Advanced Bayesian Methods - Final Project

House Price Prediction

Juyeon Park

Department of Statistics and Data Science, Yonsei University

Table of Contents

Introduction

- This dataset contains house sale prices for King County, which includes Seattle.
- It includes homes sold between May 2014 and May 2015.
- The variables can be categorized into 3 groups as below:
  - Transaction variables
    - Variables: id, date, price
  - House Features variables
    - Description: structural characteristics of the house itself
    - Variables: bedrooms, bathrooms, sqft_living, sqft_lot, floors, condition, grade, sqft_above, sqft_basement, yr_built, yr_renovated
  - Geographical variables
    - Description: spatial and surrounding environment information of the house
    - Variables: waterfront, view, zipcode, lat, long, sqft_living15, sqft_lot15

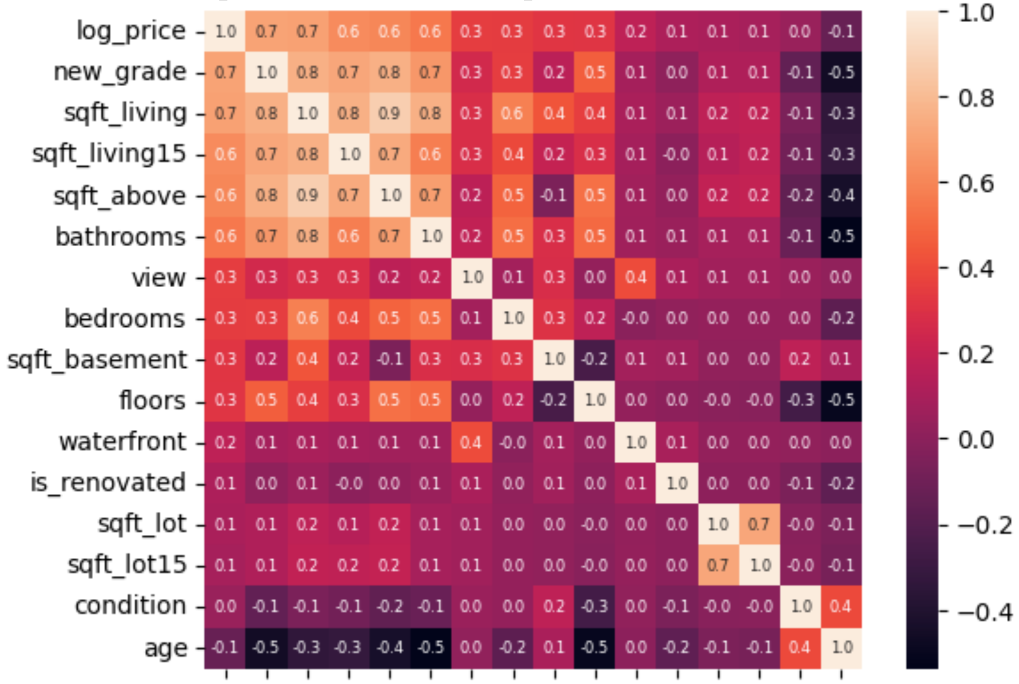
Preprocessing

- log_price
  - Logarithmically transformed the price variable.
  - New target variable.
- is_renovated
  - Indicating whether a house has been renovated. (binary)
  - If yr_renovated is 0 then 0, else 1.
  - 4.5% of the houses were renovated and most of them were built in 1900s.
- age
  - Age of each house as the month difference between the sold date and date of build or renovated.
  - month(date) – month(yr_built or yr_renovated)
- new_grade
  - Grouping spare classes of grade.
  - From 12 grades into 6 grades.

