**Chapter 1 Introduction to the Problem**

* Discussion of the context

With the development of the modern economy, “the housing market is closely linked to consumer spending” (Bank of England 2020), which suggests the interwovenness between the housing market and the national economy. Although the transaction of private properties does not affect the GDP, “the accompanying costs of a house transaction” still benefit the economy (Bank of England 2020). For example, the construction of houses not only creates job opportunities for the labour market in the construction industry but also stimulates the development of the raw material manufacturing industry, which reflects the financial contribution of the housing market to the economy. Despite the macroscopical contribution to the economy, the houses “made a significant contribution to the total asset

of many households” (Lee 2009). According to research, nearly 55% of assets or debts in Australian households are housing and other property (Headey et al. 2005) because housing is one of the necessities of life, so the housing market affects each household specifically.

Therefore, studying the housing market for the sake of being familiar with the housing market plays an essential role in economic research. Studying the factors affecting the housing market should be the key to opening the door to the mystery economy.

* Aims of the proposed research

Due to the importance of the housing market, understanding the component of housing prices can enable the public to know what factors determine house prices. Hence, building the regression model can let people know the weights of each factor contributing to house prices in general. In addition, the model also contributes to the prediction of the house prices, so the buyers can justify the reasonableness of the asking price and the sellers can have better references about what would be the minimally acceptable selling price. Besides, the generalized regression model can be the possible best reference for the investor to understand the logic behind the housing market.

This project mainly studies what affects the price of the housing market, which motivates the project to build the models to explore how much influence each exploratory variables have on the final sale price of houses. Therefore, this project would use the 500 house sales in the last six months record to discover the logic behind the housing market.

Besides the model fitting, the data analysis of the 500 house sales in the last six months can provide insight and overview to the public, so people can better understand the current situation of the property market. For example, through different statistical indexes, people can know what the average and median prices of houses with two bathrooms for a household of three would be.

* Questions of interest

These questions below are those the project mainly focuses on:

* Are there any obvious outliers and how do you deal with them?
* What is the effect of each of the different predictors on the Final House Sale Price?
* Do you need to transform any variables? If so explain why?
* What is the best model for predicting the Final House Sale Price?
* Is the best model a linear model or a more general model?

Outliers can be abnormal points that are far away from other values and influence the result of the model, so handling the outliers would be the first biggest issue at the beginning stage. Due to the topic of finding out the best regression fitting to the data, being aware of the weight of different predictors contributes to the construction of the regression models to notice the pattern of the change in housing prices. Because of the different units of exploratory variables, the transformation of the variables could be under the consideration for the regulation of the variables. As the project has a wide range of use of different regression models, model selection can be the priority among priorities, so the project plans to make comparisons among different models based on the statistical criteria and carefully select the best model for predicting the final house sale price. Of course, the linear model can have clear interpretations of each term of exploratory variables, but perhaps a more general model can be the alternative option providing better fitness of the model.

* Description of the study and variables involved

The study mainly uses the 500 house sales in the last six months as the dataset to find out the best possible regression model for the prediction of the future house price. The dataset records the nine features of 500 sold houses, including elevation of the base of the house, distances to Amenity 1,2 and 3, number of bathrooms, square footage of the house, parking type, amount of precipitation, and final house sale price. This study would use the final house sale price as the response variable to fit the data into linear regression, generalized linear model and some more advanced models, such as Lasso regression and Ridge regression.

**Chapter 2 Description of the methods**

* Description of the statistical methods used

The project mainly uses the descriptive statistical method for the data analysis, so the project would use different statistical indexes not only to do the data cleaning and describe the content of data but also to make the comparison between different models.

For the exploratory data analysis, the boxplots are used to provide insight into the distribution of each variable of the dataset. Due to the features of the boxplot, the outliers of each variable can be detected easily. Subsequently, the histogram for continuous variables and bar charts for categorical variables are used to check how the outliers affect the distribution of data, which will negatively affect the performance of model fitting in the next step. In the next step, the cook’s distance as the significant criterion is used to decide which outliers have a great influence and should be removed. Hence, calculating the cooks' distance needs to fit the simple linear model firstly because it is “a summary of how much a regression model changes when the ith observation is removed” (Thieme 2021). For the problem of multicollinearity, the VIF (variance inflation factor) and GVIF (if exist some categorical variable) are used to measure how much variance of model coefficient estimates is “‘inflated’ due to collinearity between predictors” (DeReuiter 2019).

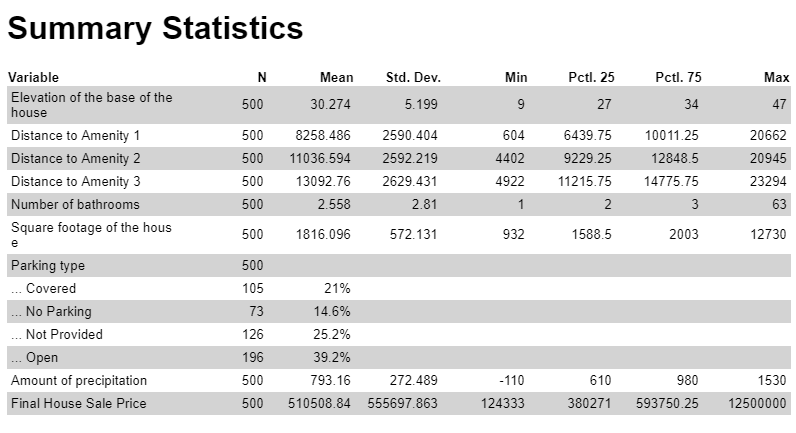
For the model fitting, there are various models used to find out the best possible regression. The project firstly uses the simple linear regression, which assumes the statistical relationship between exploratory variables and response variable is linear with the assumption that the errors are normally distributed with mean zero and constant variance . Additionally, the regression regards the categorical variables as dummy variables that take values 0 and 1, where the values indicate the presence or absence of the categorical effect. After model fitting, there needs four plots for model assumption check, which are residual plot, normal QQ plot, scale location plot, leverage plot. The residual plot requires the residuals are randomly distributed around the zero line with constant variance and without any pattern, and scale location plot expect residuals are randomly scattered around the red line with approximately equal variability, which verify that there is no clear pattern among the residuals. The normal QQ plot expects most of the points are on the line, which satisfies the assumption of data normally distributed. The leverage plot is used to detect whether there are some points are outside of the border of Cook’s distance, which are considered as influential points that should be avoided in the regression.

