Dissertation

Emmanouil Mertzianis

10/1/2021

Contents

1 Introduction **2**

1.1 The real estate market and the importance of predictive and descriptive models 2

1.2 The purpose of the paper 2

1.3 The structure of the paper 2

2 Literature **2**

3 Exploratory Data Analysis **2**

3.1 Exploring variables individually 2

3.2 Exploring relationships 6

4 Model Fitting: Selecting the best possible regression model **11**

4.1 Fitting the full model 12

4.2 Variable Selection 13

4.3 Regularised regression with Ridge and Lasso 15

5 Conclusions **15**

6 Further Work **16**

7 References **16**

1 Introduction

1.1 The real estate market and the importance of predictive and descriptive models

1.2 The purpose of the paper

1.3 The structure of the paper

~~ TODO ~~

2 Literature

~~ TODO ~~

3 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.

3.1 Exploring variables individually

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

Variable	# missing	Unique lvls	Counts
bath	0	5	1: 46, 2: 222, 3: 198, 4: 33, 63: 1
parking	0	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	# missing	Mean	SD	Min	25%	Median	75%	Max
elevation	0	30.274	5.198555e+00	9	27.00	30.0	34.00	47
dist_am1	0	8258.486	2.590404e+03	604	6439.75	8219.0	10011.25	20662
dist_am2	0	11036.594	2.592219e+03	4402	9229.25	11015.0	12848.50	20945
dist_am3	0	13092.760	2.629431e+03	4922	11215.75	13188.0	14775.75	23294
sqft	0	1816.096	5.721306e+02	932	1588.50	1770.5	2003.00	12730
precip	0	793.160	2.724887e+02	-110	610.00	790.0	980.00	1530
price	0	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"*.

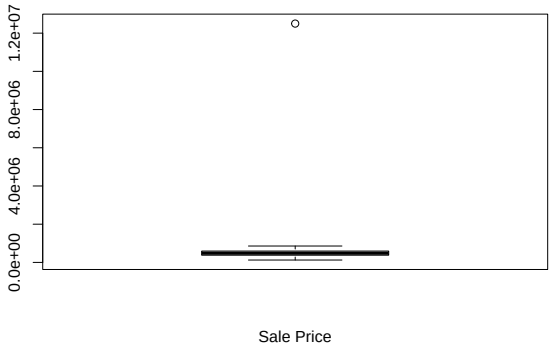


Figure 1: Boxplot of sale price.

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	# missing	Mean	SD	Min	25%	Median	75%	Max
elevation	0	30.27255	5.20367	9	27.0	30	34.0	47
dist_am1	0	8233.62926	2532.61067	604	6434.5	8210	9984.5	16233
dist_am2	0	11016.73747	2556.47368	4402	9219.5	11006	12842.0	18281
dist_am3	0	13072.31663	2591.98862	4922	11215.5	13179	14771.0	20263
sqft	0	1794.22445	297.20037	932	1588.0	1770	2002.5	2667
precip	0	792.48497	272.34339	-110	610.0	790	980.0	1530
price	0	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. Interestingly, the boxplots suggest that the sample distributions of all numerical variables are fairly symmetrical as we observe the median to lie almost at the middle of the IQR box and roughly equal tails at the top and the bottom. This observation is backed by the computed numerical statistics, where the median is reported to be quite close to the mean value for every numerical variable. **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.**