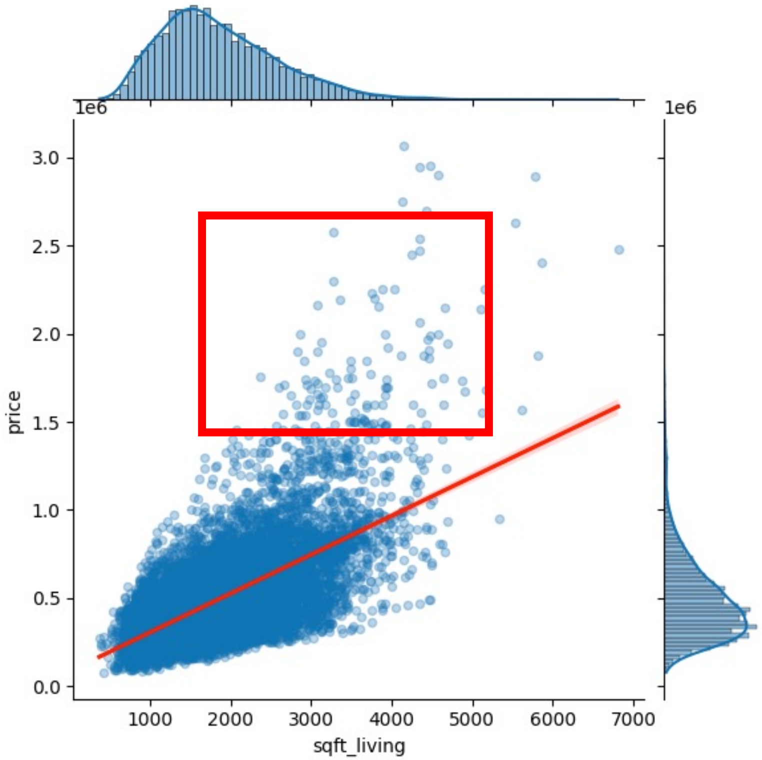
EDA

- "House features variables" show a high correlation with the price.
  - new_grade, sqft_living, sqft_above, bathrooms
  - However, Condition was expected to be highly correlated, but it wasn't.
- "Geographical variables" such as view, riverside, and waterfront have low correlation overall.
  - It is necessary to add variables that can explain spatial information.

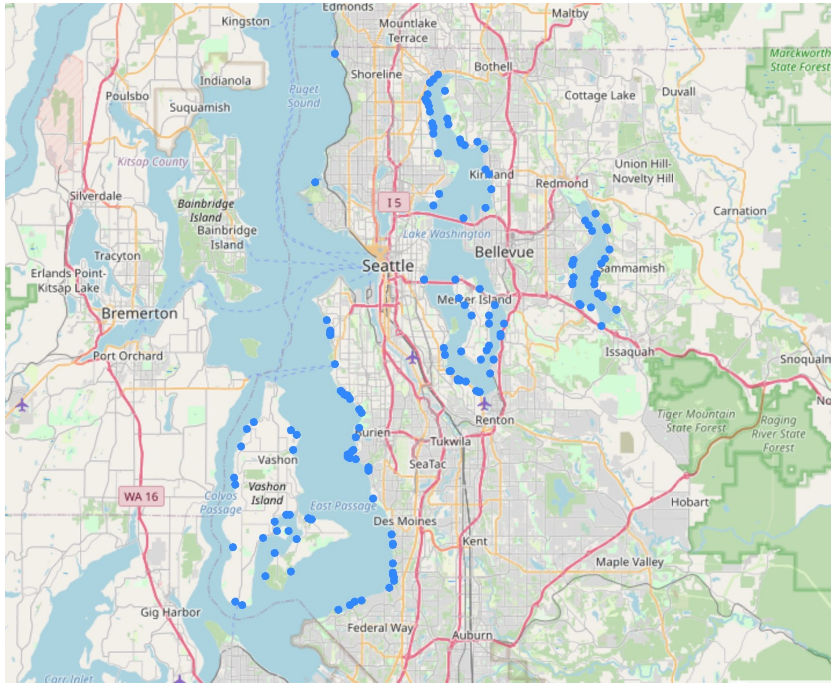
[Correlation matrix]



- Through the joint plot of sqft_living and price, a strong linear relationship can be confirmed.
- However, there exists the observations that prices deviate from the linear relationship.
  - Even thought the house has same living area(sqft_living), it shows a different price.
  - Most of cases, they locate near the river. But the waterfront is not enough.



