Dissertation on Housing Market Analysis

Best Possible Regression

Emmanouil Mertzianis, Student ID: 2600474

10/10/2021

Contents

1 Introduction

The housing market is an important factor of the global economy both directly as a form of investment and indirectly, because of its strong ties with the overall robustness of the system and its crisis recovery ability [https://www.imf.org/en/News/Articles/2015/09/28/04/53/sp060514]. As Min Zhu, Deputy Managing Director of IMF, stated, *"IMF research shows that of the nearly 50 systemic banking crises in recent decades, more than two thirds were preceded by boom-bust patterns in house prices"* [https://www.imf.org/en/News/Articles/2015/09/28/04/53/sp060514].

The primary objective of this paper is to find the best possible regression model to accurately *predict* the sale price of houses. It is desirable that the model allows for a meaningful interpretation, however our main goal is prediction. For this purpose, 500 observations of house sales in the last six months have been collected by a firm, with measurements on the following variables:

- **elevation**: Elevation of the base of the house
- **dist_am1**: Distance to Amenity 1
- **dist_am2**: Distance to Amenity 2
- **dist_am3**: Distance to Amenity 3
- **bath**: Number of bathrooms
- **sqft**: Square footage of the house
- **parking**: Parking type
- **precip**: Amount of precipitation
- **price**: Final House Sale Price

The rest of the material in this paper is arranged as follows: in the next section we explore the available variables and their relationships via numerical and graphical summaries to make educated decisions when constructing our models. The following section contains the formal design, application and assessment of statistical regression models using our data with the ultimate goal of finding the best possible predictive model. Finally, we summarise our findings and we note ideas that could extend the analysis of our paper.

~~ TODO ~~

2 Exploratory Data Analysis

Before we start searching for the best regression model through formal data analysis and model fitting, it is important to explore our data through numerical and graphical summaries. This will allow for a better understanding of the patterns in and the structure of our data and it will enable us to make educated decisions during model fitting. For this purpose, we start by exploring each variable individually and, then, we focus on the relationships between the variables with emphasis on the ones related to the sale price, which is the variable of interest.

2.1 Exploring variables individually

Table 1: Summary statistics for the categorical variables in the initial data set.

Variable	Unique lvls	Counts
bath	5	1: 46, 2: 222, 3: 198, 4: 33, 63: 1
parking	4	Covered: 105, No Parking: 73, Not Provided: 126, Open: 196

Table 2: Summary statistics for the numerical variables in the initial data set.

Variable	Mean	SD	Min	25%	Median	75%	Max
elevation	30.274	5.198555e+00	9	27.00	30.0	34.00	47
dist_am1	8258.486	2.590404e+03	604	6439.75	8219.0	10011.25	20662
dist_am2	11036.594	2.592219e+03	4402	9229.25	11015.0	12848.50	20945
dist_am3	13092.760	2.629431e+03	4922	11215.75	13188.0	14775.75	23294
sqft	1816.096	5.721306e+02	932	1588.50	1770.5	2003.00	12730
precip	793.160	2.724887e+02	-110	610.00	790.0	980.00	1530
price	510508.840	5.556979e+05	124333	380271.00	481042.0	593750.25	12500000

As a first step, we are interested in the summary statistics of the individual numerical and categorical variables in our data. The tables 1 and 2 contain useful statistics about the variables, prior to making any alterations to the original data set. We note that there are no missing values for any of our variables in the data.

Regarding the categorical variables, we observe that there are five and four unique levels for the categorical variables *"bath"* and *"parking"*, respectively. Table 1 shows that most of the sale entries refer to houses with two baths or an "open" type parking. However, the most important observation to note here is a single entry with 63 bathrooms, which is exceedingly higher than all the rest observations in our data that are limited to just 4 bathrooms at maximum. Such an observation is likely to be an outlier and the exploratory analysis to follow further underpins this assumption.

Table 2 shows statistics about the numerical variables. It becomes apparent that the numerical variables are measured in different numerical scales with differences in the magnitude of their values. In terms of magnitude and standard deviation in ascending order:

- *"elevation"* presents the smallest values that do not exceed the value of 47 and the smallest standard deviation.
- *"precip"* and *"sqft"* come second and third, respectively, with the latter having almost double the standard deviation of the former.

- The three variables representing the distance from three chosen amenities (i.e. *dist_am1*, *dist_am2* and *dist_am3*) exhibit almost the same standard deviation. However, it seems that the 75^{th} percentile of *"dist_am1"* is, relatively close to the 25^{th} percentile of *"dist_am2"*, while the 25^{th} percentile of *"dist_am3"* is a bit higher than the median of *"dist_am2"*. This could indicate that, on average, the distance of houses from *"Amenity 1"* could be significantly smaller than that from *"Amenity 2"* and equally for the distances of houses from the *"Amenity 2"* and *"Amenity 3"*.

- The numerical scale and the standard deviation of *"price"* are the largest among all numerical variables. Also, it is interesting to point out that there exists a high outlier in *"price"*, even relative to its large magnitude, as it is 21.5755574 times the standard deviation greater than the mean value. The boxplot in figure 1 further illustrates the extreme outlier in *"price"*.

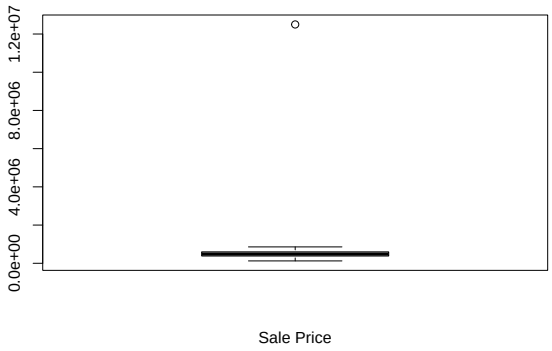

Figure 1: Boxplot of sale price.

Table 3 focuses on the aforementioned extreme outlier in *"price"*. The table reveals that the outlier (that is, the 348^{th} observation) contains extreme values in other variables as well. Specifically, that same observation is the one related to the 63 bathrooms, which we have already noted as a possible extreme value, and through further exploration we can show that its value of 12730 square feet is also extremely high. Those findings suggest that observation 348 could have probably come from a different population compared to the rest of the observations in the data. In any case, we lack enough data between this extreme observation and the rest ones, to the point that any model fitting with this outlier included would result in speculating after some range of values and would probably lead to a heavily influenced model. Therefore, we conclude that **we have enough evidence to support our decision on removing observation 348 before we move on any further**.

Table 3: The extreme observation in the variable *'price'*.

	elevation	dist_am1	dist_am2	dist_am3	bath	sqft	parking	precip	price
348	31	20662	20945	23294	63	12730	Covered	1130	12500000

Table 4: Summary statistics for the numerical variables after removing the extreme outlier.

Variable	Mean	SD	Min	25%	Median	75%	Max
elevation	30.27255	5.20367	9	27.0	30	34.0	47
dist_am1	8233.62926	2532.61067	604	6434.5	8210	9984.5	16233
dist_am2	11016.73747	2556.47368	4402	9219.5	11006	12842.0	18281
dist_am3	13072.31663	2591.98862	4922	11215.5	13179	14771.0	20263
sqft	1794.22445	297.20037	932	1588.0	1770	2002.5	2667
precip	792.48497	272.34339	-110	610.0	790	980.0	1530
price	486481.80361	142096.23500	124333	380125.0	480167	593167.0	857667

After removing the outlier, our conclusions about the variables *"elevation"*, *"dist_am1"*, *"dist_am2"*, *"dist_am3"* and *"precip"* are similar to the ones we derived earlier. However, we observe a significant drop in the maximum value of *"sqft"* along with a significant decrease in its standard deviation, which has now become relatively close to that of *"precip"*. Also, the maximum value and the standard deviation of *"price"* incurred a large drop.

