HOUSING PRICE IN KING COUNTY IN 2014-2015

LingXiang Zou 301289420

JunChao Liu 301299668

JiaJun Zhang 301327368

TABLE OF CONTENTS

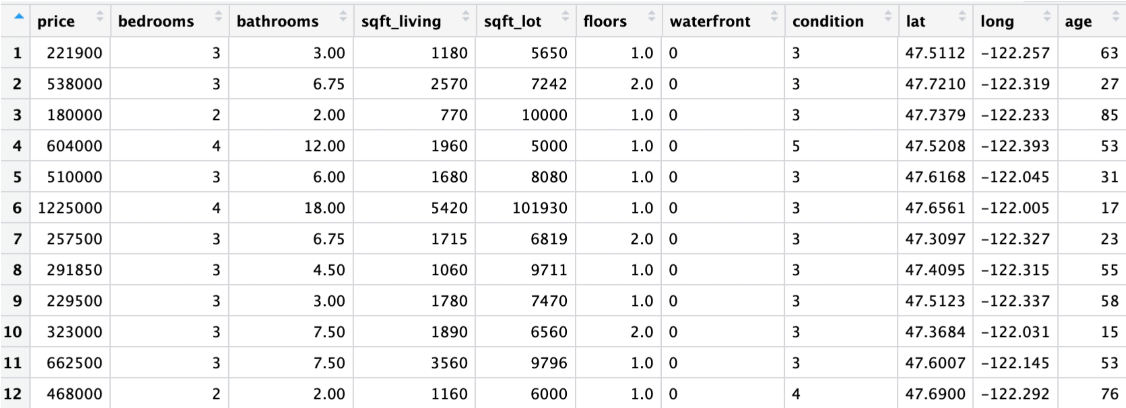
ABSTRACT

Regarding all the factors affecting the housing price, some argue locations are the primary reason and the other say that lot size is dominating among other factors. The answers are diversified and complex. However, due to the subjective impressions of the public, they do not have enough evidences or theatrical reasons for the influential factors. The purpose of this paper is to perform objective analysis of the housing price in King County,Washington, by building a multiple linear regression model.

INTRODUCTION

We obtain our dataset from Kaggle with 21 variables and over 21 thousands observations originally. After data cleaning, 11 variables are selected. Our response variable is the price of the house, and the remaining are the explanatory variables. We apply all forward, backward, and piecewise selection on the multiple linear regression model. After that, we improve the model using Box-Cox transformation and implement two machine learning techniques, Lasso and Ridge regression.

DATASET



	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	condition	lat	long	age
1	221900	3	3.00	1180	5650	1.0	0	3	47.5112	-122.257	63
2	538000	3	6.75	2570	7242	2.0	0	3	47.7210	-122.319	27
3	180000	2	2.00	770	10000	1.0	0	3	47.7379	-122.233	85
4	604000	4	12.00	1960	5000	1.0	0	5	47.5208	-122.393	53
5	510000	3	6.00	1680	8080	1.0	0	3	47.6168	-122.045	31
6	1225000	4	18.00	5420	101930	1.0	0	3	47.6561	-122.005	17
7	257500	3	6.75	1715	6819	2.0	0	3	47.3097	-122.327	23
8	291850	3	4.50	1060	9711	1.0	0	3	47.4095	-122.315	55
9	229500	3	3.00	1780	7470	1.0	0	3	47.5123	-122.337	58
10	323000	3	7.50	1890	6560	2.0	0	3	47.3684	-122.031	15
11	662500	3	7.50	3560	9796	1.0	0	3	47.6007	-122.145	53
12	468000	2	2.00	1160	6000	1.0	0	4	47.6900	-122.292	76

RESPONSE VARIABLE

y = Price:	Price of the house (in US dollars)

EXPLANATORY VARIABLE

x1 = Bedrooms:	Number of bedrooms per house
x2 = Bathrooms[1]:	Number of bathrooms per bedrooms
x3 = sqft_living:	Area inside the house (in square feet)
x4 = sqft_lot:	Area onside the house (in square feet)
x5 = floors[2]:	Number of floors ("1", "1.5", "2", "2.5", "3", "3.5")
x6 = waterfront:	Whether the house has water view? ("0", "1")
x7 = condition:	Condition of the house? ("1", "2", "3", "4", "5")
x8 = age:	Age of the house
x9 = lat:	Latitude coordinate
x10 = long:	Longitude coordinate