[Map the waterfront to Seattle]



EDA

- For each observation i, make indicator of whether waterfront exists:

$$new\_waterfront_i = \begin{cases} 1 & if\ \ waterfront_i = 1\ or\ view_i > 1 \\ 0 & othersiwe \end{cases}, \qquad i = 1, \dots, n$$

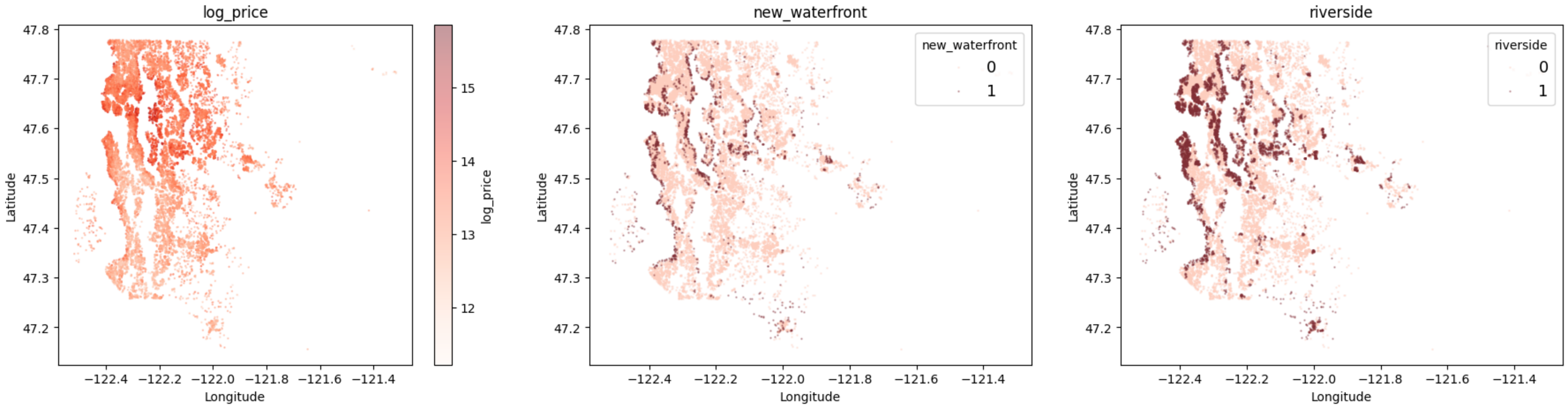
- Calculate Euclidean distance to observation k ($new\_waterfront_k = 1$):

$$distance_{i,k} = \sqrt{(long_i \ - long_k)^2 + (lat_i - lat_k)^2}, \qquad k = 1, \dots, m$$

- Then, a new features riverside is:

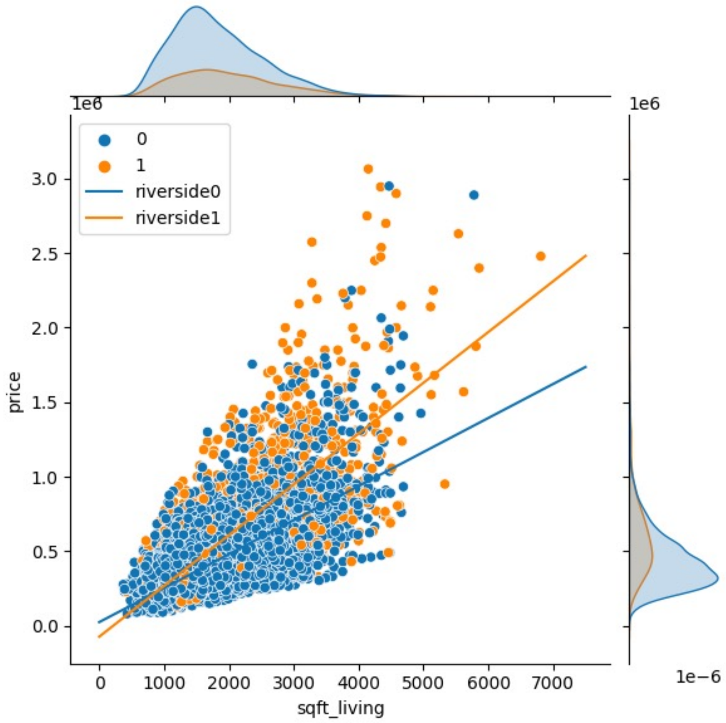
$$riverside_i = \begin{cases} 1 & if \min(distance_{i,k}) < threshold^* \\ 0 & othersiwe \end{cases}$$