There are two methods used in the variable selection for the linear regression, which are stepwise method and Bayesian Model Averaging. For both direction stepwise method, it firstly fir the data into full model, and then the algorithm would remove the most insignificant variable or add a significant variable at a time, and this process would do it again and again until the model reaches the minimum AIC (Akaike information criterion), and the lower AIC is, and the better model would be.

**Chapter 3 Analysis of the data**

* Exploratory data analysis

The data relating to the 500 sales of the houses in the last six months records the eight features of each house, and there is no missing entry in the dataset. The features include elevation of the base of the house, distance to Amenity 1, 2 and 3, number of bathrooms, square footage of the house, parking types, amount of precipitation, and specifically the parking types are covered, no parking, not provided, open. For the first insight of the summary statistics table, the average price of houses is 51,0508.84, but with a great variation in the price among the 500 house sales. The standard deviation of that is even higher than the mean value, which leaves the suspicion of the outliers existing, and there is the same situation happening to the number of bathrooms. For the parking types as the only categorical feature in the data, most of the houses come with open space parking lots while over a fifth of the houses has roof-covered parking lots. However, there is nearly 40% of house owners have the problem of parking due to no parking lot or parking lots not provided. Another observation is about the average distances to amenities 1,2 and 3, and on average the houses are closed to amenity 1.

****Figure 1. Summary Statistics

Based on the boxplots, there exist outliers in each variable, and especially, the variables including the number of bathrooms, square footage, and price have the outliers with great influence, which “compressed” the shape of the boxes because the outliers have the extreme values which make other values seem to be insignificant in the plots. Besides that, other boxplots seem to be normal even with the outliers.

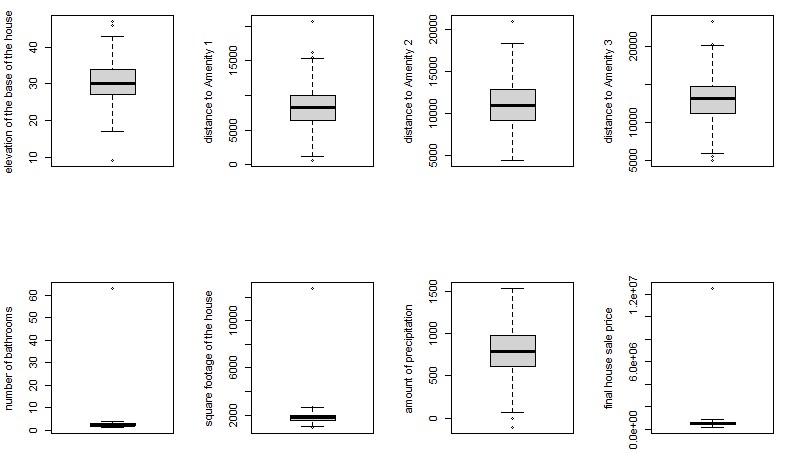


Figure 2. Boxplots

To decide on which outliers should be dropped, the cook’s distance is the criteria to detect the outliers’ influence on the dataset. It takes into consideration both the leverage and residual of each observation “when performing a least-squares regression analysis” (Wikipedia 2022). As shown in figure 3, observation 348 have a huge influence by comparison to other observation, so it would be deleted. After dropping it, figure 4 displays that the observations of continuous variables are approximately normally distributed without obvious skewness.