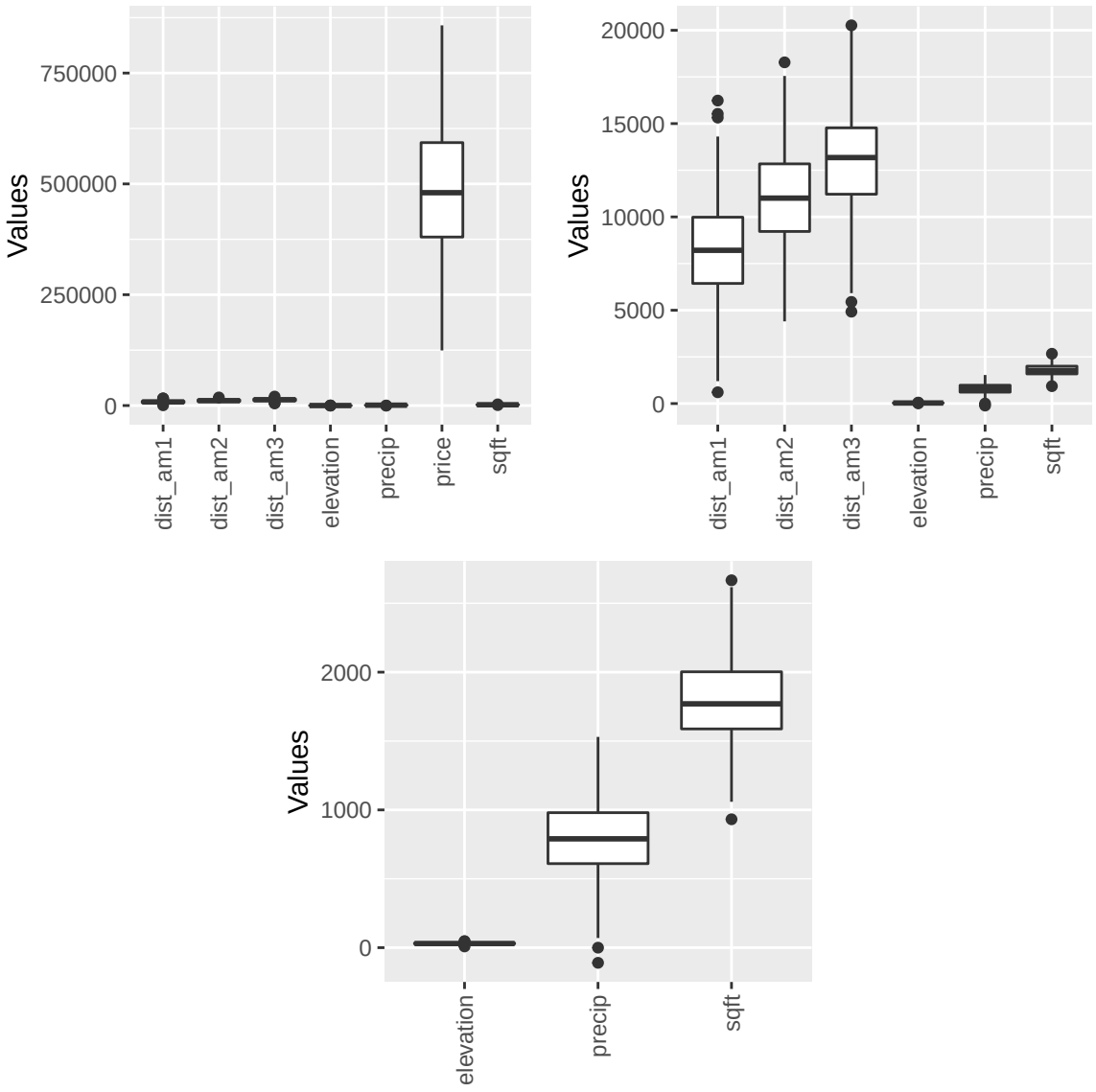


Figure 2: Boxplots on all numerical variables without the extreme outlier.