The boxplots in figure 2 present graphically the already discussed differences in the magnitude and the variation between the numerical variables, by gradually removing variables from plot to plot. **A closer view using histograms in figure 3 reveals that the sample distribution of all numerical variables in the data resembles that of a sample coming from a Normal Distribution.**

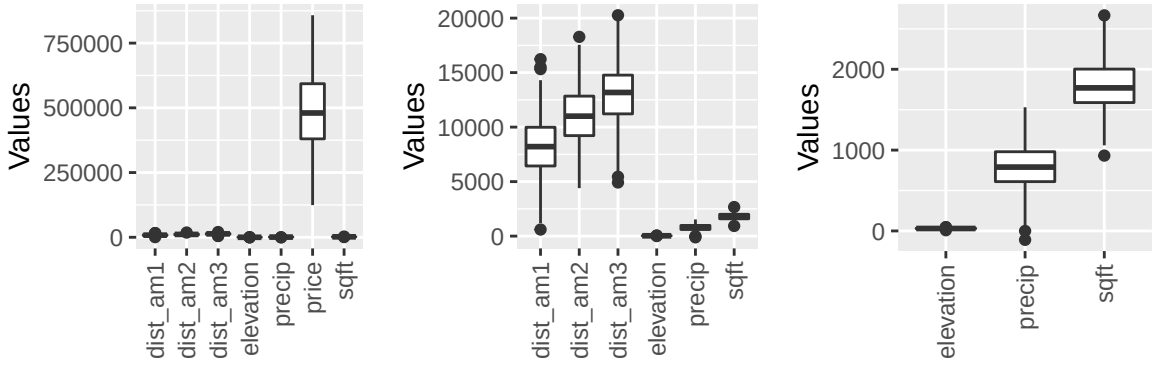

Figure 2: Boxplots on all numerical variables without the extreme outlier.

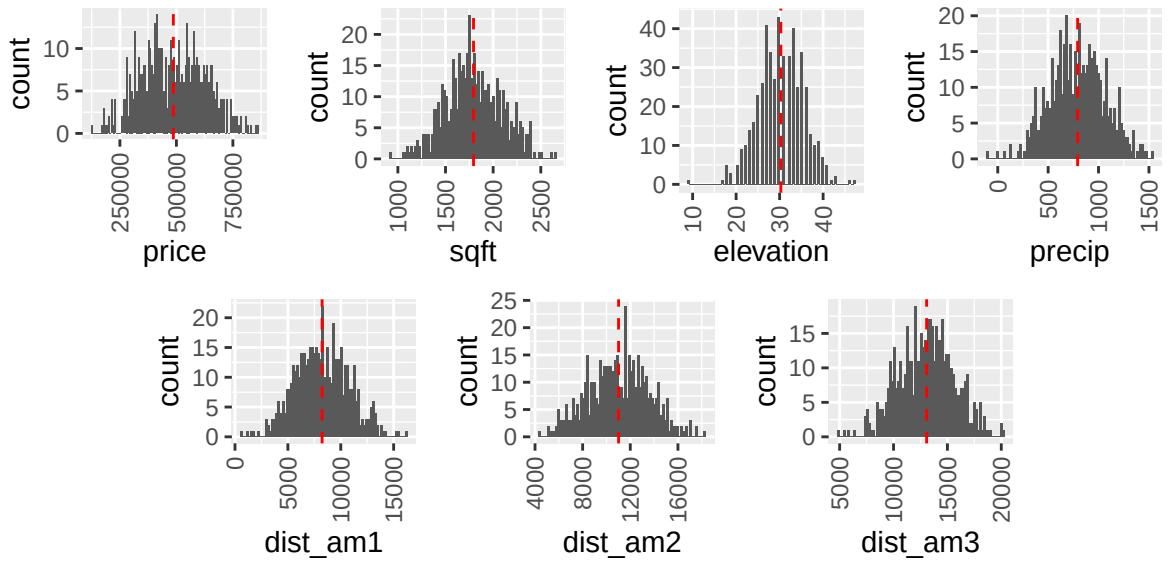

Figure 3: Histograms on all numerical variables without the extreme outlier. The red dashed line represents the sample mean of each variable.

2.2 Exploring relationships

The primary objective of this paper is to create the most accurate predictive model about *"price"* using the rest of the available variables in the data. Therefore, it is of interest to explore the relationships of *"price"* against the other variables. For this purpose, we first explore the relationships of *"price"* against all other numerical variables and, then, we focus on the categorical variables. Finally, we analyse the relationships of *"price"* against the numerical variables on different levels of the categorical variables.

2.2.1 Numerical explanatory variables

Figure 4 depicts the relationships of *"price"* against all the rest numerical variables in the form of scatterplots with a simple linear regression line superimposed. From the scatterplots, we observe a random scattering of the data points across all values with no obvious patterns suggesting that there is little association between *"price"* and any one of the other numerical variables. Additionally, the small slope of the superimposed regression lines along with the little correlation revealed in table 5 show that there is no linear relationship between any of the numerical variables and *"price"*. However, we observe a high correlation between *"dist_am3"* - *"dist_am1"* and a moderate correlation between *"dist_am1"* - *"dist_am2"* and *"dist_am2"* - *"dist_am3"* indicating that there is possibly multicollinearity between the distance variables.

A more detailed investigation on multicollinearity can be conducted by fitting linear regression models using each one of the distance variables as the response each time. The values of $R^2_{(adj)}$ reveal that we can explain 72.08% of the variability in *"dist_am3"* using the other two

distance variables. Moreover, when we treat any of the other two variables as our response and modelling it with the other two, *"dist_am3"* is almost solely responsible for observing a high $R^2_{(adj)}$ value. In fact, when we use just *"dist_am2"* to model *"dist_am1"* we can explain just 19.26% of its variability in contrast to 63.27% when we use just *"dist_am3"*.

In terms of *"price"*, the R^2 score from using just *"dist_am3"*, **although very small**, is better than the one we obtain from using *"dist_am1"* and *"dist_am2"*, together or alone. Also, the p-value for the coefficient estimate of *"dist_am3"* is 0.0502, which is very close to the 5% significance level we chose to use in this paper.

Finally we note that, before we decide on removing any of the distance variables in order to overcome multicollinearity, it is important to assess their relationship to *"price"* from their interactions with the categorical variables in the data. The exploration of interactions is presented in subsection 2.2.3.

