Least Squares Housing Price Prediction

# Goals

The goal of this project was to conduct a study on the importance and uses of the least squares methods on a linear regression problem. Using a public dataset on housing data, we created a regression model based on the method of least squares and performed an error analysis on the resulting model. In addition to the least squares method, we tested related modeling techniques to compare the strengths and weaknesses of each prediction.

# Introduction

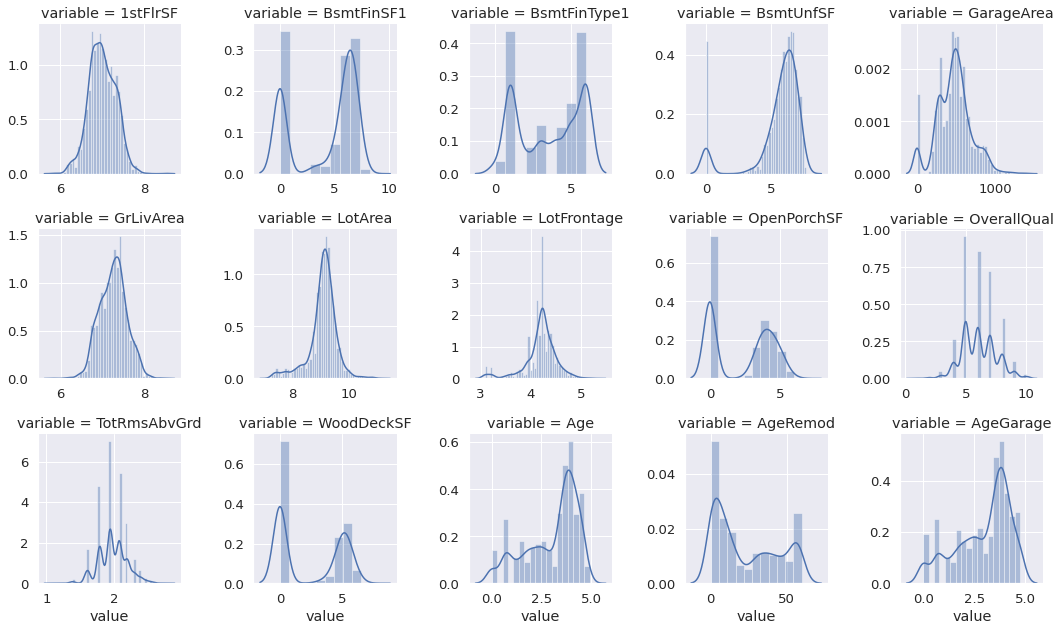
The least squares method is a common approach for dealing with linear regression problems. It is a technique that provides approximate solutions for overdetermined systems, i.e. sets of data where the number of datapoints exceeds the number of features / variables on the data points. The method focuses on minimizing the sum of squared residuals on the approximate solution.

With this in mind, we wanted to choose a data set that had a high number of features that we could explore and selectively pick ones that we deemed could be reasonable as inputs to a regression model. It was also necessary to choose a data set that had a suitable output target for prediction. With these points in mind, we chose to train a prediction model on housing data with the goal of outputting the price of a home, using sale prices as the stand-in for true values.