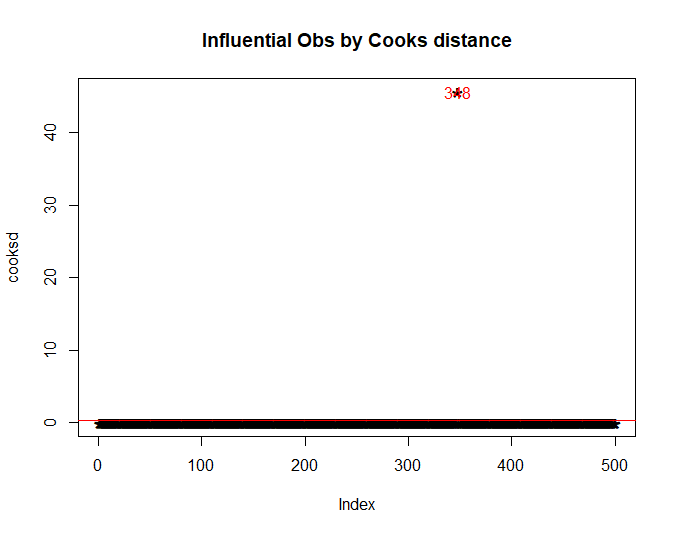
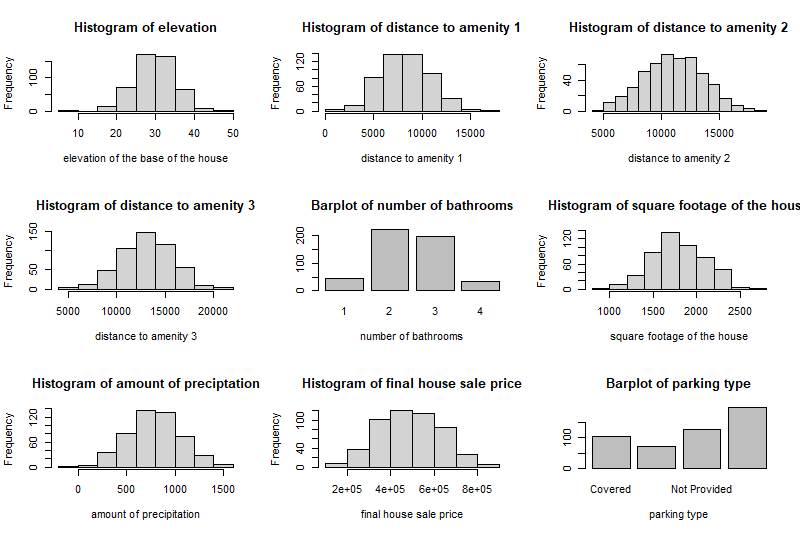
 

Figure 3 Cooks Distance Figure 4 Distribution

As Figure 5 shown, multicollinearity could be a concern in the model fitting, and there exists a high correlation between variables dist\_am1 and dist\_am3, and medium correlations happen to dist\_am2 with dist\_am1 and dist\_am3.

To address this issue, VIF (VIF = ) can be used to “measure the factor by which variance of model coefficient estimates is “inflated” due to collinearity between predictors”, and GVIF is introduced when the regression includes categorical predictors (DeReuiter 2019). To obtain the values of GVIFs, fitting all variables to the simple linear model is the first step, and calculating each variable scaled GIVF based on its degree of freedom. However, the GVIF table (Figure 6) eliminates this concern due to all squared scaled GVIFs () being less than 4 (DeRuiter 2019).

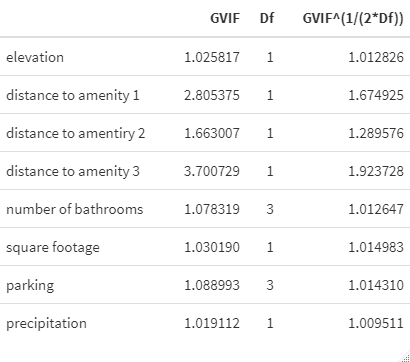
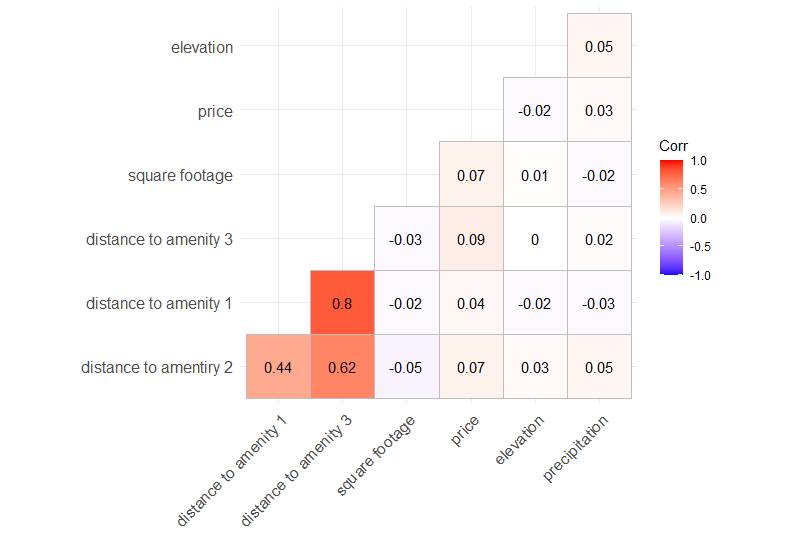


Figure 5. Correlation Pair Plot Figure 6. VIF Table

Based on the plots from Figure 7, there isn’t a clear relationship between price and each variable, except the one about the number of bathrooms which suggests the more bathrooms the houses have, the higher price the houses can be sold, and the correlation between variables bath and price are 0.9270249, which suggests that the number of bathrooms is strongly correlated to the price of the houses.