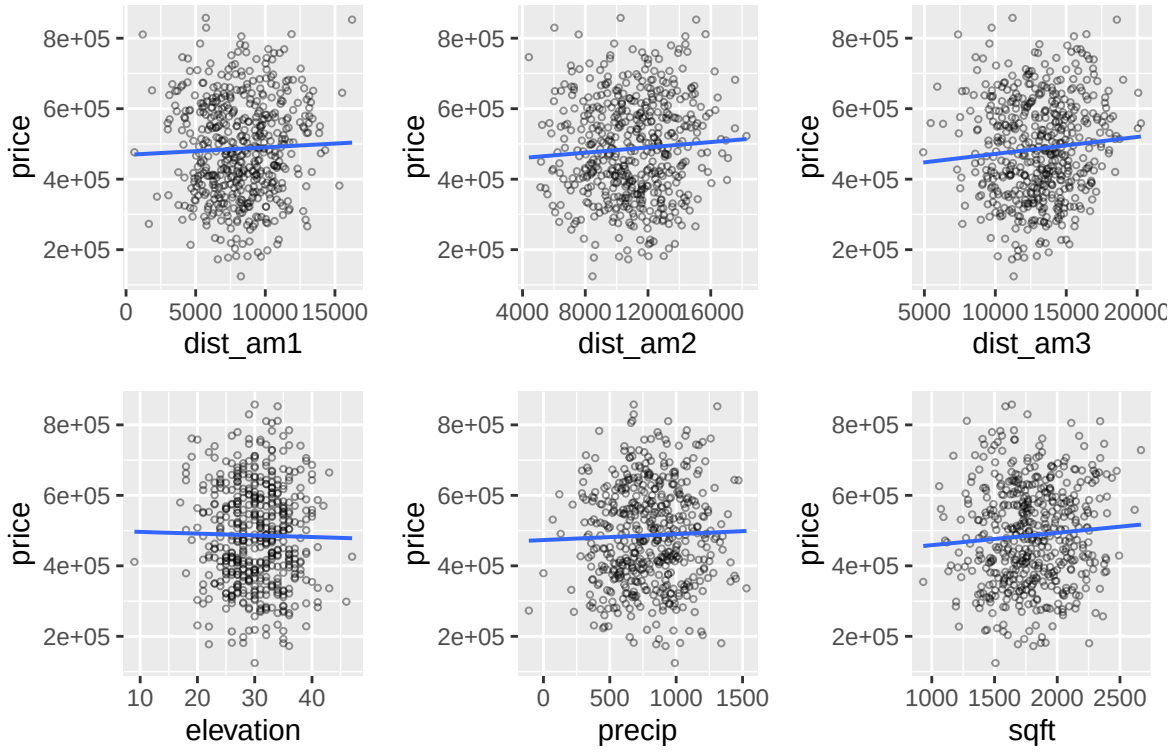

Figure 4: Scatterplots of 'price' against all the rest numerical variables. The simple linear regression line is superimposed on the plots.

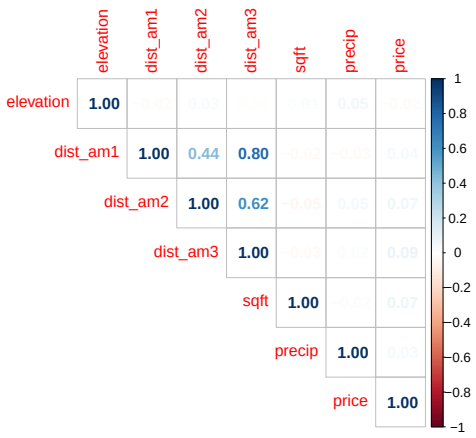

Figure 5: Correlation between all numerical variables.

2.2.2 Categorical explanatory variables

Figure 6 depicts the sample distribution of *"price"* for each level of *"bath"* or *"parking"*. Regarding *"parking"*, we observe quite a significant overlap between the boxplots with small differences between them suggesting that there is little to no difference in the sale price for houses in different *"parking"* categories. In contrast, we see that there is no overlap between the boxplots of *"price"* in different *"bath"* categories with the sale price actually increasing as the number of bathrooms increases. As it can be seen more clearly in figure 7, the seemingly normally-distributed sample of *"price"* is completely partitioned based on *"bath"* into 4 non-overlapping chunks. Also, the p-values of the estimated parameters when using a linear regression model with *"bath"* as our sole predictor are 0, while $R^2_{(adj)}$ reports that we can already explain 86.13% of the total variability in *"price"*. These findings indicate strongly that the categorical variable *"bath"* is a significant predictor with a positive relationship with *"price"*.

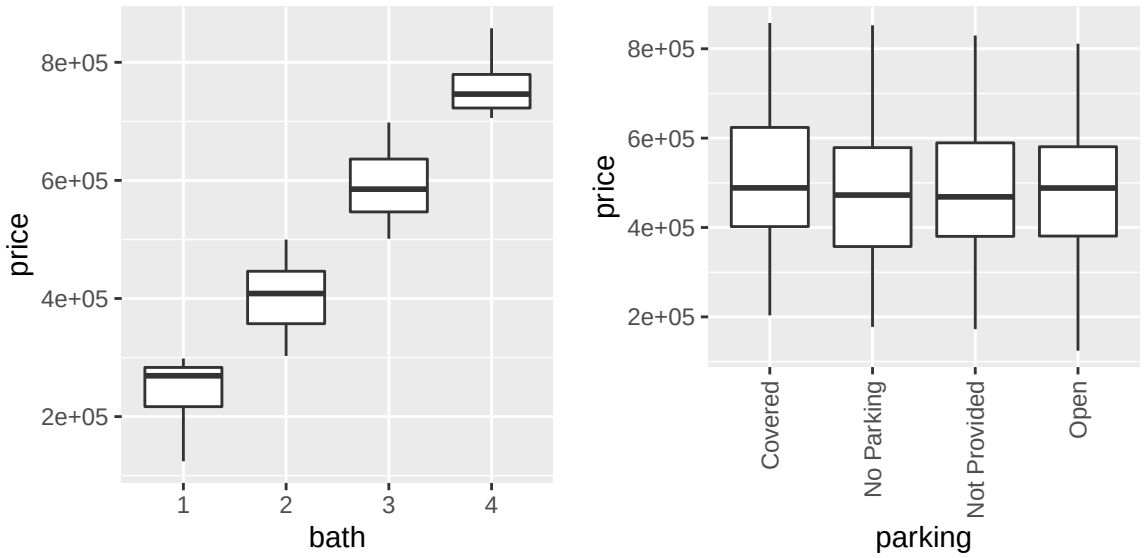

Figure 6: Boxplots of 'price' by 'bath' and by 'parking'.

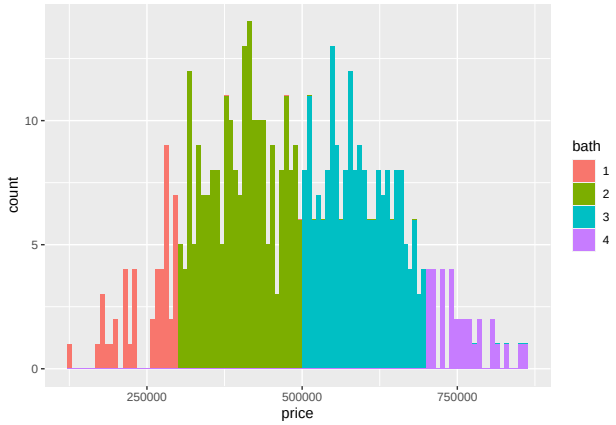

Figure 7: Histogram of 'price' coloured by 'bath'.

2.2.3 Interactions

Finally, it is important to investigate any interactions between categorical and numerical explanatory variables. Figure 9 shows the relationship of each numerical variable with *"price"*, on each *"bath"* level, in the form of scatterplots with simple regression lines superimposed. On all plots, we observe four distinct layers of data corresponding to each *"bath"* level. As we have already explained, these layers correspond to the full partitioning of the sample distribution of *"price"* into four non-overlapping chunks when considering the variable *"bath"*. It is apparent that, regardless of the *"bath"* category they belong to, the observations in the data extend fairly across the whole range of values on all numerical explanatory variables.