# Data

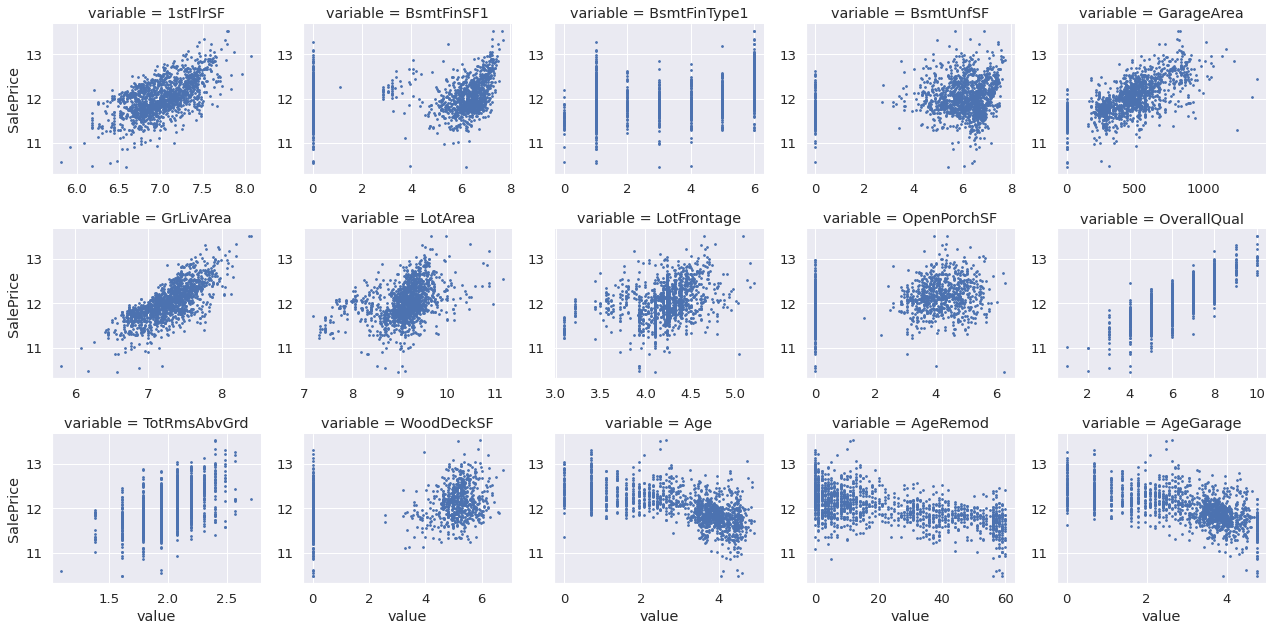
The data for this project was extracted from a kaggle competition “House Prices - Advanced Regression Techniques” [1]. The dataset contains 79 different housing features that are both quantitative and qualitative. These features range from reasonable “Lot Area” all the way to “2nd Garage Area”. Data preprocessing was accomplished using an open source notebook found on GitHub [2]. This notebook accomplishes all relevant data transformation, data normalization, and feature extraction. Once complete, the new dataset contains 52 original and new features.

The initial analysis conducted involved extracting some of the more interesting features and only running the analysis on them. The features chosen can be seen in Fig. 1 below. The figure plots the distributions of the selected features and gives us insight into how the data is grouped. As you can see from the figure many of the features are unimodal, but there are a couple of features, such as: basement square footage and porch square footage, that are bimodal. This can easily be explained by reasoning that not every house has a basement or porch, which leads to the distribution having two modes.



**Fig. 1: Interesting Feature Distributions**

The second form of analysis done on the interesting features was to plot the feature values with their respective sale prices. Fig. 2, seen below, graphs the scatter plots for each of the variables. As we can see from the figure, features like first floor square footage and living area are highly correlated with sale price. This means that as their values increase the price of the dwelling will also increase. This is extremely useful in predicting a sale price because if there is a large living area then it can be expected that the sale price should also be large. Other features and age since last renovation are negatively correlated with sale price. Meaning that as their values increase the sale price will decrease. This is again useful since the model will learn that larger ages will negatively impact sale price. The final type of variable seen in Fig. 2, are variables with no correlation such as lot area. These types of variables do not offer any viable information regarding sale price because they do not positively or negatively impact it. Table 1 below offers are more in depth analysis on the mean, standard deviation, and percentages of the correlation features just discussed.



**Fig. 2: Interesting Feature Scatter Plots**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Mean** | **Std** | **Min** | **25%** | **50%** | **75%** | **Max** |
| 1st Floor SF | 7.00 | 0.32 | 5.81 | 6.78 | 6.99 | 7.23 | 8.54 |
| Living Area | 7.26 | 0.32 | 5.81 | 7.03 | 7.27 | 7.46 | 8.53 |
| Lot Area | 9.09 | 0.50 | 7.17 | 8.92 | 9.15 | 9.35 | 11.2 |
| Age | 3.04 | 1.32 | 0.00 | 2.08 | 3.58 | 4.03 | 4.992 |
| Renovation Age | 23.6 | 20.9 | 0.00 | 4.00 | 15.0 | 43.0 | 60.0 |

**Table 1: Positively, Negatively and No Correlation Feature Summary**

The final analysis conducted examined the mean, standard deviation, and percentages of positively and negatively correlated features. The positively and negatively correlated features chosen were living area and age respectfully. Tables 2 and 3, give a full description of their summary statistics. As expected, if we examine the mean of both features, as the age of the house increases the price of the house decreases and vice versa for living area. This was expected though based on the scatter plots for each of the features. This can also be reasoned with logic since an older home is less desirable and a large home is more desirable.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Age Range** | **Count** | **Mean** | **Std** | **Min** | **25%** | **50%** | **75%** | **Max** |
| 0 to 1 | 162 | 12.4 | 0.35 | 11.3 | 12.2 | 12.4 | 12.7 | 13.3 |
| 1 to 2 | 176 | 12.3 | 0.27 | 11.8 | 12.1 | 12.3 | 12.5 | 13.1 |
| 2 to 3 | 219 | 12.3 | 0.29 | 11.4 | 12.1 | 12.3 | 12.4 | 13.5 |
| 3 to 4 | 516 | 11.9 | 0.26 | 11.0 | 11.8 | 11.9 | 12.0 | 12.9 |
| 4 to 5 | 380 | 11.7 | 0.36 | 10.5 | 11.6 | 11.7 | 11.9 | 13.1 |