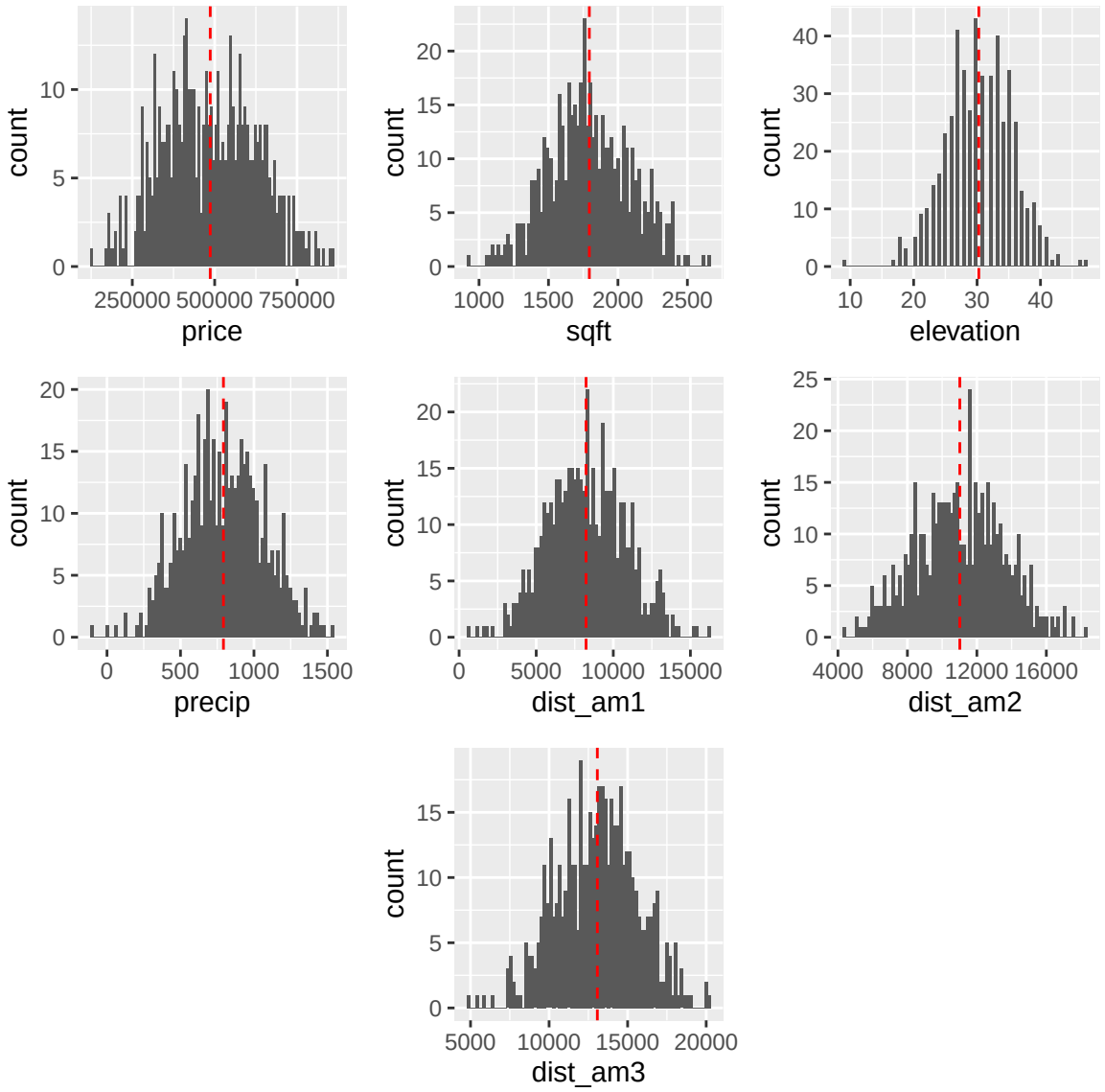


Figure 3: Histograms on all numerical variables without the extreme outlier. The red dashed line represents the mean value of the variable's values.

3.2 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.

3.2.1 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.