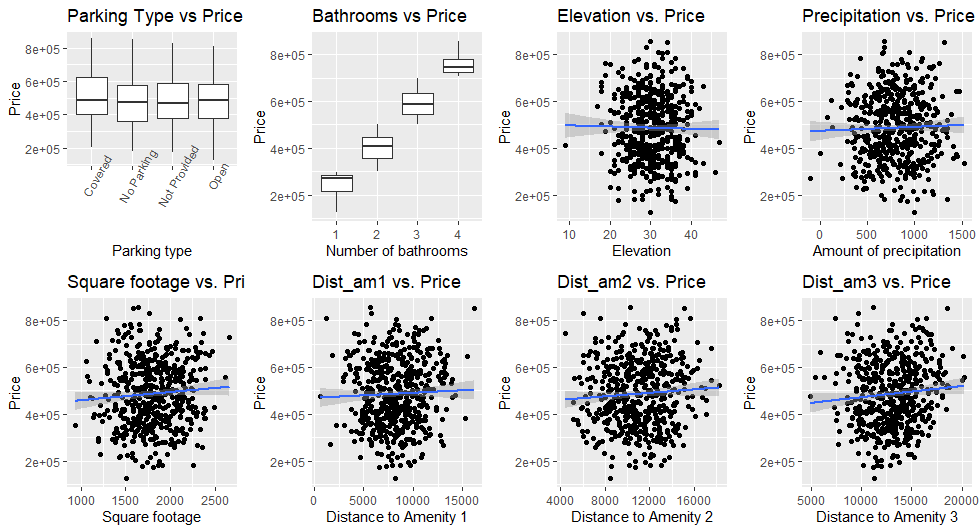


Figure 7. The relationship of price with each variable

* **Analyses and model checks**

Linear Regression Model (Stepwise AIC Model Selection)

With the use of the stepwise AIC model selection method, it produces the best possible linear regression as price = 37985.07 + 175189.60 \* bath + 12.03\*sqft, which suggests that the price will increase 175189.60 pounds with one more bathroom and increase by 12.02 pounds with one square footage increase holding other factors constant for each case. By comparison to the full model on the right, the AIC of the best possible linear model is lower, and the best possible has the less independent variables to explain the same percentage of variation as the full model does.

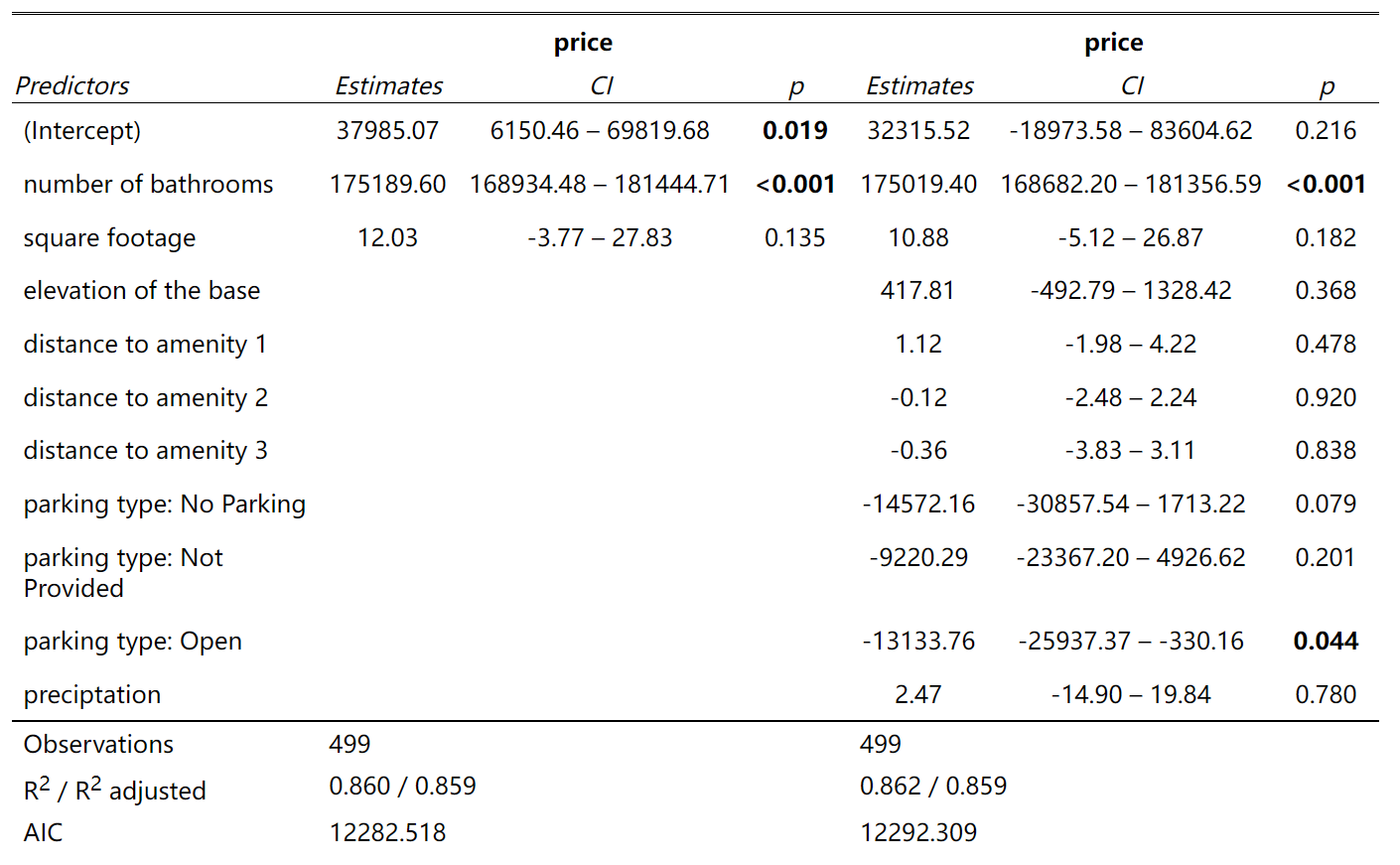


Figure 8 Stepwise Linear Regression

After fitting the data to the linear regression model, the four plots below indicate that the model does not violate the assumptions. The residual plot shows that the points seem to divide into four groups because the variable bath has a strong correlation with the response variable price, and there are only four levels for the variable bath.