Apart from *"precip"*, the nature of the relationship between the rest of the explanatory

variables and *"price"* is roughly the same on each level of *"bath"* and the relationships do not seem to improve for any of the variables when we take *"bath"* into consideration (as seen from figure 9 and the high p-values). In the case of the relationship between *"precip"* and *"price"*, we observe a slightly positive trend on all levels of *"bath"* except for the houses with one bathroom, where the relationship becomes negative. Therefore, the aforementioned relationship seems to improve and become important when considered under each *"bath"* category. This fact is further supported by the p-values shown in table 5 which are either close to the chosen 5% significance level or below it.

Table 5: The regression coefficient estimates, p_values and 95% C.Is from fitting the linear model with just an interaction between 'precip' and 'bath.'

term	estimate	p_value	lower_ci	upper_ci
intercept	286198.952	0.000	244982.778	327415.126
precip	-49.749	0.061	-101.738	2.241
bath: 2	106649.457	0.000	60201.806	153097.108
bath: 3	299251.062	0.000	251913.864	346588.260
bath: 4	440604.462	0.000	365170.250	516038.675
precip:bath2	62.042	0.035	4.321	119.763
precip:bath3	56.242	0.062	-2.908	115.392
precip:bath4	84.430	0.065	-5.294	174.153

When we assess the same relationships under each category of *"parking"*, it can be shown that there is an almost complete overlap between the layers of data of the different *"parking"* levels, with a random scattering of the data points and the p-values on those relationships are fairly large. This suggests that the single regression line model should be suitable for each relationship with *"price"*, with no difference in the slopes or the intercept terms among the categories (i.e. no interaction with *"parking"*).

However, it appears that in the relationship of *"price"* and *"dist_am1"* there is a significant difference in the coefficient estimate of *dist_am1* between the *"parking"* categories "Covered"(baseline category) and "No Parking" and a significantly large p-value for the difference in the coefficient estimate between "Covered" and "Not Provided." This can be seen from both the slopes of the simple linear regression lines in figure 8 and table 6. Therefore, it is important to study formally the possible interaction of *"dist_am1"* and *"parking"*.

Continuing with the results about multicollinearity in subsection 2.2.1, the interaction between *"parking"* and *"dist_am1"* demonstrates a better $R^2_{(adj)}$ score and better p-values when modelling *"price"* than *"parking"*, *"dist_am3"* or the interaction between the two. Therefore, **we decide on dropping variable *"dist_am3"* from our models**. This decision solves the problem of multicollinearity as the remaining two distance variables have a 0.44 correlation coefficient which is rather moderate. Also, the remaining two variables can explain 72.08% of the variability in *"dist_am3"*, which suggests that *"dist_am3"* is fairly represented by the other two and we do not lose that much information in the data.

Finally, we note that other transformations we applied on the numerical explanatory variables and on *"price"* did not seem to improve the relationships between them.

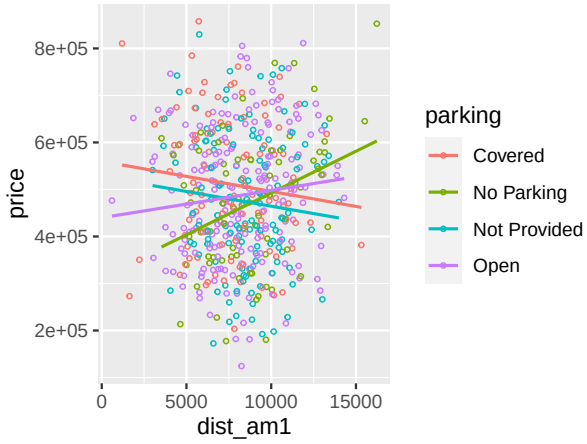

Figure 8: 'Price' against 'dist_am1' on all 'parking' levels.

Table 6: The regression coefficient estimates, p_values and 95% C.Is from fitting the linear model with just an interaction between 'dist_am1' and 'parking.'

term	estimate	p_value	lower_ci	upper_ci
intercept	559765.820	0.000	472950.234	646581.406
dist_am1	-6.387	0.243	-17.128	4.354
parking: No Parking	-244711.358	0.001	-383238.686	-106184.029
parking: Not Provided	-32718.514	0.622	-163116.737	97679.709
parking: Open	-119940.155	0.031	-228814.569	-11065.742
dist_am1:parkingNo Parking	24.133	0.003	8.292	39.974
dist_am1:parkingNot Provided	0.125	0.987	-15.307	15.557
dist_am1:parkingOpen	12.205	0.070	-1.016	25.426

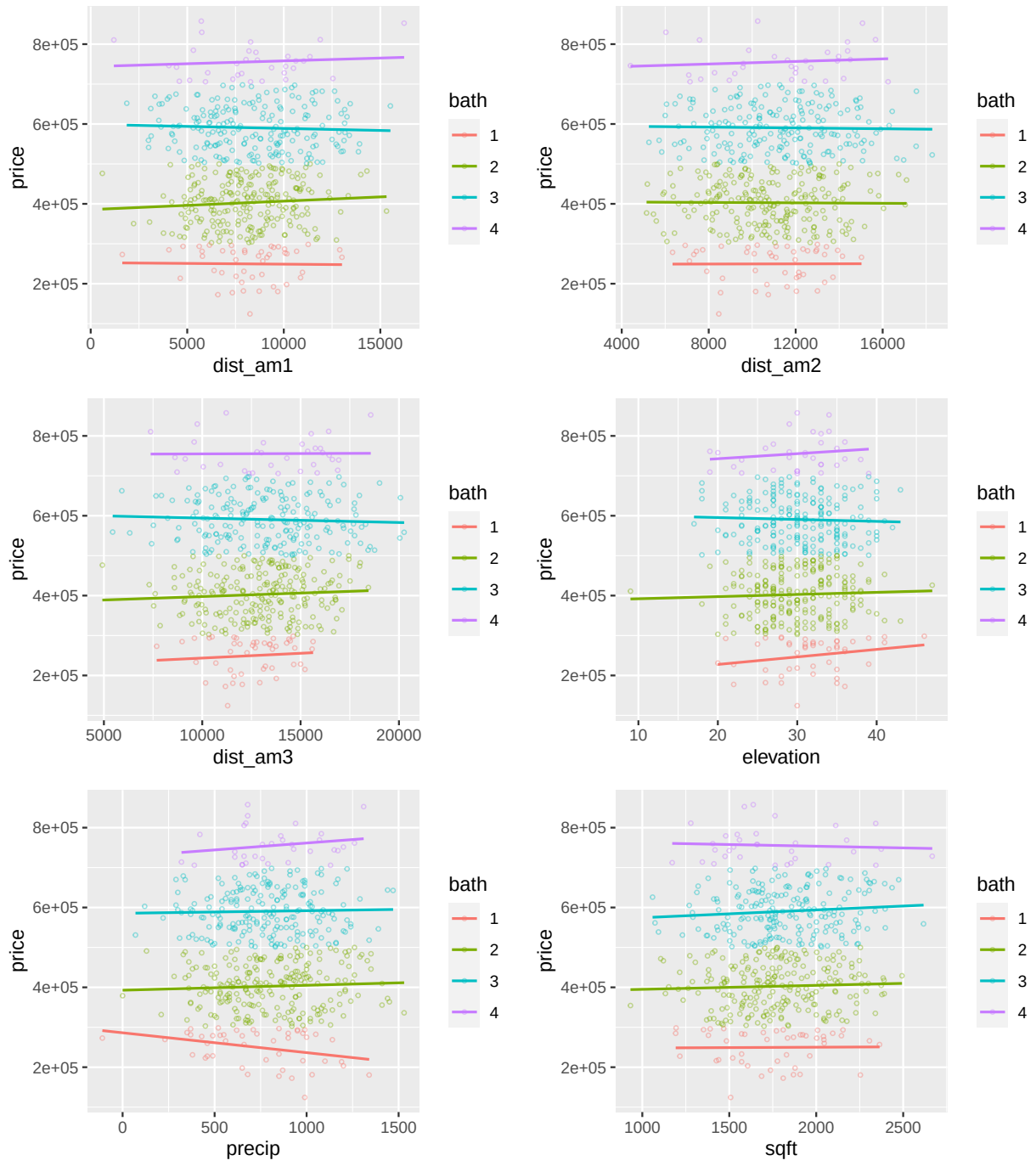

Figure 9: Scatterplots of 'price' against all the rest numerical variables and a simple linear regression line superimposed, coloured by 'bath'.