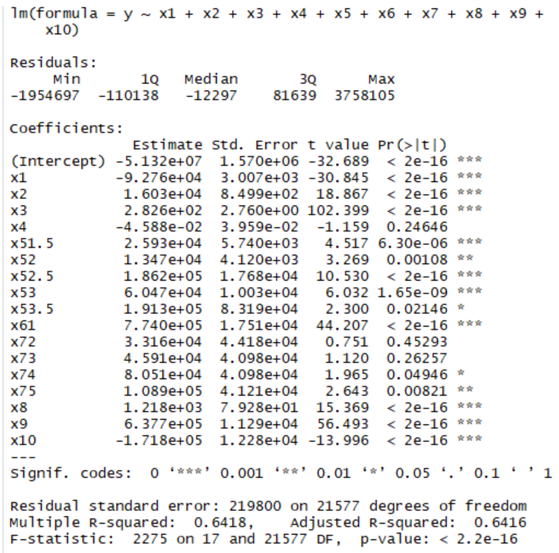
MODELS

Full model:

$$\hat{y} = \beta0 + \beta1 * x1 + \beta2 * x2 + \beta3 * x3 + \beta4 * x4 + \beta5 * x5 + \beta6 * x6 + \beta7 * x7 + \beta8 * x8 + \beta9 * x9 + \beta10 * x10$$

[1] Bathrooms Full = Tub, sink, toilet; 3/4 = Shower without tub, sink, toilet; 1/2 = Sink and toilet

[2] Floors 1.5 floors = lower floor is twice the area of upper floor

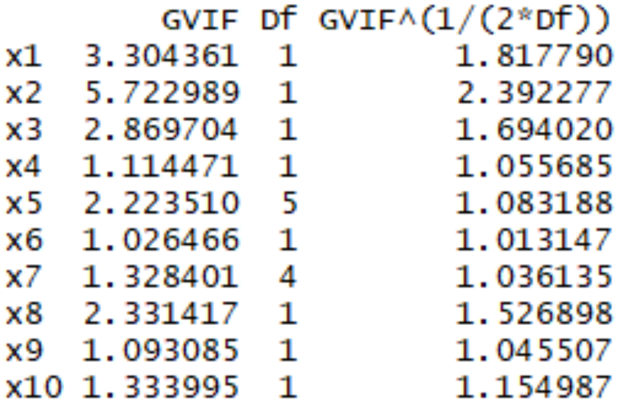


```
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + 
    x10)

Residuals:
     Min       1Q   Median       3Q      Max 
-1954697  -110138   -12297    81639  3758105 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.132e+07  1.570e+06 -32.689  < 2e-16 ***
x1          -9.276e+04  3.007e+03 -30.845  < 2e-16 ***
x2           1.603e+04  8.499e+02  18.867  < 2e-16 ***
x3           2.826e+02  2.760e+00 102.399  < 2e-16 ***
x4          -4.588e-02  3.959e-02  -1.159  0.24646    
x51.5        2.593e+04  5.740e+03   4.517 6.30e-06 ***
x52          1.347e+04  4.120e+03   3.269  0.00108 ** 
x52.5        1.862e+05  1.768e+04  10.530  < 2e-16 ***
x53          6.047e+04  1.003e+04   6.032 1.65e-09 ***
x53.5        1.913e+05  8.319e+04   2.300  0.02146 *  
x61          7.740e+05  1.751e+04  44.207  < 2e-16 ***
x72          3.316e+04  4.418e+04   0.751  0.45293    
x73          4.591e+04  4.098e+04   1.120  0.26257    
x74          8.051e+04  4.098e+04   1.965  0.04946 *  
x75          1.089e+05  4.121e+04   2.643  0.00821 ** 
x8           1.218e+03  7.928e+01  15.369  < 2e-16 ***
x9           6.377e+05  1.129e+04  56.493  < 2e-16 ***
x10         -1.718e+05  1.228e+04 -13.996  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 219800 on 21577 degrees of freedom
Multiple R-squared:  0.6418,    Adjusted R-squared:  0.6416 
F-statistic:  2275 on 17 and 21577 DF,  p-value: < 2.2e-16
```

Figure 1

In figure 1, most of the regressors in the full model are significant, the R-squared and Adjusted R-Squared are pretty good, saying that our regressors can explain more than 60% of variability in the response variable. However, since we have 10 independent variables, multicollinerity will likely to be appeared. The next step would be checking for multicollinerity and simpliflying the model.