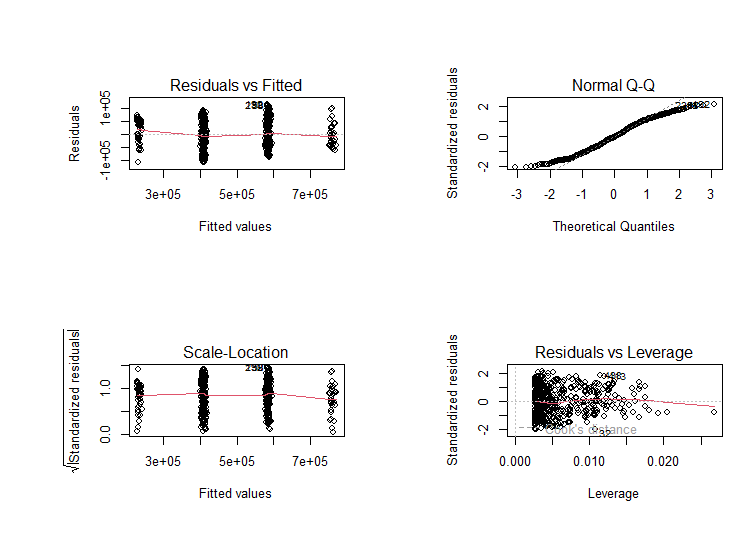


Figure 9 Stepwise Linear Regression Model Check

Linear Regression (BAS model selection)

For the stepwise model selection, some predictors are considered insignificant to the model, but those could contribute to the response variable, and whose existence in the model will not make trouble with the model fitting. Due to these features, stepwise model selection would only conclude one final model without considering more possibilities. Hence, to discover more possibilities, the Bayesian Model Averaging method is used to detect how many probabilities would the predictors exist in the best possible model based on Bayesian statistics. Figure 10 shows that the variable bath has the highest probability to be in the best possible model, and figure 11 shows the model only including the variable bath would be the top 1 option. Although the ranked one model only includes the predictor number of bathrooms, figure 11 shows the possibility that other models can be the alternative option with a lower “score” (i.e., log posterior odds) but those have more predictors to explain the variation in the data.

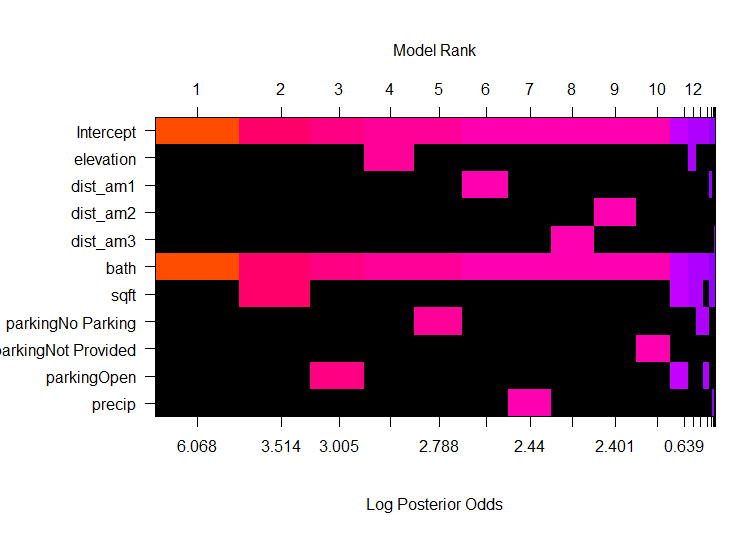
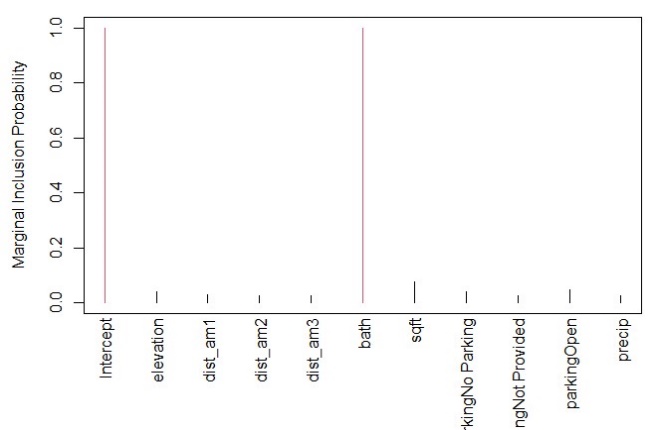
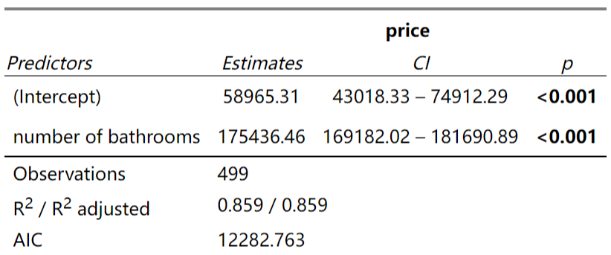


Figure 10 Marginal Inclusion Probability Figure 11 BAS Model Rank

Figure 12 BAS Linear Regression

Based on the Bayesian Model Averaging method, the best possible linear regression model is price = 58965.31 + 175436.46\*bath with 0.859 of adjusted R square. The model suggests that the price will increase by 175436.46 pounds with one more bathroom holding other factors constant. The four graphs in figure 13 do not indicate the model violet the assumption of the linear regression model.