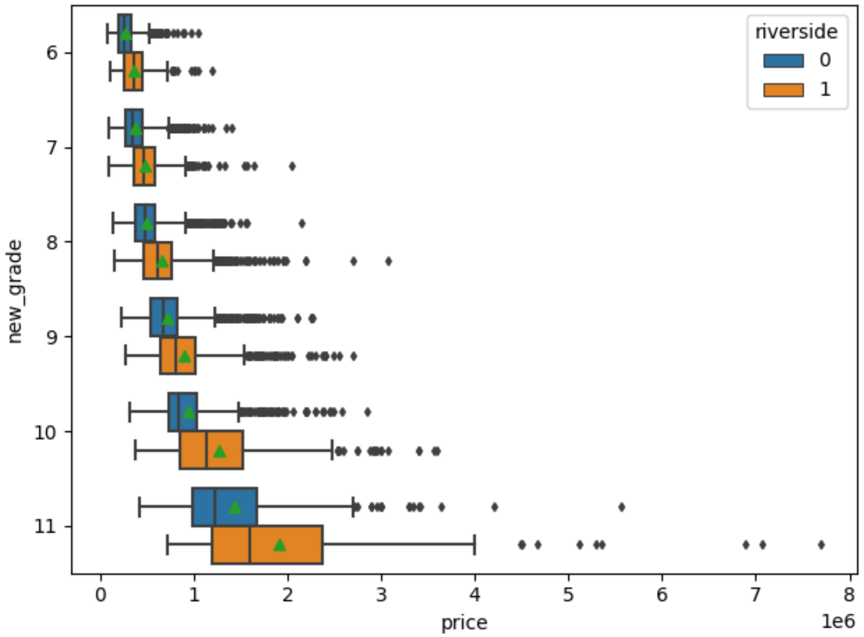
$threshold^*$=0.005



EDA

- The riverside make different linear fit for price.
- Also, there is a difference in price depending on the group divided by riverside.
- Therefore, a hierarchical model with riverside random effect is appropriate. And the predictors are
  - “House features variables”: bedrooms, bathrooms, sqft_living, floors, new_grade
  - “Geographical variables”: new_waterfront , riverside





Hierarchical Linear Model

- Hierarchical Linear Model
  - Likelihood

$$y \sim N(X_1\beta + X_2 u, \sigma^2 I_n), \qquad u \sim N_2(0, \tau^2 I_2)$$

where $X_2 = \begin{pmatrix} 1_{n_0} & 0 \\ 0 & 1_{n_1} \end{pmatrix}$, $\mathrm{n}_0 + \mathrm{n}_1 = \mathrm{n}$

  - Prior

$$\sigma^2 \sim Inv - Gamma(0.3, 0.5)$$
$$\beta \sim N(0, 1000 I_7)$$
$$\tau^2 \sim Inv - Gamma(0.05, 0.05)$$