MULTICOLLINEARITY AND MODEL SELECTION:



```
        GVIF Df GVIF^(1/(2*Df))
x1  3.304361  1        1.817790
x2  5.722989  1        2.392277
x3  2.869704  1        1.694020
x4  1.114471  1        1.055685
x5  2.223510  5        1.083188
x6  1.026466  1        1.013147
x7  1.328401  4        1.036135
x8  2.331417  1        1.526898
x9  1.093085  1        1.045507
x10 1.333995  1        1.154987
```

Figure 2

	Model selected	AIC value
Forward	y ~ x3+x9+x6+x10+x1+x7+x2+x8+x5	531276.7
Backward	y ~ x1+x2+x3+x5+x6+x7+x8+x9+x10	531276.7
Both	y ~ x1+x2+x3+x5+x6+x7+x8+x9+x10	531276.7

Table 1

In figure 2, values of GVIF$^{(1/2(Df))}$ address that there're no multicollinerity issue in the model and from Table 1, all three directions give the same AIC values and same models, excepting the orders are different in forward selection. Therefore, for convience, we choose the model in stepwise for our reduced model.

REDUCED MODEL:

$\hat{y} = \beta0 + \beta1 * x1 + \beta2 * x2 + \beta3 * x3 + \beta5 * x5 + \beta6 * x6 + \beta7 * x7 + \beta8 * x8 + \beta9 * x9 + \beta10 * x10$

BOX-COX TRANSFORMATION:

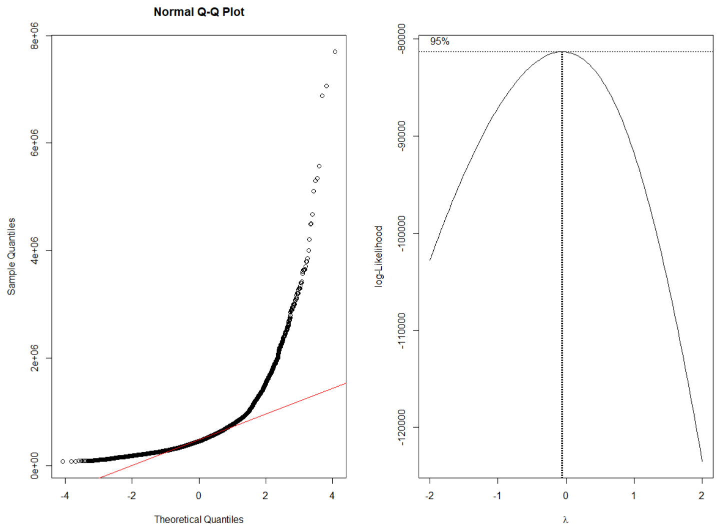


Figure 3

In Figure 3, the normality assumption of the response variable has been violated. In order to fix the problem, box-cox transformation method has been used and the best- λ is -0.0606. So y^{λ} is the suggested transformation by box-cox method.

FINAL MODEL:

$\hat{y}^{-0.0606} = \hat{y} = 2.898 + 0.001696 * x1 - 0.0002652 * x2 - 0.00001020 * x3 - 0.002714 * x51.5$

$- 0.002815 * x52 - 0.005591* x52.5 - 0.003910*x53 -0.003699*x53.5 - 0.01631* x61$

$- 0.002647 * x72 - 0.007641 * x73 - 0.009679 * x74 - 0.01129 * x75 - 0.00002212* x8$

$- 0.04206 * x9 + 0.003396 * x10$

FINAL MODEL DIAGNOSTIC:

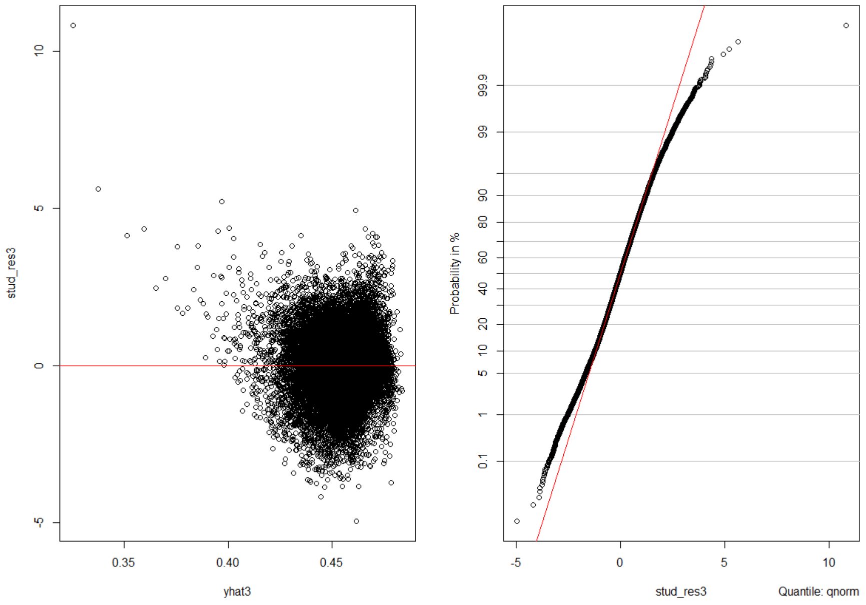


Figure 4

The linearity, constant variance and normality assumptions are satisfied. From the studentized residual versus predicted value plot in figure 4, most of the residuals are randomly distributed except few potential outliers, so linearity and constant variance conditions are met. For normality assumption, most of the points are in the diagonal line; however, there're some points outside the line in both tails but the deviations are still acceptable, comparing to the full model.

In conclusion, the final transformed model is the best model for ordinary least square estimation, the adjusted R-square also rises from 0.6416 in the full model to 0.6846 in our final model.

SHRINKAGE METHODS

The methods we choose to apply on our model are ridge and lasso regression. These methods are very similar, and they are both regulated by adding a penalty term to the sum of squared residuals.

FORMULAS

Lasso regression: RSS + $\lambda\sum_{j=1}^{p} |\beta_j|$

Where $\lambda\sum_{j=1}^{p} |\beta_j|$ is called penalty term and $\lambda \geq 0$ is a tuning parameter.

Ridge regression: RSS + $\lambda\sum_{j=1}^{p} {\beta_j}^2$

where $\lambda\sum_{j=1}^{p} {\beta_j}^2$ is called penalty term and $\lambda \geq 0$ is a tuning parameter.
If λ is 0, then both of the regressions will be an ordinary least square regression.

LASSO REGRESSION

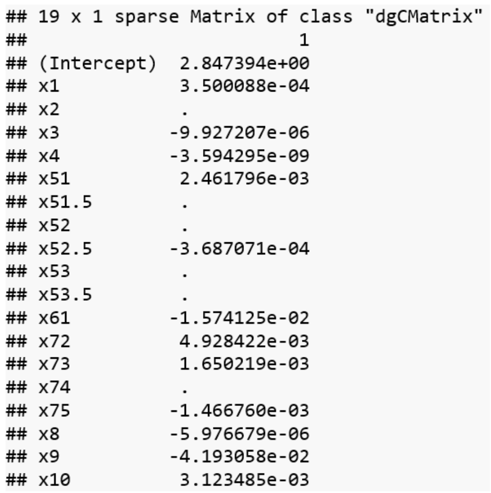
Lasso regression performs two main tasks: regularization and variable selection. According to the formula above, Lasso applies a regularization or shrinkage process where it penalizes the coefficients of the regressors all the way to zero. Other non-zero coefficients after variable selection will be selected to be parts of the model. That is due to the strength of the penalty which is controlled by the tuning parameter λ. The larger parameter λ is, the more numbers of coefficients are shrunk to zero.

RIDGE REGRESSION

Ridge regression is another shrinkage method. In the case of variable selection, it cannot reduce the numbers of parameters in the model. However, it does shrink the coefficients of the regressors towards zero, but never equal to zero. Evidently, it performs better than Lasso regression when most of the explanatory variables are highly correlated. Furthermore, this method is widely used tool to the problem of multicollinearity. It can effectively eliminate collinearity, improve accuracy, and provide more interpretable parameter estimates.