图示

描述已自动生成

Figure 13 BAS Linear Regression Model Check

GLM Gamma Model

The linear regression model includes too less predictors in the final model after variable selection. Hence, here the generalized linear model is introduced to seek a model that perhaps has more predictors to explain the variation of data more generally, and the Gamma model is used because of the continuous positive outcome variable. For the Gamma Model, via the stepwise AIC-based model selection, the identity link function is the best one among the three link functions due to the lowest AIC. And the best possible Gaussian model is price = 63492.81 + 896.46\*elevation + 170702.36\*bath + -14797.96\*parking (No Parking) + -10089.62\*parking (Not Provided) + -15832.80\*parking (Open) + -12.32\*precip. For each case holding other factors constant, it suggests that the house price will increase by 896.46 pounds with one more unit of elevation, increase by 170702.36 pounds with one more bathroom, or decrease by 12.32 pounds with one more unit of precipitation, and the house price will decrease by over ten thousand pounds unless the parking is covered. Also, due to (2) < 39.4136, the Gaussian model satisfies the condition of the goodness of fit.

Identity Link Inverse Link Log Link

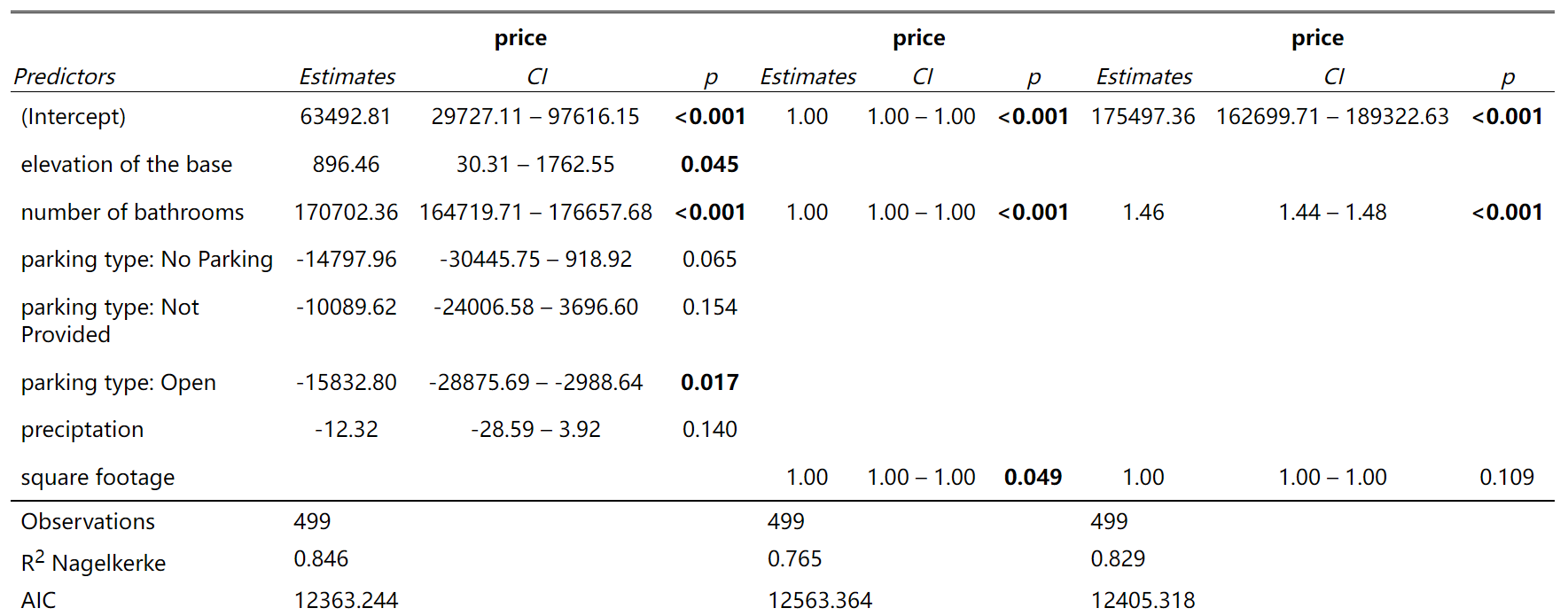


Figure 14 Gamma Model

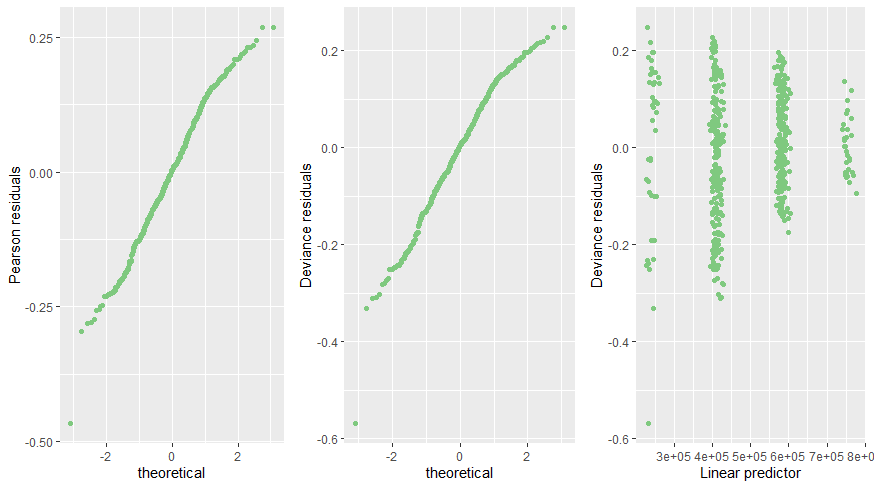


Figure 15 Gaussian Model Check

Ridge Regression

Although the linear regression and GLM perform well in the model fitting, the Regularized regression with the shrinkage can prevent the model from overfitting, and in general, it applies the penalty term to OLS (ordinary least squares) regression, which might come up with a better model. Mathematically, the objective function is to find a hyperplane in multidimensions or a line in two dimensions which minimizes the sum of squared errors between the observed values and predicted values, so the objective function can be