**Table 2: Negatively Correlated Feature Summary on Sale Price**

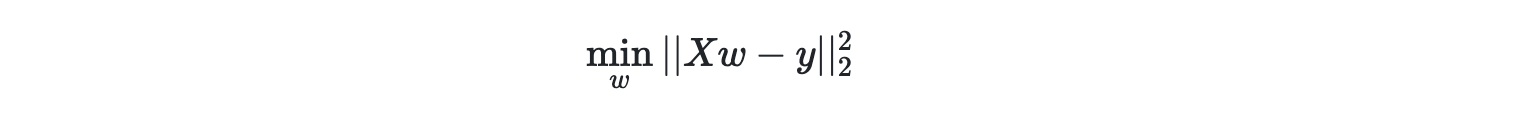
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Living Area** | **Count** | **Mean** | **Std** | **Min** | **25%** | **50%** | **75%** | **Max** |
| 0 to 7 | 325 | 11.6 | 0.27 | 10.5 | 11.5 | 11.7 | 11.8 | 12.1 |
| 7 to 7.27 | 373 | 11.9 | 0.25 | 10.6 | 11.7 | 11.9 | 12.1 | 12.9 |
| 7.27 to 7.52 | 447 | 12.1 | 0.28 | 11.3 | 12.0 | 12.1 | 12.3 | 12.9 |
| 7.52 to 7.63 | 122 | 12.3 | 0.29 | 11.6 | 12.2 | 12.3 | 12.5 | 13.0 |
| 7.63 to 9 | 186 | 12.5 | 0.38 | 11.4 | 12.3 | 12.5 | 12.7 | 13.5 |

**Table 3: Positively Correlated Feature Summary on Sale Price**

# Models

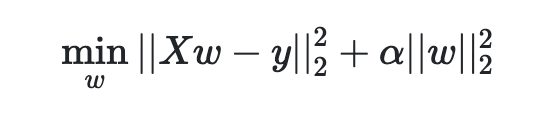
## Ordinary Least Squares

This fits a linear model to minimize the squared residuals of observed targets and predicted targets via a linear approximation. The objective function for Ordinary Least Squares is:



## Ridge Regression

This fits a linear model similar to that of ordinary least squares, but imposes a penalty on the size of the coefficients. This penalization effectively blocks the model from over fitting on a single feature. The objective function for Ridge Regression is:

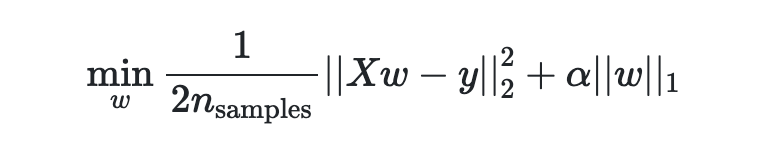


## Cross Validated Ridge Regression

This model fits a Ridge Regression model, but with the added bonus of cross validating with training data to update training weights. This model has the same objective function as Ridge Regression, but has an added training cross validation.

## Lasso Regression

This fits a linear model that estimates the amount of sparse coefficients, which is particularly useful with this dataset since the qualitative features needed to be transformed into quantitative features, which introduced sparsity. This model automatically reduces the number of features the solution is dependent on. The objective function of Lasso Regression is:



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **Description** | **Reason** | **Range** | **Function** |
| Explained Variance | Measures the discrepancy between predicted and true values | Indicates the predicted and true values are both correlated in the same way | [0, 1] |  |
| Max Error | Measures the worst case error between predicted and true values | Describes how bad the model could be | [0, Inf) |  |
| Mean Absolute Error | Measures the errors of paired observations | Describes how bad the model is overall | [0, Inf) |  |
| Mean Squared Error | Measures the average of squared errors between predicted and true values | Indicates that the errors are correlated and not to far away from each other | [0, Inf) |  |
| r2 Score | Measure the model’s goodness of fit | Indicates that the model’s errors don’t deviate | [0, 1] |  |