3 Model Fitting: Selecting the best possible regression model

The **primary objective** of this paper is to search for and design the best possible *predictive* regression model to accurately predict the sale price of a house based on the rest of its attributes. The main approach we chose for this problem is **best-subset selection on the full model via backward elimination**.

For this purpose, we decided on assessing our models' performance and conduct variable selection based on the average **Mean Squared (Prediction) Error** calculated using 5-fold cross-validation(C.V.) on the training data:

$$\hat{MSPE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

We achieve this by splitting our original data set into two subsets; training and test. The training set will be used for training and assessing the prediction performance(out-of-sample prediction) via 5-fold C.V. and the test set for calculating the expected prediction performance on the best model. The split we chose to use is:

- 450 observations for training and C.V.,
- 49 observations for assessment

The reason for choosing cross-validation instead of a simple training/validation/test split is to avoid the case of a bad split, especially with a small data set as ours. With K-fold C.V. we are able to use all of our data and calculate a more robust estimate of the average (expected) prediction error [book ESLII]. In practice, we have used both techniques and we have experienced more consistent results from cross-validation (even with 5, 6 or 10 folds) than from using the simple 3-subset split. That is, small changes in the training set or in the distribution of the splits resulted in almost the same final subset of predictors. Also, the choice of the number of folds to use is of great importance as it balances the bias-variance trade-off. However, a usual choice for that number is 5 or 10 folds [Breiman and Spector or Kohavi], which is the choice we made in this paper.

Finally, we will consider moving beyond the Ordinary Least Squares method by smoothing the best regression model's OLS estimates using Ridge regression on its subset of predictors and see whether we can improve the prediction performance even further. Also, we will investigate whether variable selection via Lasso regression yields a more powerful model than backward elimination.

3.1 Fitting the full model

In section 2, we discussed about the different available variables in our data and we explored numerically and graphically the relationships between them, especially with *"price"*. Based

on our findings, we select as our starting model the linear, additive model on the original variables excluding *"dist_am3"* and including interactions between *"precip"* - *"bath"* and *"dist_am1"* - *"parking"*. Therefore, we assume the true relationship is:

$$price_i = \beta_0 + \beta_1 \cdot \mathbb{I}bath|2_i + \beta_2 \cdot \mathbb{I}bath|3_i + \beta_3 \cdot \mathbb{I}bath|4_i +$$
$$\beta_4 \cdot \mathbb{I}parking|NoParking_i + \beta_5 \cdot \mathbb{I}parking|NotProvided_i + \beta_6 \cdot \mathbb{I}parking|Open_i +$$
$$\beta_7 \cdot sqft_i + \beta_8 \cdot elevation_i + \beta_9 \cdot dist\_am2_i +$$
$$\beta_{10} \cdot precip_i + \beta_{11} \cdot \mathbb{I}bath|2_i \cdot precip_i + \beta_{12} \cdot \mathbb{I}bath|3_i \cdot precip_i + \beta_{13} \cdot \mathbb{I}bath|4_i \cdot precip_i +$$
$$\beta_{14} \cdot dist\_am1_i + \beta_{15} \cdot \mathbb{I}parking|NoParking_i \cdot dist\_am1_i + \beta_{16} \cdot \mathbb{I}parking|NotProvided_i \cdot dist\_am1_i +$$
$$\beta_{17} \cdot \mathbb{I}parking|Open_i \cdot dist\_am1_i$$
$$+ \epsilon_i, \quad \epsilon_i \overset{indep.}{\sim} N(0, \sigma^2), \quad i = 1, ..., 450$$

, where the indicator variables(starting with $\mathbb{I}$) are equal to 1 when the i^{th} observation's corresponding categorical variable takes the relevant value(separated from the variable name with "|") and 0 otherwise. The baseline categories are **1** for *"bath"* and **Covered** for *"parking"*.

Table 7 shows the OLS estimates of the full model's regression coefficients along with their p-value and 95% Confidence Interval(C.I.), when fitted on the whole training set. The reported p-values suggest that we have enough statistical evidence that the coefficients for the categories of *"bath"* are significant, when considering all the predictors we used in the model and choosing a significance level of 5%. Furthermore, the absence of overlap between the 95% C.Is of the *"bath"* categories further underlines the importance of this predictor.

Additionally, almost all of the reported p-values related to the interaction between *"precip"* and *"bath"* are either below or close to the chosen 5% threshold, suggesting that this interaction seems important for predicting *"price"*, considering all the other predictors we used in the model. All the rest estimates have a p-value that is greater than the chosen significance level or, equivalently, they have a 95% C.I. that includes the value of 0 as a plausible value for the coefficient in the population's regression model.

Table 7: The estimated regression coefficients along with the corresponding p-values and 95% Confidence Intervals from fitting the full model on the training set.

term	estimate	p_value	lower_ci	upper_ci
intercept	271732.000	0.000	200015.888	343448.112
bath: 2	93939.818	0.000	44635.580	143244.055
bath: 3	297146.185	0.000	247907.231	346385.138
bath: 4	428734.435	0.000	350775.282	506693.588
parking: No Parking	-14760.981	0.602	-70402.966	40881.004
parking: Not Provided	-6279.779	0.812	-58059.512	45499.953
parking: Open	-10592.884	0.639	-54981.915	33796.147

term	estimate	p_value	lower_ci	upper_ci
precip	-53.081	0.051	-106.306	0.144
dist_am1	1.284	0.564	-3.085	5.653
dist_am2	-0.889	0.430	-3.100	1.323
sqft	10.548	0.216	-6.183	27.279
elevation	315.794	0.525	-659.242	1290.831
bath: 2:precip	73.954	0.016	13.829	134.078
bath: 3:precip	54.373	0.079	-6.359	115.106
bath: 4:precip	92.035	0.050	0.004	184.067
parking: No Parking:dist_am1	-0.219	0.946	-6.575	6.137
parking: Not Provided:dist_am1	-0.259	0.934	-6.364	5.847
parking: Open:dist_am1	-0.232	0.932	-5.566	5.101

Lastly, we can calculate the average Mean Squared Prediction Error of the full model via 5-fold cross-validation, which is equal to $\hat{MSPE} = 2.9577973 \times 10^9$. This metric will be used for performing variable selection as described in the following subsection.

