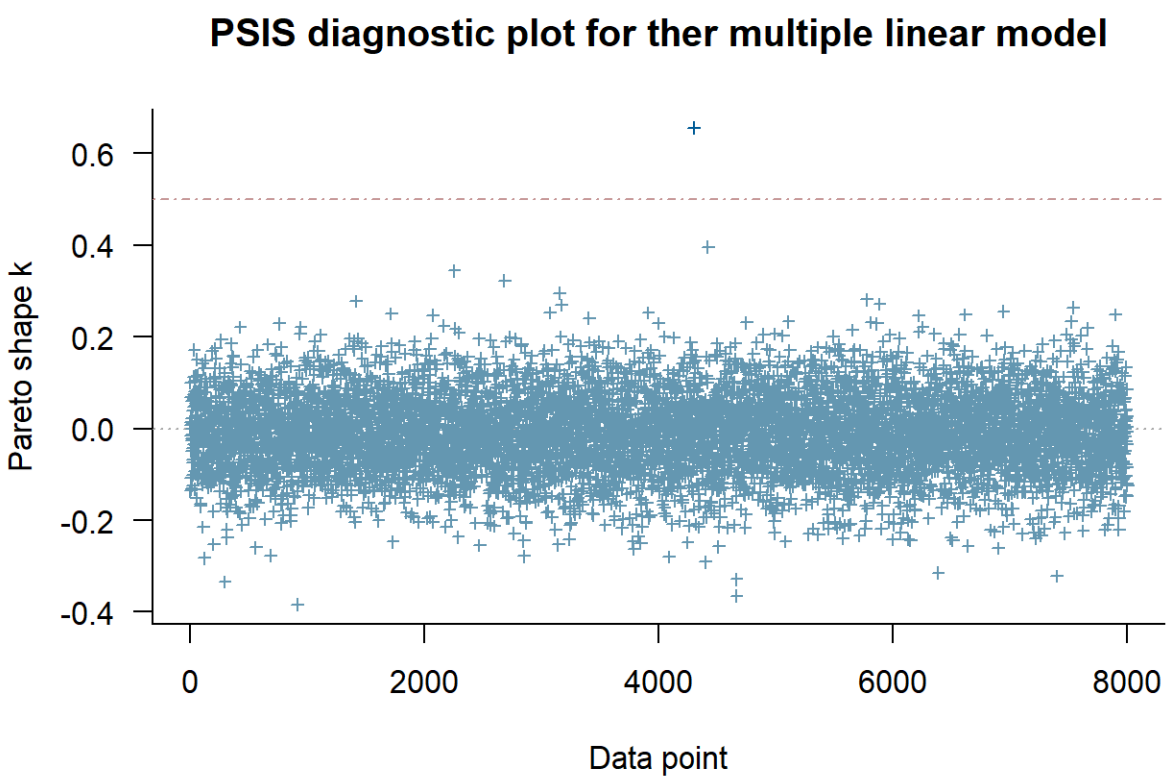
```
#elpd loo
multiple_linear_loo_lin
```

```
##
## Computed from 4000 by 8000 log-likelihood matrix
##
##          Estimate    SE
## elpd_loo  -3558.8 108.1
## p_loo        86.2   4.5
## looic      7117.7 216.2
## ------
## Monte Carlo SE of elpd_loo is 0.2.
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     7999  100.0%  682
##  (0.5, 0.7]   (ok)          1    0.0%  140
##    (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.
```

```
pareto_k_table(multiple_linear_loo_lin)
```

```
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     7999  100.0%  682
##  (0.5, 0.7]   (ok)          1    0.0%  140
##    (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).
```

```
plot(multiple_linear_loo_lin, diagnostic = c("k", "n_eff"), label_points = FALSE,
  main = "PSIS diagnostic plot for ther multiple linear model")
```



Multiple polynomial regression

```
# Extract log-likelihood
multiple_polynomial_log_lik <- extract_log_lik(multiple_polynomial_fit, merge_chains = FALSE)

# PSIS-LOO elpd values
r_eff <- relative_eff(exp(multiple_polynomial_log_lik))
multiple_polynomial_loo_lin <- loo(multiple_polynomial_log_lik, r_eff = r_eff)
```

```
## Warning: Some Pareto k diagnostic values are too high. See help('pareto-k-diagnostic') for details.
```

```
#elpd loo
multiple_polynomial_loo_lin
```