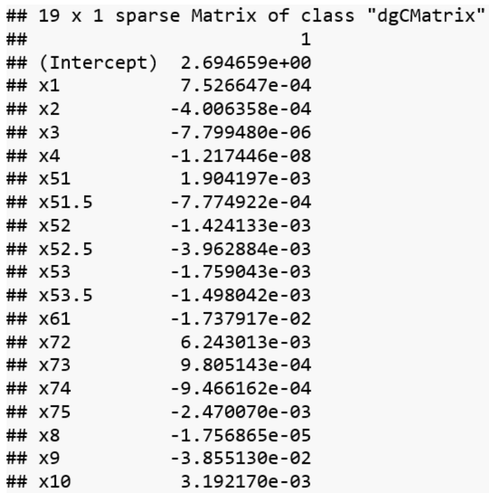
MODELS COMPARISON

After getting the model from traditional AIC selection methods, we are interested in some new machine learning techniques, such as Lasso (least absolute shrinkage and selection operator) and Ridge, to see if there is a significant improve in our model prediction and decide which model is best to use for our dataset.



```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                         1
## (Intercept)  2.847394e+00
## x1           3.500088e-04
## x2           .
## x3          -9.927207e-06
## x4          -3.594295e-09
## x51          2.461796e-03
## x51.5        .
## x52          .
## x52.5       -3.687071e-04
## x53          .
## x53.5        .
## x61         -1.574125e-02
## x72          4.928422e-03
## x73          1.650219e-03
## x74          .
## x75         -1.466760e-03
## x8          -5.976679e-06
## x9          -4.193058e-02
## x10          3.123485e-03
```

Figure 5



```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                         1
## (Intercept)  2.694659e+00
## x1           7.526647e-04
## x2          -4.006358e-04
## x3          -7.799480e-06
## x4          -1.217446e-08
## x51          1.904197e-03
## x51.5       -7.774922e-04
## x52         -1.424133e-03
## x52.5       -3.962884e-03
## x53         -1.759043e-03
## x53.5       -1.498042e-03
## x61         -1.737917e-02
## x72          6.243013e-03
## x73          9.805143e-04
## x74         -9.466162e-04
## x75         -2.470070e-03
## x8          -1.756865e-05
## x9          -3.855130e-02
## x10          3.192170e-03
```

Figure6

After using lasso regression, we can clearly see, from the coefficient table figure 5, some of the estimate of coefficients appear as "." which are considered as 0. The zero terms are x2(number of bathrooms), x5(1.5), x5(2), x5(3), x5(3.5) and x7(4), respectively. This demonstrates that, as lasso applied, since those coefficients are not significant enough in predicting, they are zeroed out in order to enhance the prediction accuracy overall. Nonetheless, we obtain a model with all the variables involved from figure 6, as ridge regression suggests. Given the three models from different techniques, we begin to think which is our best model in terms of estimation or prediction. Our goals are to see whether there are strong relationships between our observed responses and their predicted values. In this report, we mainly focus on the MSE(mean squared error) of all three models and determine the optimal model on that basis.

CROSS VALIDATION

While working with lasso and ridge regression, we divide our dataset into training set and testing set. In fact, they are obtained from our dataset after cleaning where randomly two-third of the data are assigned as training and the remaining one-third considered as testing.

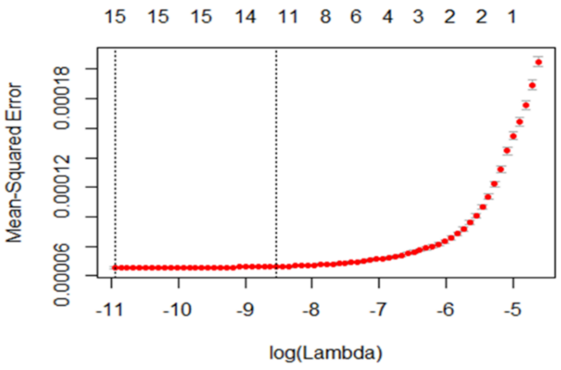
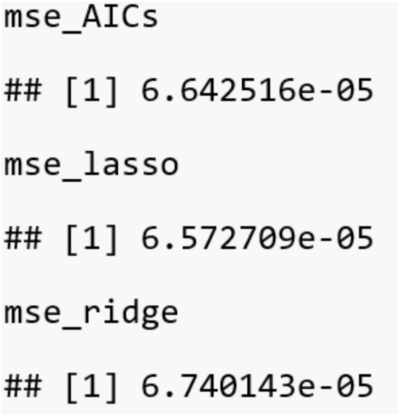