3.2 Variable Selection via Best-Subset Selection

Although the full model is a good starting point to consider all predictors that could be possibly related to the response variable, it is almost certain that we have included excessive, unnecessary complexity that leads to overfitting. Usually, such a model is very flexible, lacks stability (high variance) and captures sample-specific features that are not generalisable to other data from the same population (noise). In other words, this kind of models excel in estimating on their training data set, but fail to demonstrate a good performance in out-of-sample prediction.

Since our goal is to construct the best predictive model, we wish to minimise the average $\hat{MSPE}$ we obtain from cross-validation as much as possible and the method we chose to achieve this is **best-subset selection via backward elimination on the full model**. This iterative method starts by removing one variable at a time from the full model and calculating the average $\hat{MSPE}$ after each removal by performing 5-fold C.V. on the training set, until it has removed each variable once. Then, the method identifies the variable that resulted in the best(lowest) $\hat{MSPE}$ value after its removal and updates the model such that it does not contain this variable any longer. The method continues by seeking the next best variable to remove on the updated model. This process is repeated until either there is only one independent variable left or no more improvement can be achieved.

In simple words, we decrease the overall complexity of our model by gradually removing unnecessary variables, while we continuously monitor the improvement of the $\hat{MSPE}$ averaged across the 5 folds of our data, until we reach the best possible value. It is easy to realise that this method belongs to the family of greedy algorithms as it consists of a series of consecutive, independent and non-reversible steps at which the optimal solution is chosen each time.

Table 8: Results from performing variable selection via backward elimination on the full model. The left column contains the removed variables in the order of removal, while the right one shows the average $\widehat{MSPE}$ from 5-fold C.V. after the removal of the variable on the left. The full model with its average $\widehat{MSPE}$ is displayed in the first row for reference.

Removed variable	cross-validated $\widehat{MSPE}$ after removal
full model	2957797295.27158
dist_am1*parking	2917494111.79514
sqft	2887082216.82688
dist_am1	2871441045.5095
parking	2854039381.82848
dist_am2	2846166919.01058
elevation	2838980183.56733

Table 8 shows the process of applying the aforementioned variable selection method on our full model. Apparently, the method managed to reduce the average $\widehat{MSPE}$ down to 2.8389802×10^9 from the initial, full model's value of 2.9577973×10^9. As a result, we now update our assumption about the true relationship between *"price"* and our predictors in the population such that:

$$price_i = \beta_0 + \beta_1 \cdot \mathbb{1}bath|2_i + \beta_2 \cdot \mathbb{1}bath|3_i + \beta_3 \cdot \mathbb{1}bath|4_i +$$

$$\beta_4 \cdot precip_i + \beta_5 \cdot \mathbb{1}bath|2_i \cdot precip_i + \beta_6 \cdot \mathbb{1}bath|3_i \cdot precip_i + \beta_7 \cdot \mathbb{1}bath|4_i \cdot precip_i$$

$$+ \epsilon_i, \quad \epsilon_i \overset{indep.}{\sim} N(0, \sigma^2), \quad i = 1, ..., 450$$

(4.2.1)

The results from fitting model 4.2.1 on the training data set are shown in table 9. The reported p-values indicate that, accounting for all of the variables we used in our model, the categorical variable *"bath"* and the interaction between precipitation and *"bath"* seem to be significant predictors in terms of the sale price of a house. More specifically, our fitted model estimates that:

- **for houses with 0 precipitation**, the expected sale price of a house with just one bathroom is, on average, 288833.246. This base price increases, on average, by 98592.790, 298359.921 or 431877.270 for 2, 3 or 4 bathrooms, respectively. This gradual increase in the price as we increase the number of bathrooms combined with the non-overlapping 95% C.I. for the OLS estimates of the price differences between the *"bath"* categories show a significant positive relationship between the two variables and agree with our findings in section 2.

- **for every 1 unit increase in precipitation**, the expected sale price of a house decreases, on average, by 50.454 for houses with 1 bathroom. However, the expected price seems to increase, on average, by 18.277, 3.206 or 39.697 per each unit increase in precipitation when the houses have 2, 3 or 4 bathrooms, respectively. The p-values of the estimates suggest that there is a statistically significant difference in the slopes of *"precip"* between the categories of 1 and 2 bathrooms. Although we do not have enough evidence of a significant difference in the slopes between the rest of the categories, the reported p-values are very small(<10%) and some are close to 5%.

It is, also, worth mentioning that $R^2_{(adj)}$, which takes the complexity of the used model into account, improves from the full model's value of 0.8612811 to 0.8622949, suggesting that the predictors in model 4.2.1 are able to explain 86.23% of the variation in the sale price.

Table 9: The final model fitted on the training set after performing best-subset selection.

term	estimate	p_value	lower_ci	upper_ci
intercept	288833.246	0.000	246468.064	331198.427
bath: 2	98592.790	0.000	49797.811	147387.770
bath: 3	298359.921	0.000	249575.078	347144.763
bath: 4	431877.270	0.000	355067.249	508687.291
precip	-50.454	0.060	-103.126	2.218
bath: 2:precip	68.731	0.024	9.156	128.307
bath: 3:precip	53.660	0.080	-6.504	113.823
bath: 4:precip	90.151	0.051	-0.435	180.738

However, the interpretation of model 4.2.1 and its application in predicting sale prices are meaningless if the assumptions we made prior to fitting our model are violated. In reality, there is no formal way of checking the validity of those assumptions. However, the plots of the residuals in figure 10 are commonly used in order to check the assumptions of our multiple linear regression model 4.2.1, as described in the book *"Linear Models with R"* by J.J. Faraway[book].

The "Residuals vs. Fitted" plot above shows 4 distinct vertical groups of data points. This behaviour is expected as our model relies heavily on the 4 disjoint *"bath"* categories to calculate the fitted price, while precipitation plays a smaller role in explaining the remaining variability. Also, the plot depicts a random and even scattering of the data points above and below the horizontal 0 line with no obvious patterns suggesting that the assumption that the errors' mean is 0 holds and that there is no significant non-random structure left unexplained in our data. Plots of residuals against the excluded predictors further support this claim, as they show a random scattering of the data points without any obvious patterns.

Finally, we observe a fairly equal scattering of the points about the 0 line across the range of the fitted values which indicates that the assumption of constant variance(homoscedasticity) is valid. One could argue that there is smaller variability observed for the residuals of the

first and last column in the plot. However, it is important to consider that the data set contains remarkably fewer observations for the categories of 1 or 4 bathrooms, which might be the reason behind this seemingly different variability.