written as . When lambda equals zero, there is no effect, and it is just the normal OLS regression. When lambda approaches infinity, it applies the heavy penalty to the regression that forces the coefficients toward zero. Figures 16 and 17 provide insight into how the coefficients change under the different lambda and degrees of freedom.

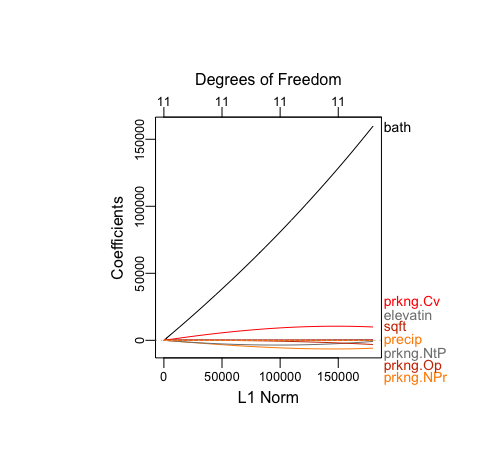
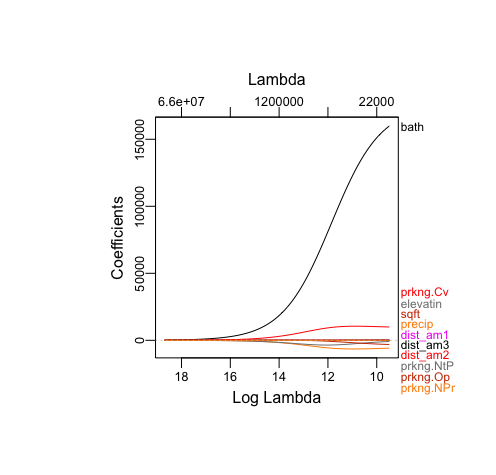


Figure 16 Ridge log lambda vs coefficients Figure 17 Ridge DF vs coefficients

Figure 18 shows the optimal lambda which can minimize the mean squared error, so the optimal lambda is 13159.47. Based on the optimal lambda, figure 19 shows how much influence each variable has on the response variable. Since the Ridge regression does not have the feature of variable selection, the regression fits all variables to the final model, but by comparison of the scale of each coefficient, the variables indicating the number of bathrooms and parking type have the dominant effect on the houses price, which suggests that the price would increase by 159873.2 with one more bathroom holding other factors constant, and the price would only increase by 10327.08 when the parking is covered, but other than that, the price would decrease holding other factors constant.

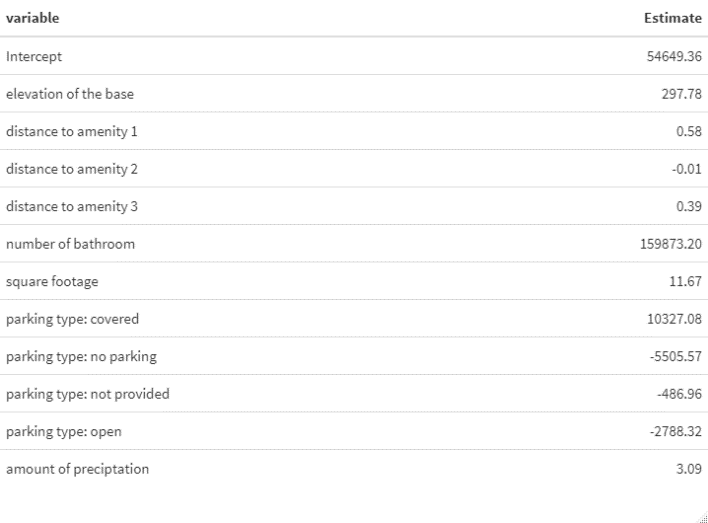
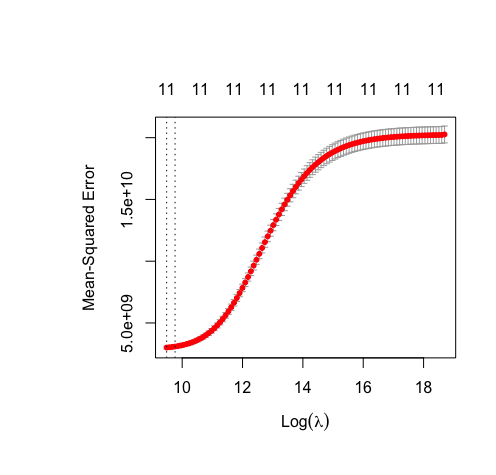


Figure 18 Ridge MSE vs Log lambda Figure 19 Ridge Model

Lasso Regression

In comparison to Ridge Regression, the Lasso Regression have the penalty term , and more importantly, it has the features of model selection. With the use of Lasso Regression, the shrinkage parameter and the parameter of the variable would be the key to the model, and figures 20 and 21 below display how the coefficients of parameter behave under the different values of lambda, so based on the graph, it is beneficial to have insight about the lambda and variable selection for the best possible regression.