Figure 7

The idea of cross validation is using our training data to find the optimal lambda and use that lambda value in our further prediction. In figure 7, we obtain a plot of all lambda values by lasso regression. As a result, it provides two significant lambda values where the dotted line on the left suggests the value of lambda.min and line on the right gives the value of lambda.1se. The use of these two lamba is somewhat different, while lambda.min provides a best model but might be complex and overfitted, lambda.1se gives a simplest model within one standard error of the best model. Lambda.1se tends to be used when lasso regression applied. However, in order to keep the consistency, we use lambda.1se for both lasso and ridge. Then, we use our training data in a function with lambda.1se to produce the predicted values of the response. Now, our focus becomes the difference between the observed and predicted values.

MEAN SQUARE ERROR COMPARISON

Therefore, the MSE of lasso and ridge are resulted below where the MSE of lasso is slightly smaller than ridge’s.



```
mse_AICs

## [1] 6.642516e-05

mse_lasso

## [1] 6.572709e-05

mse_ridge

## [1] 6.740143e-05
```

Figure 8

The fact that these two values are esstientially small leads us to think both models are satisfactory, considering the small variation between the prediction and true values. However, we also need to determine if the model from AIC selection performs better. Same idea applied as we divide the data into training and testing, but this time we remove the observation of x4(square feet of the lot) before divison given that it is removed from AIC selection method. Ultimately, we have the MSE result from all three model, as shown in figure 8, where the MSE from lasso regression is the smallest. This indicates, of all three models, lasso performs the best since the smallest MSE agrees that there is almost no difference between prediction and reality.

CONCLUSION

Finally, based on a reasonable MSE comparison, we decide the model from lasso regression is our best model. Since we previously obtain the coefficients of lasso regression, we can now express our model as $\hat{y}^{*} = 2.847 + 0.00035*x1 - 0.0000099*x3 - 0.00000000359*x4 + 0.00246*x5(1) - 0.000369*x5(2.5) - 0.0157*x6(1) + 0.00493*x7(2) + 0.00165*x7(3) - 0.00147*x7(5) - 0.00000598*x8 - 0.0419*x9 + 0.00312*x10$ where $\hat{y}^{*} = \hat{y}^{-0.0606}$. Based on our final model, an example can be interpreted as a house in King County, Washington with 3 bedrooms, 2000 square foot of living, 6000 square foot of lot, 2.5 floors,not waterfront, at the condition of 4, 50 years old and locats at 47.5 latitude, -122.2 longitude, the predicted price will be $432131.