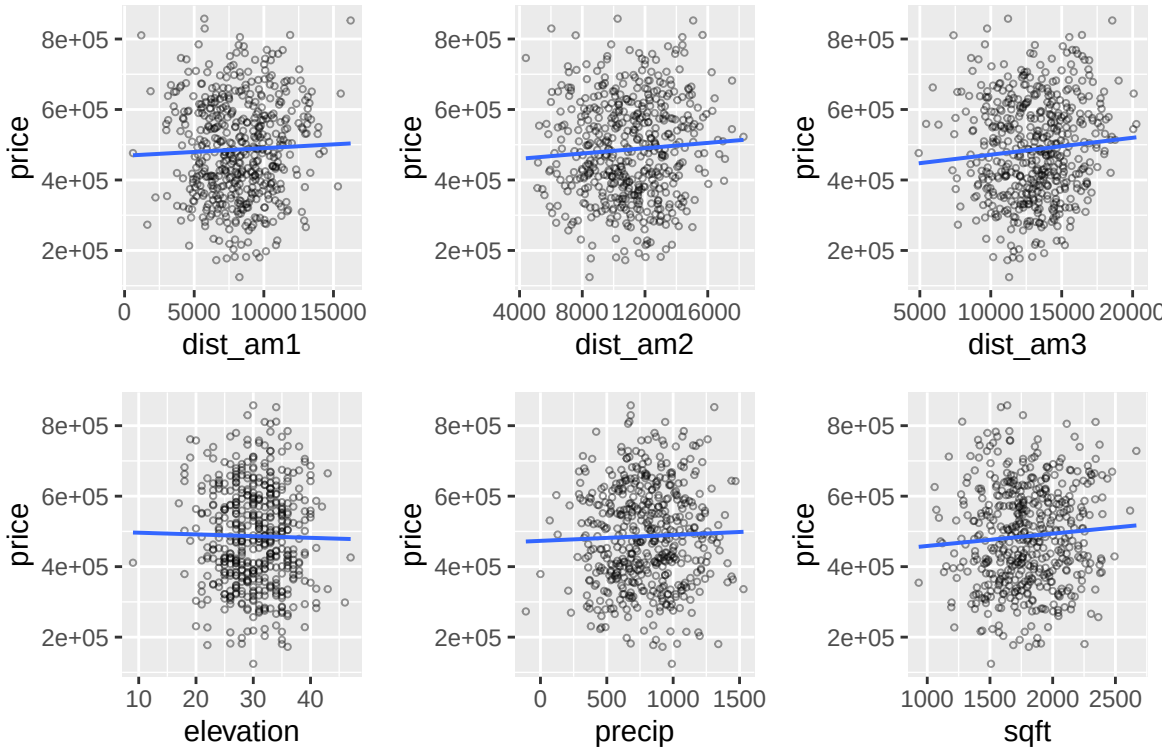


Figure 4: Scatterplots of 'price' against all the rest numerical variables. The simple linear regression line is superimposed on the plots.

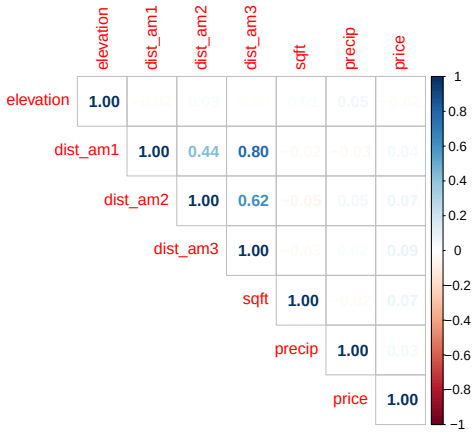


Figure 5: Correlation between all numerical variables.

3.2.2 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. These findings indicate strongly that the categorical variable *"bath"* is a significant predictor with a positive relationship with *"price"*.

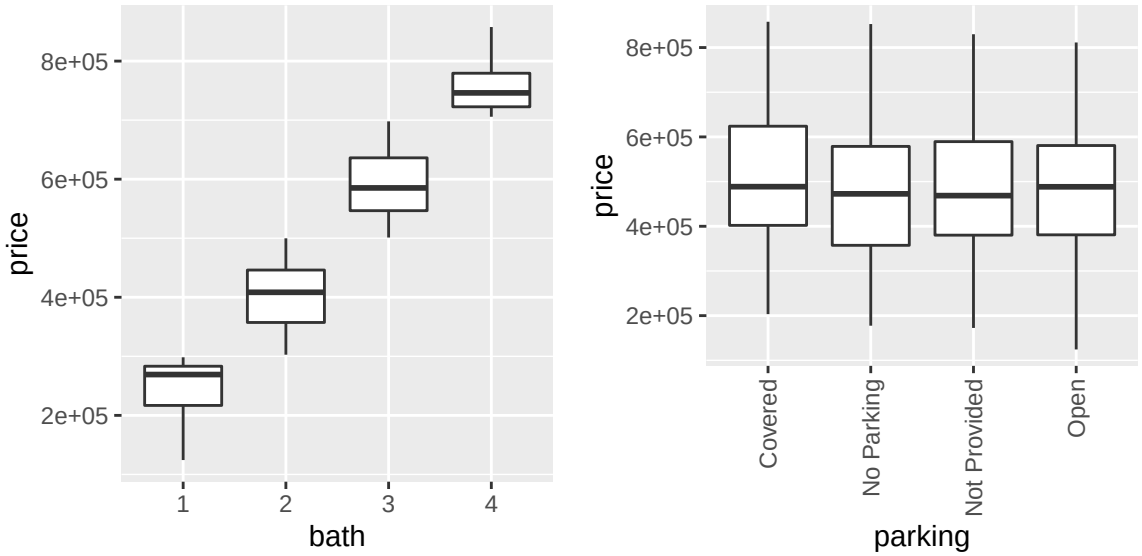


Figure 6: Boxplots of 'price' by 'bath' and by 'parking'.