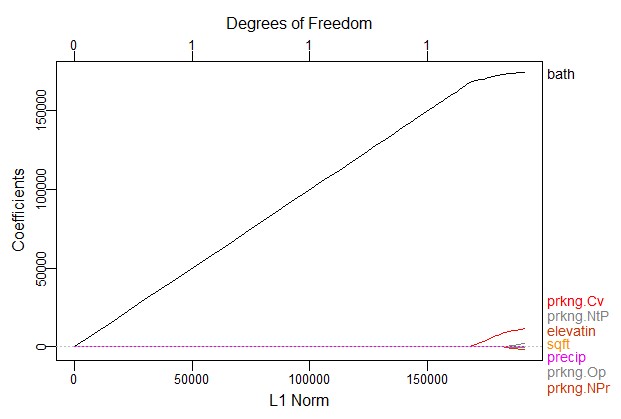
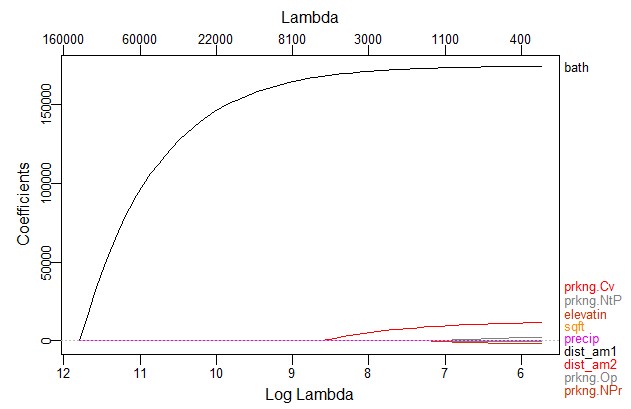


Figure 20 Lasso log lambda vs coefficients Figure 21 Lasso Degree of Freedom vs coefficients

To minimize the sum of squares based on the law of Lasso regression, graph 22 below is to help to find out the optimal shrinkage parameter, which is 2644.044. Based on the variable selection by Lasso, the variables with their coefficients are displayed in figure 19, and the R square is 0.8601869. Hence, the price would increase by 172213.88 pounds with one more bathroom and by 4.05 pounds with additional square footage, and also if the parking type is covered, the price would increase by 6984.49 pounds holding other factors constant for each case.

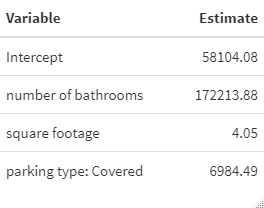
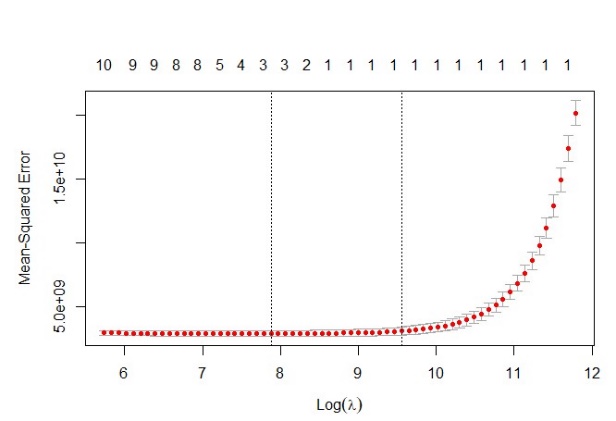


Figure 22 Lasso MSE vs Log lambda Figure 23 Lasso Model

Reference

“How Does the Housing Market Affect the Economy?” Bank of England, Bank of England, 18 Mar. 2020, <https://www.bankofengland.co.uk/knowledgebank/how-does-the-housing-market-affect-the-economy#:~:text=The%20housing%20market%20is%20closely,or%20pay%20off%20other%20debt>.

Lee, Chyi Lin. “Housing Price Volatility and Its Determinants.” International Journal of Housing Markets and Analysis, Emerald Group Publishing Limited, 27 Mar. 2009, <https://www.emerald.com/insight/content/doi/10.1108/17538270910977572/full/pdf?title=housing-price-volatility-and-its-determinants>.

Headey, Bruce, et al. “The Structure and Distribution of Household Wealth in Australia.” The Australian Economic Review, vol. 38, no. 2, 2005, pp. 159–175., <https://doi.org/10.1111/j.1467-8462.2005.00363.x>.

DeRuiter, Stacy. “Stat 245 Course Notes.” Chapter 11 Collinearity and Multicollinearity, Calvin University, 2 Dec. 2019, <https://stacyderuiter.github.io/s245-notes-bookdown/collinearity-and-multicollinearity.html>.

Wikipedia. “Cook's Distance.” Wikipedia, Wikimedia Foundation, 11 July 2022, <https://en.wikipedia.org/wiki/Cook%27s_distance>.

Thieme, Christian. “Identifying Outliers in Linear Regression - Cook's Distance.” Medium, Towards Data Science, 16 June 2021, <https://towardsdatascience.com/identifying-outliers-in-linear-regression-cooks-distance-9e212e9136a#:~:text=Cook's%20Distance%20is%20an%20estimate,the%20ith%20observation%20is%20removed>.