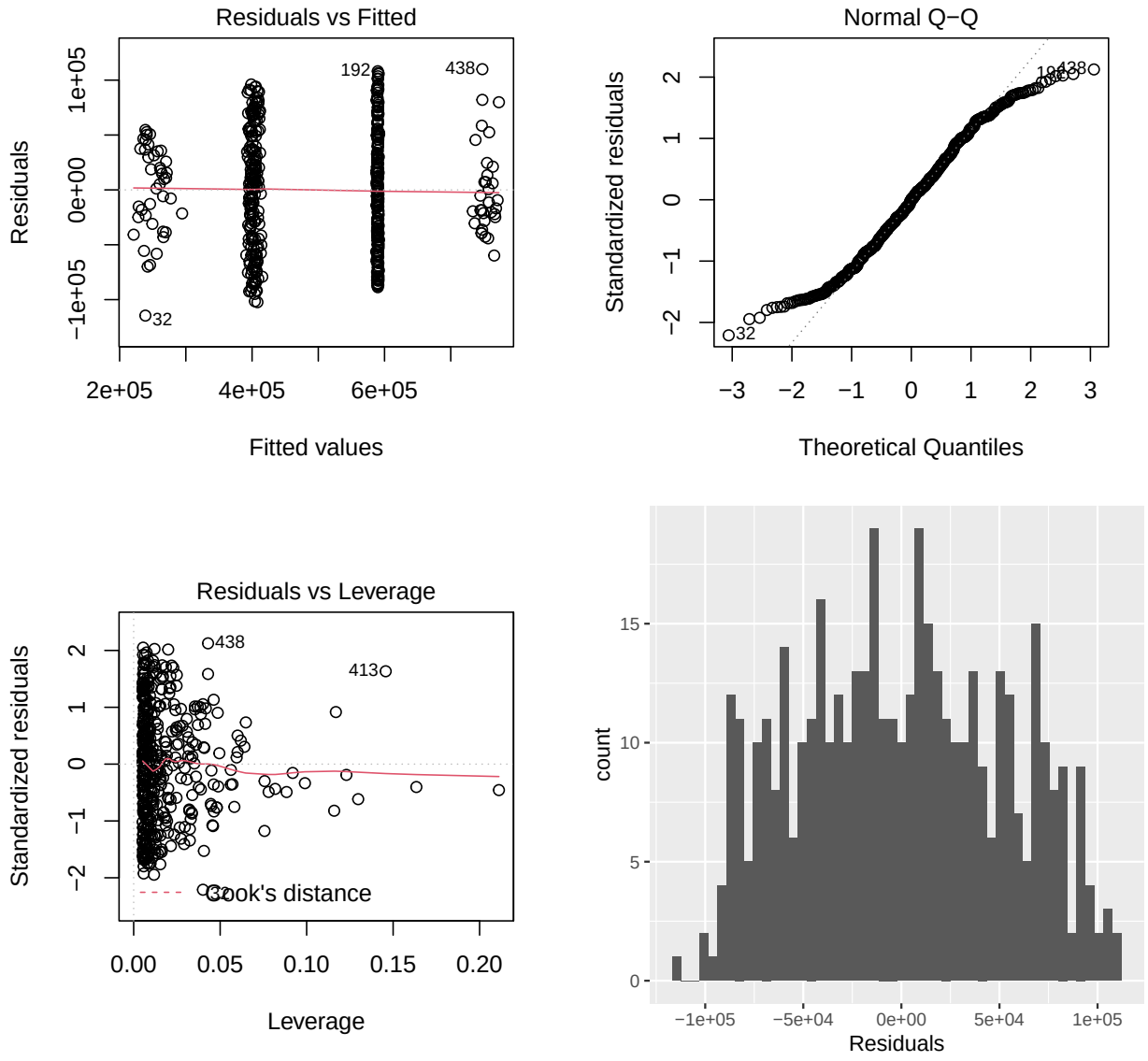

Figure 10: Plots of the residuals from fitting model 4.2.1 on the training set.

Using the "Normal Q-Q" plot, we can check the assumption of normally distributed errors. The plot exhibits a light and a heavy tail and it does not have the *ideal* shape. However, given the fact that we are modelling real-life data, it is reasonable to think that the points follow the dashed line adequately and the assumption holds. The same can be inferred from the supplemental histogram of residuals, which resembles the bell shape of the Normal distribution but not perfectly.

Lastly, we wish to look for any influential observations that we might need to discard as they could incorrectly draw the direction of the relationship towards them. Although the rest of the plots report consistently the observations "32" and "438" as possible outliers, the plot of residuals against leverage shows no observations within the contour lines of Cook's distance. Therefore, we choose not to remove any additional observations.

3.3 Regularised regression with Ridge and Lasso

As mentioned in book [ESLII], "Best-Subset Selection" can be considered as a *discrete* technique, in the sense that complexity reduction is achieved by choosing to either include a variable in the model or exclude it from it.

However, there are methods that provide a *continuous* and more flexible complexity reduction. The estimated parameter values from fitting the model using the least squares approach can be seen as the weighted influence of each predictor on the predicted outcome. As a result, there is opportunity to decrease the absolute value of those parameters in order to reduce the influence of some predictors(*smoothing*), thus resulting in a *continuous* method of complexity reduction without necessarily removing the included predictors completely. This idea leads to "Regularised Regression" and both **Ridge** and **Lasso** belong to this family of regression techniques.

In the Ordinary Least Squares(OLS) approach, which is the typical linear regression technique and the one we used so far in this paper, we want to estimate the regression parameters in such a way that the resulting regression line has in total the closest distance possible from the data points, given the selected predictors. This implies the use of a loss function that calculates the total discrepancy between the estimated line and the data points, which allows for finding the optimal solution through loss minimisation. Assuming that the parametric form of the chosen model is correct(true), the OLS estimates are unbiased[ESLII]. This means that, if we could fit the same model on a sufficiently large amount of different data sets from the same population and average the results, the average model would be asymptotically equal to the true relationship in the population. In fact, the Gauss-Markov Theorem states that, among all linear *unbiased* estimators, the OLS estimator has the smallest variance and, therefore, the smallest Mean Squared Error.

Although the OLS estimates can be considered as unbiased when the correct parametric model is selected, in reality there can be a biased estimator with better prediction performance[ESLII]. Since the prediction error is a mixture of the bias and the variance of the model, we ultimately want to achieve the best balance between the two.

In Lasso and Ridge, we exploit the loss function by integrating a penalty term for each parameter estimate that is related to its absolute value. More specifically, parameter estimates that are higher in absolute value result in more loss because of their penalty term, thus leading to the need of an equilibrium between avoiding penalisation and achieving the closest fit. Consequently, the use of penalties constrains the magnitude of the estimates and biases them towards 0, making them less variable. Essentially, we try to strike the correct balance of bias and variance by introducing a small amount of bias in exchange to much less variance.

The loss function of the regularised regression has the following form:

$$L(\vec{\theta}) = \sum_{i=1}^{n} [y_i - (\theta_0 + \vec{\theta} \cdot \vec{x_i})]^2 + \lambda \cdot r(\vec{\theta}), \qquad \vec{\theta} = (\theta_1, \theta_2, ..., \theta_p), \quad \theta_0 : intercept$$

, where the sum is the Residual Sum of Squares, $r(\vec{\theta})$ is the penalty function of the regression parameters and λ is a hyperparameter called *regularisation parameter* that adjusts the weight of the penalty and, therefore, the amount of constraint imposed on the parameter estimates.

The difference between the two regularised regression methods is in their penalty function:

- Ridge regression uses the squared $L_2 - norm$, $r(\vec{\theta}) = \sum_{j=1}^{p} \theta_j^2$, of the parameters' vector, while
- Lasso uses the $L_1 - norm$, $r(\vec{\theta}) = \sum_{j=1}^{p} |\theta_j|$.

This small difference results in a different behaviour. As λ increases, Ridge regression will start applying a significant penalty on the parameters, forcing them towards 0. However, as mentioned in [An Introduction to Statistical Learning], unless λ is extremely large ($\lambda \to \infty$), Ridge will never actually remove any of the predictors in the model(compute a 0 coefficient estimate). On the other hand, Lasso is able to draw some of the coefficients exactly to 0 when λ is sufficiently large and, therefore, it can be considered as a biased regression method that can perform both coefficient penalisation and continuous variable selection. For that reason, we can see Ridge regression as a smoothing technique that decreases model variance, while Lasso can be used for performing flexible variable selection.