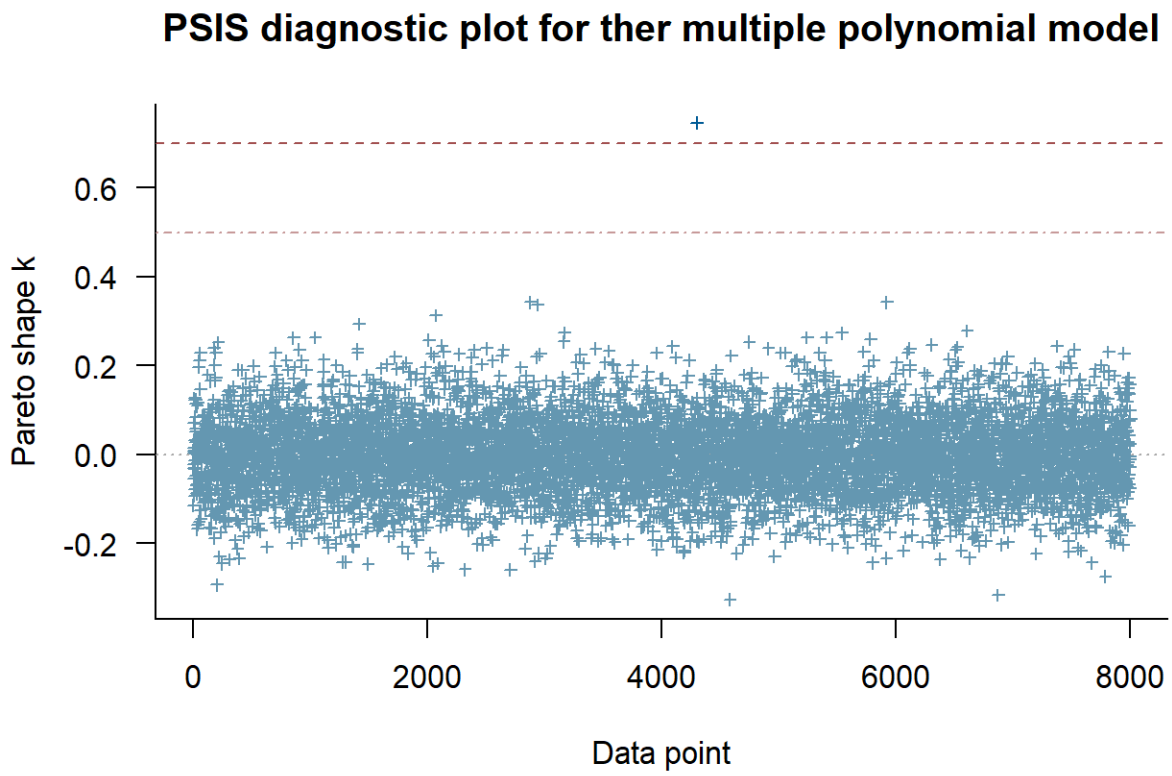
```
##
## Computed from 4000 by 8000 log-likelihood matrix
##
##          Estimate    SE
## elpd_loo  -3315.1  96.7
## p_loo        91.3   3.1
## looic      6630.3 193.4
## ------
## Monte Carlo SE of elpd_loo is NA.
```

```
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     7999  100.0%  515
##  (0.5, 0.7]   (ok)          0    0.0%  <NA>
##    (0.7, 1]   (bad)         1    0.0%  502
##    (1, Inf)   (very bad)    0    0.0%  <NA>
## See help('pareto-k-diagnostic') for details.
```

```
pareto_k_table(multiple_polynomial_loo_lin)
```

```
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     7999  100.0%  515
##  (0.5, 0.7]   (ok)          0    0.0%  <NA>
##    (0.7, 1]   (bad)         1    0.0%  502
##    (1, Inf)   (very bad)    0    0.0%  <NA>
```

```
plot(multiple_polynomial_loo_lin, diagnostic = c("k", "n_eff"), label_points = FALSE,
  main = "PSIS diagnostic plot for ther multiple polynomial model")
```



elpd_loo comparison

Polynomial Model is sligtly better according to elped information criteria.

```
loo_compare(x = list(multiple_linear_loo_lin, multiple_polynomial_loo_lin))
```

```
##        elpd_diff se_diff
## model2    0.0       0.0
## model1 -243.7      51.2
```

7. Discussion

In this report we have explored linear and polynomial regression models for predicting house prices. The differences between the results from the models are small, but the polynomial model performs a bit better. The mean absolute error for both models is over hundred thousand, but considering the mean of the prices is around five hundred thousand, the error rate is small considering the simplicity of the model.

In the future we could consider varying slope parameter by zipcode. This would mean the different predictors would have different effects in the model depending on the zipcode they belong to, which would help us to better undestand how different factors effect housing prices in geograhical areas. However this has few techical drawbacks. There are 70 groups, so using a different beta value for each parameter for each group would increase the number of parameters of the model considerably, likely slowing the model. In addition, the number of data usable for each beta value would shrink, which likely would lead to problems in convergence and biased estimates.