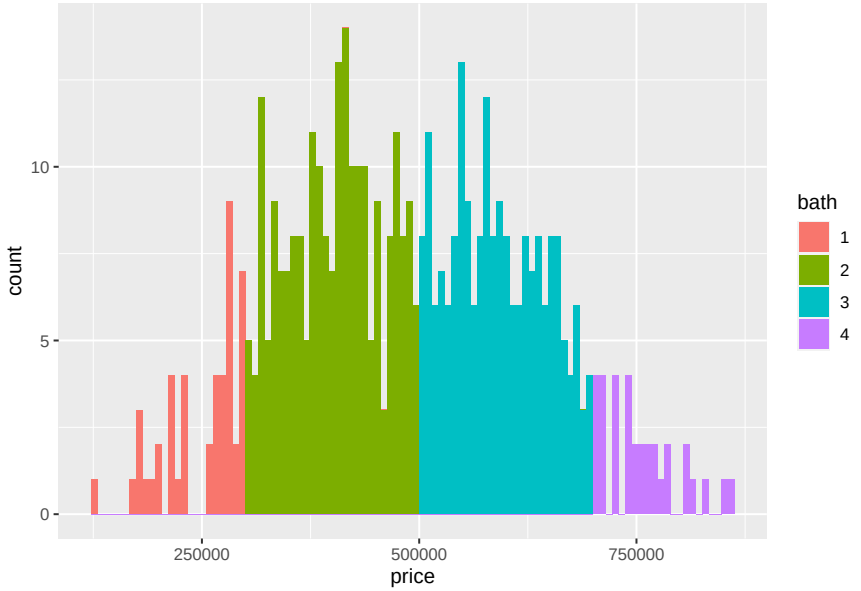


Figure 7: Histogram of 'price' coloured by 'bath'.

3.2.3 Interactions

Finally, it is important to investigate any interactions between categorical and numerical explanatory variables. Figure 8 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.

Excluding *"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 the scatterplots and the high p-values). In the case of the relationship between *"precip"* and *"price"*, we observe a slightly positive trend on all levels except for the houses with one bathroom, where the relationship becomes negative. Therefore, the aforementioned relationship seems to improve and become interesting 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 slope of the simple linear regression lines among the *"parking"* categories, except for the slopes in the categories "Covered" and "Not Provided" where they are roughly equal. This can be seen from both the figure 9 and the table 6. Therefore, it is important to study the possible interaction of *"dist_am1"* and *"parking"*.

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.

