1. Introduction

Over the past decade, rising prices in China's housing market have been increasingly overwhelming to the young families trying to settle down in major cities in China. As a group of university students potentially facing this issue in our near future, we want to understand how much housing prices in a Chinese metropolis like Beijing can be predicted and explained by some visible factors in the market, for example, characteristics of a house itself.

In view of what we have learned in this course, we set our goal as building a multiple regression model capable of making predictions of the housing price with hope that this model could also provide us with further insights into the housing market.

1. Methods
   1. Data cleaning

The source of data is  from [this page](https://www.kaggle.com/ruiqurm/lianjia) on Kaggle. All the data in this dataset was fetched by its author Qichen Liu from [this page](https://bj.lianjia.com/chengjiao) on the Chinese housing website [Lianjia.com](https://bj.lianjia.com).

The most recent data (2010 to 2017) in the data source is extracted as the dataset for this project. The missing data is removed, as well as the relative predictors are changed to numeric or factors.

The extracted dataset is separated into two groups: training data and testing data. The ratio of these two groups is 6:4 (training to testing). The training data is used for model searching, and the chosen model will be tested with testing data.

* 1. Model searching

1. Observations from the pairs() plot (polynomial relationship) (Model 0) 2. Backward AIC selection from the full model (Model 1) 3. Reducing the selected model to an "all-significant" model (Model 2) 4. Building a small, all-significant model from common sense, observations from the pairs() plot and trials (Model 3)

5. Forward AIC selection from the small model (Model 4)

6. Outlier removal & log transformation (2nd version of Model 1~4)

7. Comparision between Model 1~4

* 1. Chosen Model

1. Results
2. Discussion
3. Is hierarchy necessarily for a predictive model (under the circumstance that a lower-degree term is non-significant)?
4. How well does the final model predict?
5. Which methods were most useful in this experiment?
6. Outliers role in training and testing
7. Appendix

Functions for testing