- Simple Linear model
  - Likelihood

$$y \sim N(X\beta, \sigma^2 I_n)$$

  - Prior

$$p(\beta, \sigma^2) \propto \sigma^{-2}$$

Hierarchical Linear Model

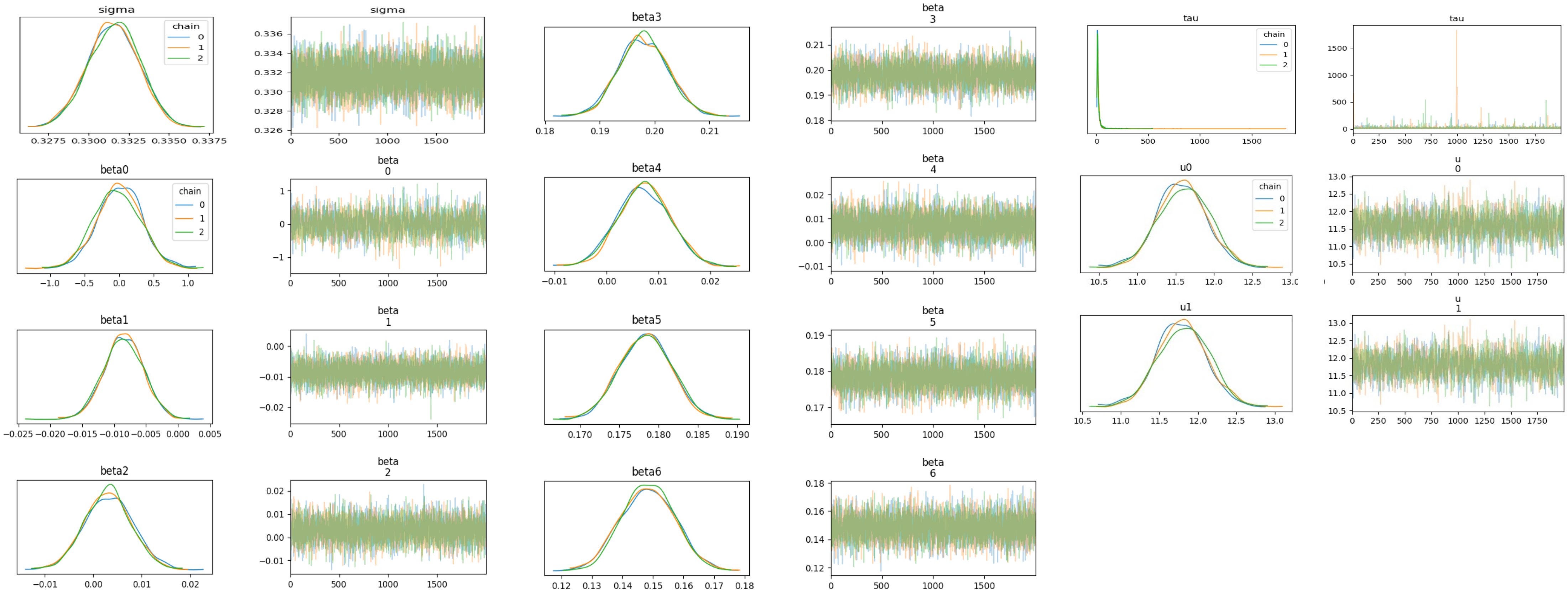
- Posterior mean
  - 3 chains, each with iteration 4000 and warm-up 2000.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
tau1	7.4e-3	8.5e-5	7.3e-3	1.8e-4	2.2e-3	5.2e-3	0.01	0.03	7302	1.0
tau2	9.1	1.2e-3	0.09	8.93	9.04	9.1	9.16	9.27	5490	1.0
beta[1]	-1.5e-3	8.0e-3	0.33	-0.66	-0.22	-2.6e-3	0.22	0.65	1730	1.0
beta[2]	-8.5e-3	4.0e-5	3.1e-3	-0.01	-0.01	-8.5e-3	-6.4e-3	-2.5e-3	5823	1.0
beta[3]	3.1e-3	7.2e-5	4.9e-3	-6.5e-3	-2.7e-4	3.1e-3	6.3e-3	0.01	4705	1.0
beta[4]	0.2	7.1e-5	4.5e-3	0.19	0.19	0.2	0.2	0.21	3906	1.0
beta[5]	7.2e-3	7.2e-5	5.1e-3	-2.8e-3	3.8e-3	7.2e-3	0.01	0.02	5024	1.0
beta[6]	0.18	4.9e-5	3.3e-3	0.17	0.18	0.18	0.18	0.18	4633	1.0
beta[7]	0.15	1.2e-4	8.8e-3	0.13	0.14	0.15	0.15	0.17	5010	1.0
u[1]	11.59	8.0e-3	0.33	10.93	11.37	11.59	11.81	12.24	1728	1.0
u[2]	11.8	8.0e-3	0.33	11.14	11.58	11.8	12.02	12.45	1728	1.0
tau	21.7	1.47	46.04	6.14	9.88	13.82	21.47	75.57	981	1.0
sigma	0.33	2.2e-5	1.6e-3	0.33	0.33	0.33	0.33	0.33	5489	1.0

- The intraclass correlation is 0.99.
  - There is strong within-group variability that would benefit from a random effect.

$$\mathrm{ICC} = \frac{\tau^2}{\tau^2 + \sigma^2} = \frac{21.7^2}{21.7^2 + 0.33^2} = 0.99$$