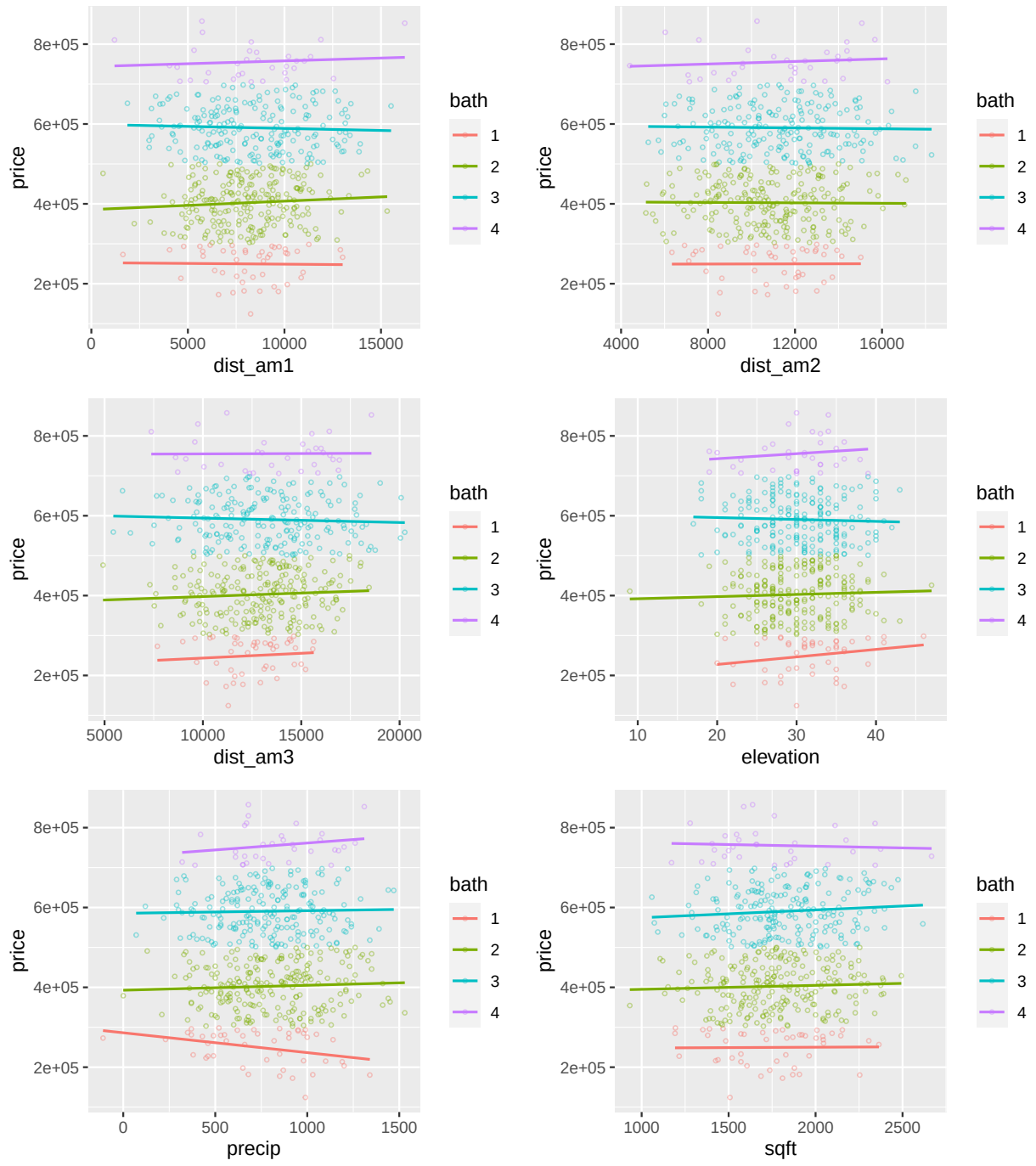


Figure 8: Scatterplots of 'price' against all the rest numerical variables and a simple linear regression line superimposed, coloured by 'bath'.

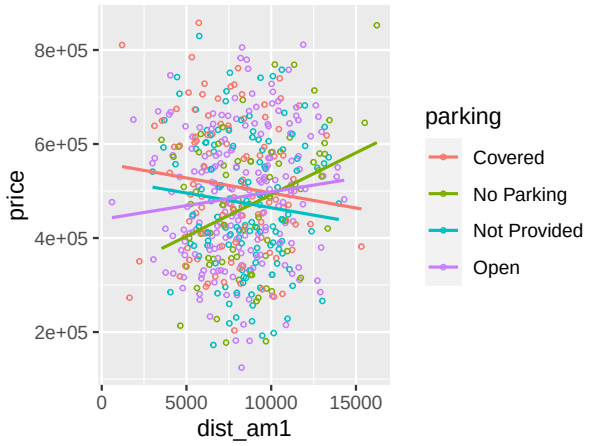


Figure 9: 'Price' against 'dist_am1' on all 'parking' levels.

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

4 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 **variable 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 solid results from variable selection via cross-validation (even with 5, 6 or 10 folds) than via the simple 3-subset split; small changes on the training set 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 linear models by smoothing the best regression model's least squares 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.