**Table 4: Experiment Metrics**

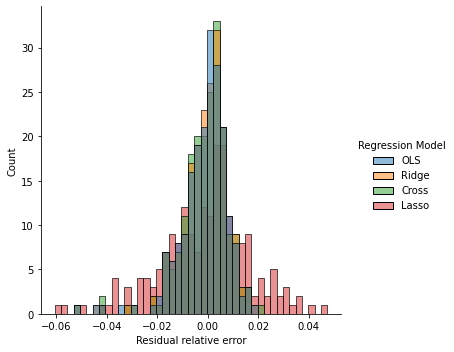
# Results

The results of the experiments conducted can be seen in Table 5 below. As we can see from inspecting the r2 score for each model, they all did exceptionally well with the exception of Lasso Regression. This can be explained by Lasso removing some of the sparser features during training that could have provided valuable information in edge cases. Ridge Regression did the best with a 99% accuracy over the entire testing dataset. This means that the model is very accurate at determining sale price. Ordinary Least Squares and Cross Validated Ridge regression did about the same with a 90% accuracy. Both models didn’t perform as well as Ridge Regression but each of them beat it in another category. Cross Validated Ridge Regression had the smallest max error meaning that in the worst case it won’t make an as bad prediction and Ordinary Least Squares had the smallest mean absolute error meaning overall it’s predictions are least likely to be overestimated or underestimated.

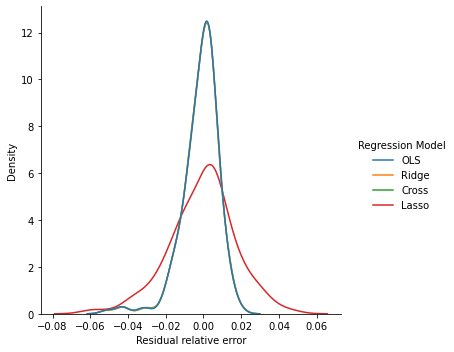
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Explained Variance** | **Max Error** | **MAE** | **MSE** | **R2** |
| Ordinary | 0.9021 | 0.6114 | **0.0827** | 0.0139 | 0.8996 |
| Ridge | 0.9023 | 0.6105 | 0.0828 | 0.0138 | **0.9900** |
| Cross Ridge | **0.9024** | **0.6097** | 0.0829 | 0.0138 | 0.9001 |
| Lasso | 0.5417 | 0.6806 | 0.1596 | 0.0435 | 0.5414 |

**Table 5: Experiment Result**

Comparing the results of the residual vectors for each least squares model, we found that the relative errors on the residual vector for OLS, Ridge regression, and cross validated ridge regression to be very similar, with similar distributions for error values. The lasso regression in comparison had more values along it’s tails, indicating that the lasso regression method had much greater variance in its predictions.



**Figure 6: Relative residual error on predictions**



**Figure 7: Probability density function estimated by kernel density estimation**

The kernel density estimation above better encapsulates the extreme similarity in predictions between OLS, Ridge Regression, and Cross Ridge. All three appear to follow a similar density function with similar mean and variance. The lasso regression method in comparison has a similar mean, but much greater variance as shown by the larger tail areas in the graph.

# Learning Outcomes

During our experiments, we learned how to apply the least squares method to a real-world problem. The housing data set sourced from Kaggle proved to be a well-suited problem for using the least squares method on. With a large dataset, we were able to visualize and select the most promising features for applying the least squares method on, and created prediction models that were fairly accurate.

This project also enabled exploration of other types of linear regression models, and observed the subtle changes in results stemming from changes in the cost function compared to the ordinary least squares method. From our experimentation, we learned that ridge regression and cross validated ridge regression showed comparable results to the ordinary least squares method when predicting housing price data, and that the lasso regression method was much worse when applied to the same problem.

# Contributions

Nicholas Miller: Data Preprocessing, All Model Fittings, Correlation Analysis and Feature Selection

Kevin Nguyen: Feature Engineering, Data Analysis and Least Squares Research, Analysis of results

# References

[1] https://www.kaggle.com/c/house-prices-advanced-regression-techniques

[2] https://github.com/liyenhsu/Kaggle-House-Prices

[3] https://www.investopedia.com/terms/l/least-squares-method.asp

[4] https://www.sciencedirect.com/topics/nursing-and-health-professions/least-square-analysis

[5] Michael T. Heath, 2002, Scientific Computing: An Introductory Survey, Second Edition