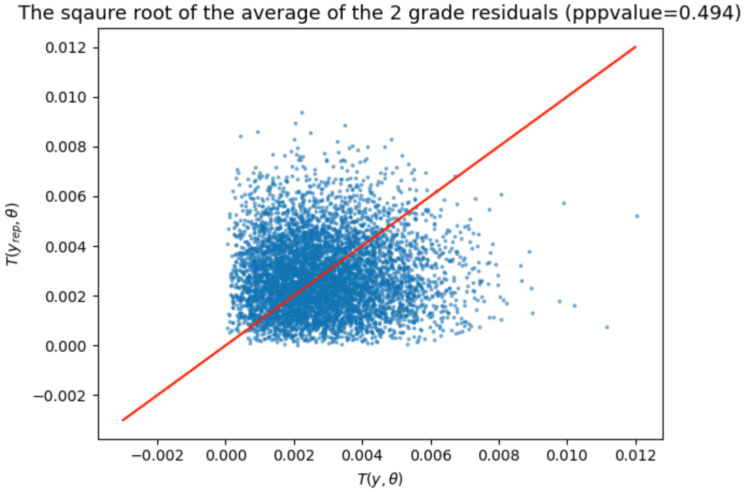
- MCMC Convergence check of Hierarchical Model



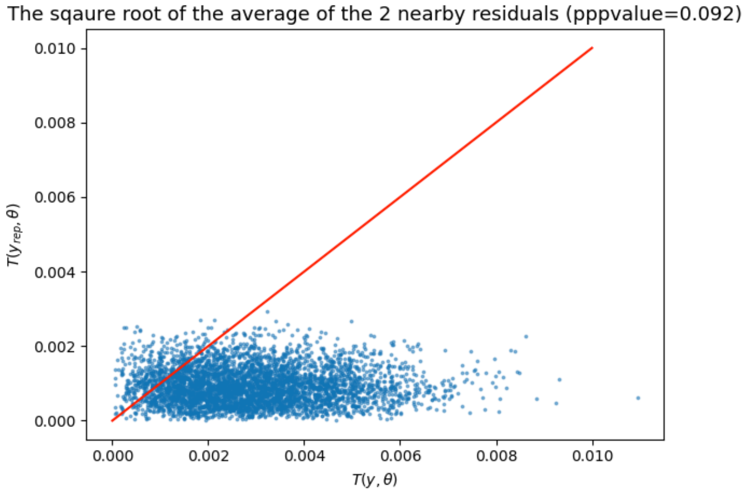
Hierarchical Linear Model

- Predictive posterior check
  - The test quantity is the square root of the average of the residuals per nearby effect:

$$T_2(y,\theta) = \sqrt{\frac{1}{2}\sum_{i=1}^{2}\left(\frac{1}{n_i}\sum_{j\,\in\,nearby_i}(y_j - X_j\beta)\right)^2} \qquad i = 0, 1, \qquad j = 1, \dots n_i$$



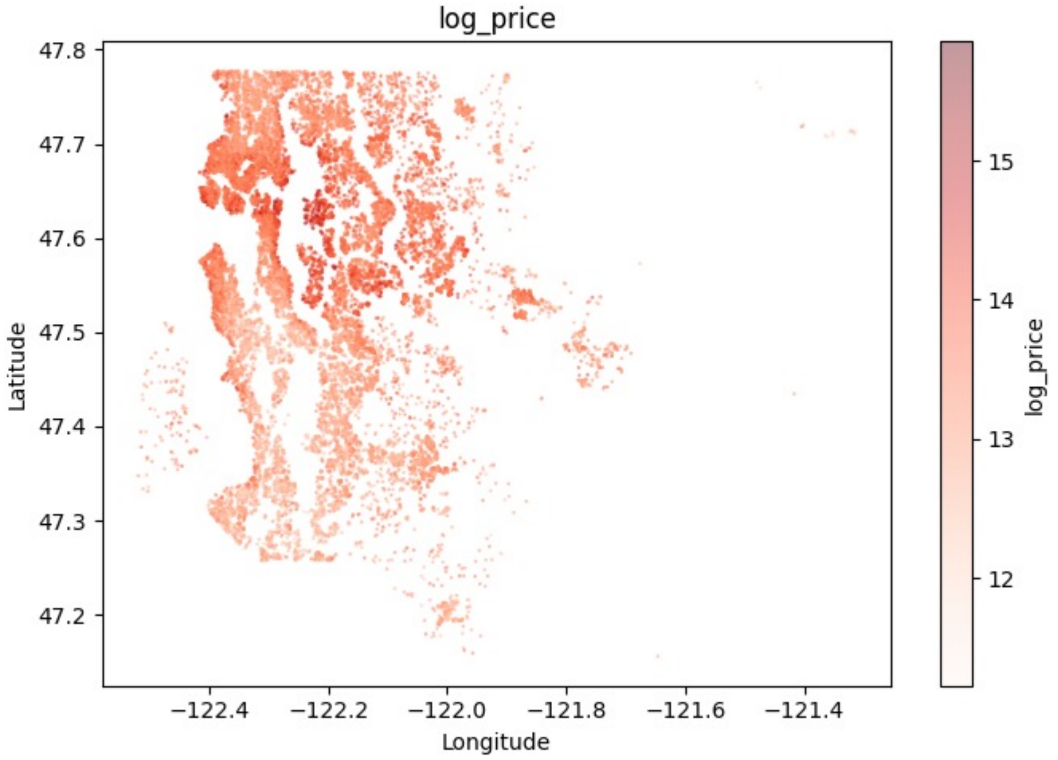
[Hierarchical Model]