4.1 Fitting the full model

In section 3, we discussed about the different available variables in our data and we explored numerically and graphically the relationships between them and especially with *"price"*. Based on our findings, we select as our starting model the linear, additive model on the original variables, including interactions between *"precip"* - *"bath"* and *"dist_am1"* - *"parking"*. Therefore, we assume the true relationship is:

$$
\begin{aligned}
price_i = \beta_0 + \beta_1 \cdot \mathbb{I}bath|2_i + \beta_2 \cdot \mathbb{I}bath|3_i + \beta_3 \cdot \mathbb{I}bath|4_i + \\
\beta_4 \cdot \mathbb{I}parking|NoParking_i + \beta_5 \cdot \mathbb{I}parking|NotProvided_i + \beta_6 \cdot \mathbb{I}parking|Open_i + \\
\beta_7 \cdot sqft_i + \beta_8 \cdot elevation_i + \beta_9 \cdot dist\_am2_i + \beta_{10} \cdot dist\_am3_i + \\
\beta_{11} \cdot precip_i + \beta_{12} \cdot \mathbb{I}bath|2_i \cdot precip_i + \beta_{13} \cdot \mathbb{I}bath|3_i \cdot precip_i + \beta_{14} \cdot \mathbb{I}bath|4_i \cdot precip_i + \\
\beta_{15} \cdot dist\_am1_i + \beta_{16} \cdot \mathbb{I}parking|NoParking_i \cdot dist\_am1_i + \beta_{17} \cdot \mathbb{I}parking|NotProvided_i \cdot dist\_am1_i + \\
\beta_{18} \cdot \mathbb{I}parking|Open_i \cdot dist\_am1_i \\
+ \epsilon_i, \quad \epsilon_i \overset{indep.}{\sim} N(0, \sigma^2), \quad i = 1, ..., 450
\end{aligned}
$$

, 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 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, 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	271886.973	0.000	198975.549	344798.396
bath: 2	93953.790	0.000	44578.854	143328.727
bath: 3	297159.437	0.000	247851.175	346467.699
bath: 4	428750.065	0.000	350689.585	506810.545
parking: No Parking	-14771.862	0.603	-70485.804	40942.080
parking: Not Provided	-6293.913	0.812	-58146.888	45559.062
parking: Open	-10592.691	0.640	-55033.450	33848.068
precip	-53.092	0.051	-106.387	0.203
dist_am1	1.312	0.601	-3.618	6.242
dist_am2	-0.874	0.499	-3.411	1.664
dist_am3	-0.044	0.981	-3.678	3.589
sqft	10.548	0.217	-6.202	27.299
elevation	316.051	0.525	-660.348	1292.449
bath: 2:precip	73.971	0.016	13.759	134.183
bath: 3:precip	54.387	0.079	-6.427	115.201
bath: 4:precip	92.066	0.050	-0.107	184.239
parking: No Parking:dist_am1	-0.215	0.947	-6.585	6.154
parking: Not Provided:dist_am1	-0.256	0.934	-6.373	5.861
parking: Open:dist_am1	-0.232	0.932	-5.572	5.108

Lastly, we can calculate the average Mean Squared Prediction Error of the full model via 5-fold cross-validation, which is equal to $\hat{MSPE} = 2.9767139 \times 10^9$. This metric will be used for performing variable selection as described in the following subsection.

4.2 Variable 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. 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 **variable 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	average $\widehat{MSPE}$ after removal
full model	2976713912.5101
dist_am1*parking	2917944935.90957
sqft	2888542067.74259
dist_am3	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. Following the order of the removals, the method decreased the complexity of the model by dropping the interaction between *"dist_am1"* and *"parking"*, *"sqft"*, *"dist_am3"*, *"parking"*, *"dist_am2"* and, finally, *"elevation"*. That way, the method managed to reduce the average $\widehat{MSPE}$ down to 2.8389802×10^9 from the initial value of 2.9767139×10^9. Therefore, according to our findings we now update our assumption about the true relationship between *"price"* and our predictors in the population such that:

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

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

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

The results from fitting the above model on the training data set is 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 precipitaion and *"bath"* seem to be significant predictors regarding the sale price of a house. More specifically, our fitted model estimates that:

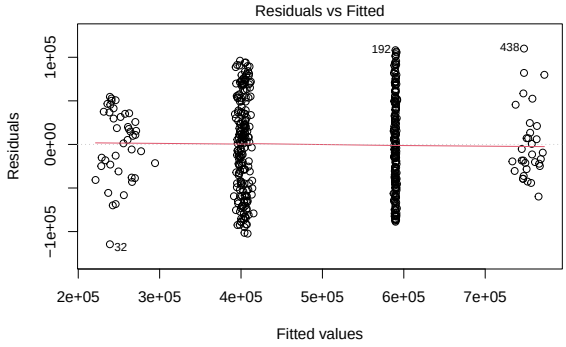
- **for houses with 0 precipitation**, the expected sale price of a house with just one bathroom is, on average, 288833.246. This base price increases, on average, by 98592.790, 298359.921 or 431877.270 for 2, 3 or 4 bathrooms, respectively. This gradual increase in the price as we increase the number 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 3.
- **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 by 18.277, 3.206 or 39.697 per unit of precipitation when the houses have 2, 3 or 4 bathrooms, respectively.

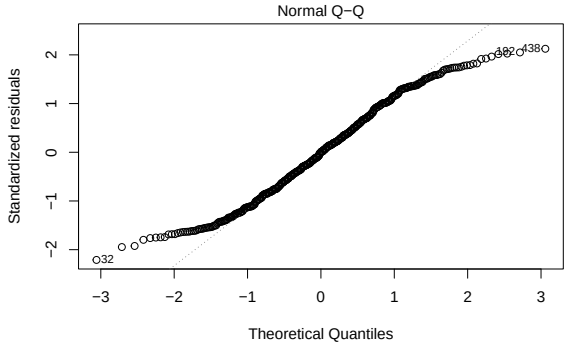
It is, also, worth mentioning that $R^2_{(adj)}$, which takes the complexity of the used model into account, improves from the full model's value 0.8612813 to 0.8622949, suggesting that we are able to explain 86.23% of the variation in the sale price.

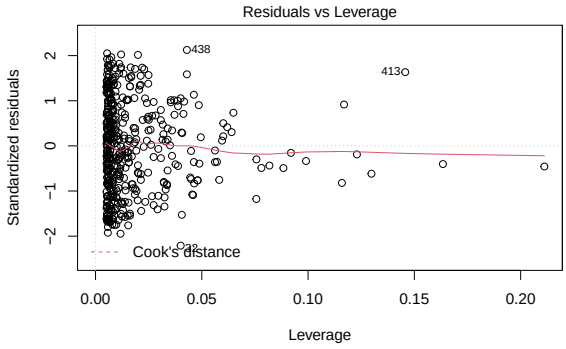
Table 9: The final model fitted on the training set after performing variable selection.

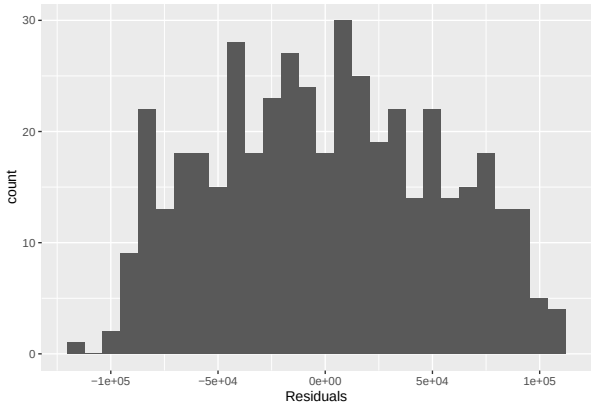
term	estimate	p_value	lower_ci	upper_ci
intercept	288833.246	0.000	246468.064	331198.427
bath: 2	98592.790	0.000	49797.811	147387.770
bath: 3	298359.921	0.000	249575.078	347144.763
bath: 4	431877.270	0.000	355067.249	508687.291
precip	-50.454	0.060	-103.126	2.218
bath: 2:precip	68.731	0.024	9.156	128.307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

Assumptions: mention that residuals vs. not included predictors show no relation









~~ TODO ~~

4.3 Regularised regression with Ridge and Lasso

variable selection: discrete. Shrinkage methods are continuous allowing for more flexible complexity reduction. ~~ TODO ~~

5 Conclusions

~~ TODO ~~

6 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 ~~

7 References