According to the above perspective, we wish to apply Ridge regression on the best subset model 4.2.1 to assess whether a reduction in variance can lead to more improvement, meaning an even lower $\hat{MSPE}$. Additionally, we would like to try Lasso on the full model to check if the resulting flexible variable selection demonstrates greater performance than the discrete best-subset method.

However, before we apply either Ridge or Lasso, it is important to tune the hyperparameter λ in both cases. In order to ensure the fairness of the results, we perform 5-fold CV on the same 5 folds we used in best-subset selection. Similarly to the approach presented in [An Introduction to Statistical Learning], we calculate the cross-validated average $\hat{MSPE}$ on a grid of $\lambda = 10^i$, where i takes 40000 equidistant values from -3 to 16. Then, we select the value of λ that yielded the smallest $\hat{MSPE}$.

Regarding Ridge on the best-subset model 4.2.1, we obtain the smallest $\hat{MSPE} = 2.8388517 \times 10^9$ when $\lambda = 60.8750347$, which is indeed better than the $\hat{MSPE}$ obtained from simply performing best-subset selection. The left plot in figure 11 shows the change in the estimated $\hat{MSPE}$ as we gradually depart from the OLS results(that is, when $\lambda \to 0$) by increasing λ. Although there is a dip shown in the plot, this dip is not prominent and the relative difference in the cross-validated $\hat{MSPE}$ between the best-subset model before and after Ridge is only $\frac{\hat{MSPE}_{bestSubset} - \hat{MSPE}_{Ridge}}{\hat{MSPE}_{bestSubset}} \times 100 = 0.0045257$ %. This suggests that the Ridge regression has only a small effect on model 4.2.1 and that the least squares solution seems already adequate [introduction to statistical learning].

In terms of applying Lasso on the full model, $\lambda = 196.636748$ results in the lowest possible $\hat{MSPE}$ of 2.9372985×10^9. However, in the case of the Lasso it is common to use the *"one standard error"* rule. According to this rule, we select the least complex model possible(the largest value of λ) that yields an $\hat{MSPE}$ value which is no larger than one estimated standard error away of the minimum. In our case, the largest possible λ is 2370.3841852 which results in $\hat{MSPE} = 3.0217057 \times 10^9$.

In our research, we assess both cases of Lasso regression. Table 10 shows the coefficient estimates from performing Lasso on the full model with both choices of λ. Since we are interested in using Lasso mainly as a variable selection method, we will use its results only to exclude variables from the full model. Therefore, the minimum λ results indicate merely the removal of the interaction between *"dist_am1"* and *"parking"*, while the ones we obtain from using the 1se λ suggest the additional removals of *"elevation"*, *"dist_am2"* and *"dist_am1"*.

Assessing the full model when we exclude the variables suggested by Lasso in both cases, we have that the minimum λ leads to a value of $\hat{MSPE} = 2.9174941 \times 10^9$ and 1se λ results in 2.8847468×10^9. We observe that the cross-validated $\hat{MSPE}$ is worse than the one obtained from the best-subset model and we point out that applying Ridge on those models does not improve these results.

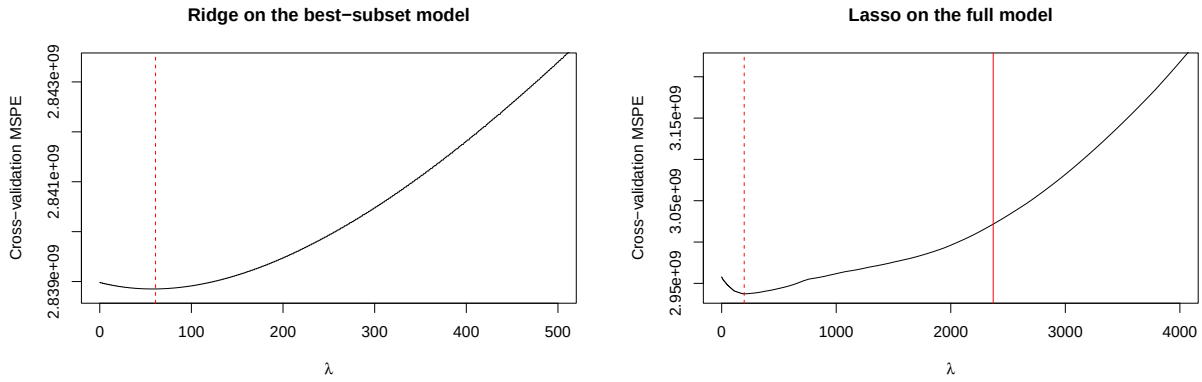

Figure 11: The change in the estimated MSPE as we increase the regularisation parameter, λ. The dashed vertical line corresponds to the chosen λ for each regularised regression method, that yields the best CV MSPE. The solid vertical line corresponds to the largest value of λ such that the resulted $\hat{MSPE}$ is within 1 estimated standard error of the minimum $\hat{MSPE}$.

Table 10: The parameter estimates from performing Lasso regression on the full model using the training data.

	Estimates with min λ	Estimates with 1se λ
(Intercept)	2.691959e+05	275619.600961
precip	-4.259010e+01	0.000000
bath2	9.923547e+04	102862.619492

	Estimates with min λ	Estimates with 1se λ
bath3	3.020949e+05	302590.359601
bath4	4.337946e+05	432091.517242
dist_am1	9.577968e-01	0.000000
parkingNo Parking	-1.475203e+04	-2854.606390
parkingNot Provided	-6.627268e+03	0.000000
parkingOpen	-1.075973e+04	0.000000
dist_am2	-7.713939e-01	0.000000
sqft	1.021823e+01	4.962045
elevation	2.482101e+02	0.000000
precip:bath2	6.301327e+01	15.599924
precip:bath3	4.379674e+01	0.000000
precip:bath4	8.088890e+01	34.884875
dist_am1:parkingNo Parking	0.000000e+00	0.000000
dist_am1:parkingNot Provided	0.000000e+00	0.000000
dist_am1:parkingOpen	0.000000e+00	0.000000

TODO: Change lambda grid from 20000 to 40000

4 Conclusions and Further Work

Our goal throughout this paper was to find the best possible regression model to predict the sale price of houses using the rest of their available attributes. First, we explored our data where we detected and removed an extreme outlier and we uncovered the significant importance of the number of bathrooms a house has on its final sale price. Then, we assessed the relationships between our variables that revealed a possibly significant interaction between the number of bathrooms and the precipitation of a house and forced us to deal with the problem of multicollinearity. Finally, we formalised the results of our analysis by constructing statistical models and performing variable selection and regularised regression to search for the best predictive model.

According to our findings, we conclude that the best regression model is the best-subset model 4.2.1 that consists of the interaction between the precipitation of the house and the amount of bathrooms it has. Specifically, we found out that as the number of bathrooms increases so does, on average, the sale price and, in fact, the house moves up to a higher price category. As for houses with the same amount of bathrooms, our model predicts that, on average, the price drops with precipitation for houses with 1 bathroom, while it increases when the number of bathrooms is 2 or greater. A more detailed interpretation of the final fitted model has already been provided in subsection 3.2. Finally, we report that the expected prediction performance of the best model in terms of its MSPE is 2.7621819×10^9.

Further work: 1. More variables 2. Go beyond regression models:

- PCA on distance variables to counter multicolinearity and regression on the full model with the PC instead of distances
- k-NN and calculate average neighbours(but first, feature scaling and possibly weighted distance).

~~ TODO ~~

5 References