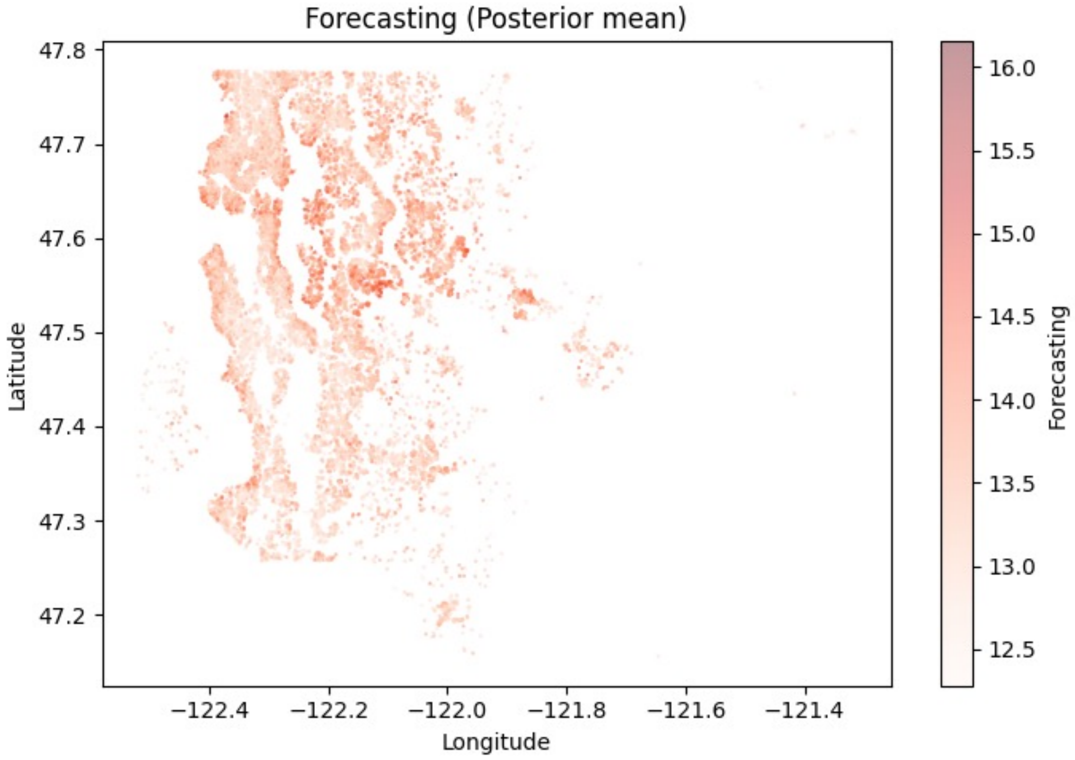


[Simple Linear Model]

- Hierarchical model accurately fit the observed data than simple linear model.

- Forecasting result:



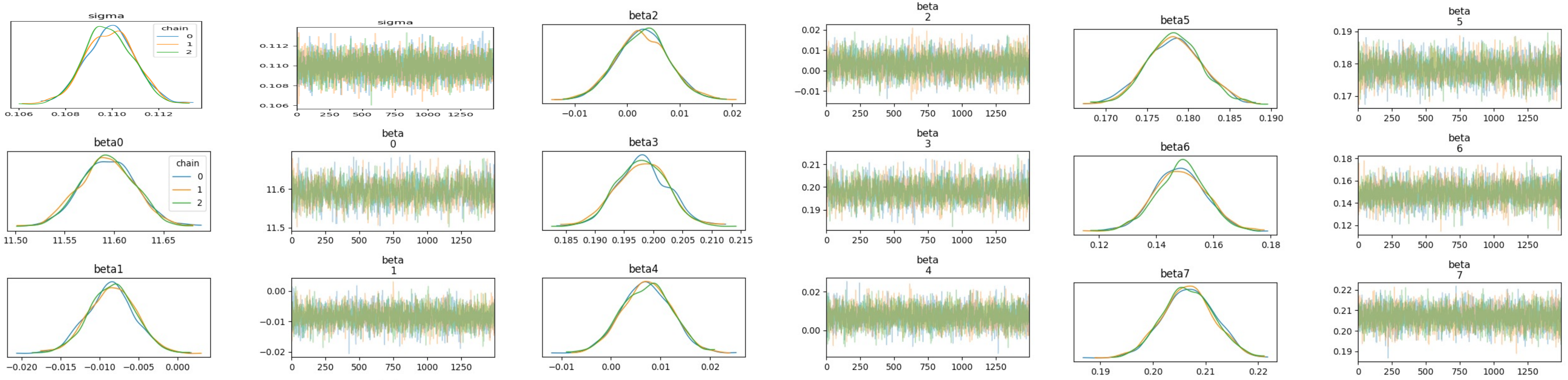


Appendix – Result of Simple Linear model

- Posterior mean

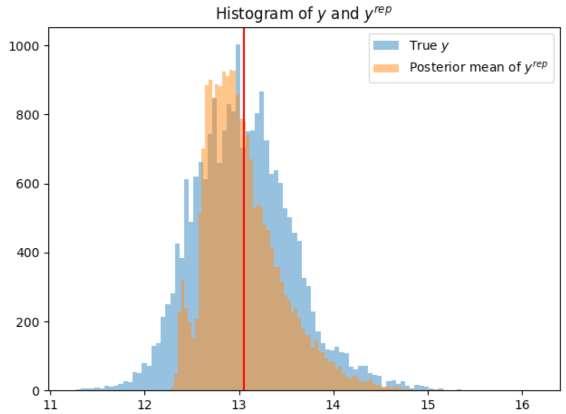
	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
beta[1]	11.59	7.1e-4	0.03	11.54	11.57	11.59	11.61	11.65	1546	1.0
beta[2]	-8.5e-3	5.3e-5	3.1e-3	-0.01	-0.01	-8.4e-3	-6.4e-3	-2.5e-3	3319	1.0
beta[3]	2.9e-3	8.2e-5	4.9e-3	-6.8e-3	-4.5e-4	2.9e-3	6.1e-3	0.01	3594	1.0
beta[4]	0.2	1.1e-4	4.5e-3	0.19	0.2	0.2	0.2	0.21	1745	1.0
beta[5]	7.2e-3	7.7e-5	5.0e-3	-2.6e-3	3.8e-3	7.2e-3	0.01	0.02	4208	1.0
beta[6]	0.18	7.6e-5	3.4e-3	0.17	0.18	0.18	0.18	0.19	1979	1.0
beta[7]	0.15	1.5e-4	8.9e-3	0.13	0.14	0.15	0.15	0.17	3613	1.0
beta[8]	0.21	7.7e-5	4.9e-3	0.2	0.2	0.21	0.21	0.22	4001	1.0
sigma	0.11	1.6e-5	1.1e-3	0.11	0.11	0.11	0.11	0.11	4326	1.0

- MCMC Convergence check

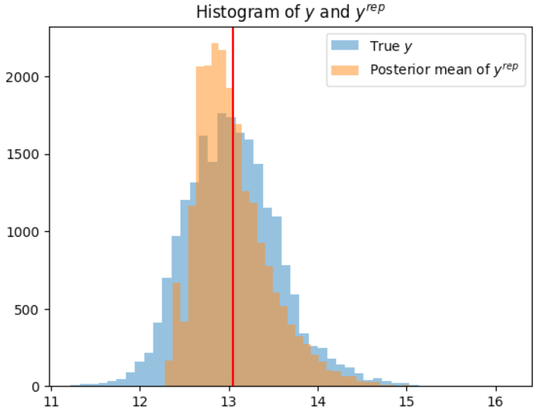


Appendix

- Histogram of y and posterior mean of y^{rep}



[Hierarchical Model]



[Simple Linear Model]

- Correlation matrix for predictors