DISCUSSION

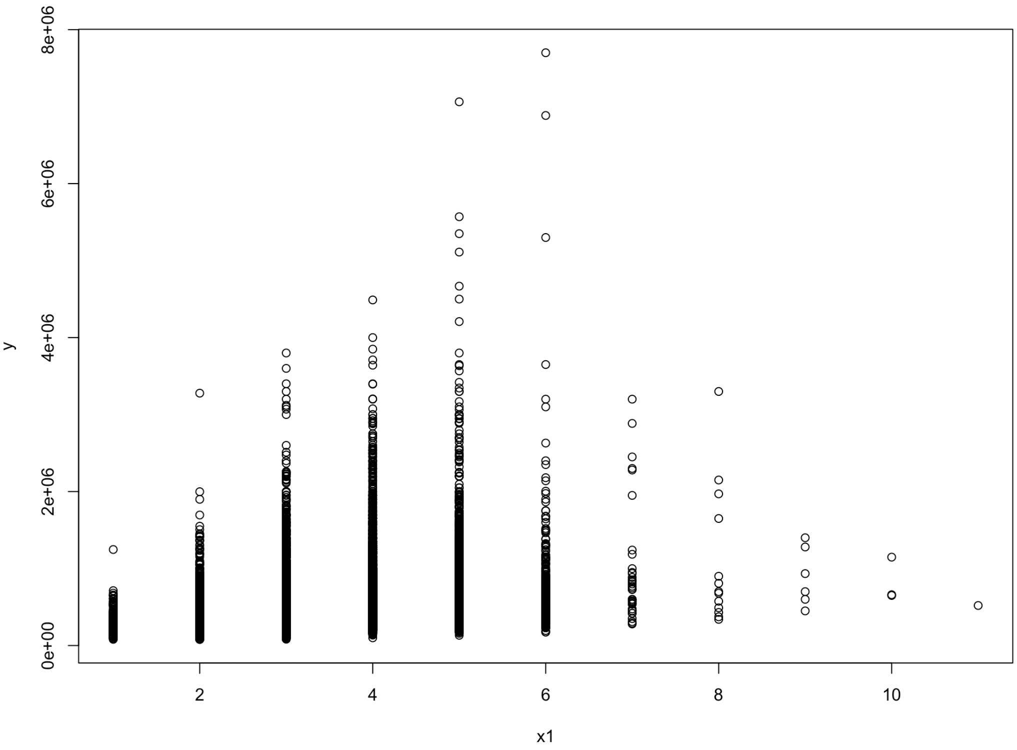


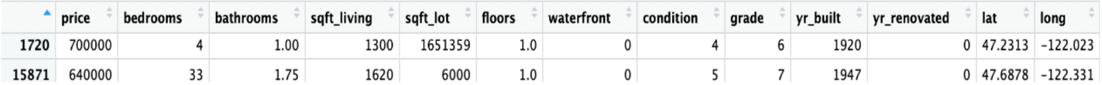
Figure 9

In all models, the regressor x1(bedrooms) has a negative relationship with the predicted price which is not expected, in this case, with all same conditions, a house with 3 bedrooms will be cheaper than a house with 4 bedrooms, which makes no sense. In figure 9, we plot the x1 verses y, we can see the price of houses with 4 bedrooms and 5 bedrooms are generally more expensive than houses with 6 to 10 bedrooms which is interesting.So we conclude that The appearance of the wrong sign may due to the missing of important regressors, since the data doesn’t include the housing type, and the price of apartment and house varies a lot in King County, Washington, so the data may need further discussion with more information.

APPENDIX

Data cleaning

- Data cleaning is the process of detecting and correcting errors from a dataset in order to improve its quality and accuracy.
- Removed the variables that are uncorrelated to our objective.
    - Id, Date, grade, sqft_above, sqft_basement, yr_built, yr_renovate, zipcode, sqft_living15 , sqft_lot15
- Delete the rows that have missing values.
- Manipulated the variable, bathrooms, to number of bathrooms per house.
- Replaced the variable, yr_built, with the variable, yr_renovated, if the house was renovated.
- Delete the rows that have irrelevant observations.
    - The observations 1720 and 15871 are either input errors or outliers. The variable, sqft_lot, of observation 1720 is way bigger than the other observations. Same for the variable, bedrooms, of observation 15871.



	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	condition	grade	yr_built	yr_renovated	lat	long
1720	700000	4	1.00	1300	1651359	1.0	0	4	6	1920	0	47.2313	-122.023
15871	640000	33	1.75	1620	6000	1.0	0	5	7	1947	0	47.6878	-122.331

- Converted the variable, yr_built, to age of the house

REFERENCE

A gentle introduction to logistic regression and lasso regularisation using R. (2017, October 6). Retrieved https://eight2late.wordpress.com/2017/07/11/a-gentle-introduction-to-logistic-regression-and-lasso-regularisation-using-r/

Harlfoxem. (2016). House Sales in King County, USA. Retrieved from https://www.kaggle.com/harlfoxem/housesalesprediction

Ridge Regression | Columbia University Mailman School of Public Health. (n.d.). Retrieved from https://www.mailman.columbia.edu/research/population-health-methods/ridge-regression

Valeria, F. (2017). FEATURE SELECTION USING LASSO. Retrieved from VRIJE UNIVERSITEIT https://beta.vu.nl/nl/Images/werkstuk-fonti_tcm235-836234.pdf

Free Seattle Wallpapers 4K (1920x1080 px) http://zyzixun.net/image-download/3397117.html