Linear Ordinary Least Squares and Ridge Regression to Predict Melbourne Housing Prices

Cameron Sequeira-Hogg 20722977

SYDE 312, Winter 2021

This report details an attempt to construct a Linear Regression model to predict housing prices in Melbourne based on several factors, including number of bedrooms, bathrooms, and distance from downtown. Initial analysis and pre-processing were completed in Python, to select important features and remove null rows from the dataset. An Ordinary Least Squares model was computed in Matlab using QR Decomposition, and a Ridge Regression model was fit using the Normal Equations. The Ridge Regression model performed better, but both models had high Root-Mean-Square Error, meaning that neither predicted new observations well.

1. Background and Discussion of the Application Area

Linear Least Squares, commonly referred to as Linear Regression by statisticians, is a method used to find a function that approximates a set of data points well. The term "Least Squares" stems from the fact that the sum of squared residual errors, the difference between the real and predicted values for each observation in a set of data, is minimized in the determination of the model. This method of least squares was pioneered by several mathematicians and statisticians in the 17th and 18th centuries **[1]**. Several different types of Regression have arisen since, including models such as Ridge Regression that attempt to reduce variance and improve accuracy in their predictions by adding a regularization term.

The best use of Linear Regression is to develop models for processes and datasets that can be well-described by a linear model. In this report, Linear Regression is applied to real estate data from Melbourne to predict house prices from other factors.

2. Use of Linear Algebra in the Analysis

The Ordinary Least Squares problem is to find the best approximation for the vector β in the equation $X\beta = y$, where the best approximation is the solution where the residuals are minimized:

$$\min_{x \in R^n} \sum (X\beta - y)^2$$

Where X is an $m \times n$ matrix, β is an $n \times 1$ column vector, and y is an $m \times 1$ column vector. There are three fundamental cases of the problem depending on the dimensions of the matrix X. If $m = n$ and the matrix X is non-singular, the solution simply becomes $\beta = X^{-1}y$. If $m < n$, there are more unknowns than equations and the problem is considered underdetermined. The most common case is when $m > n$, in which the problem is considered overdetermined, with more equations than unknowns. The overdetermined problem is the most common case encountered in Linear Regression, and the problem presented later in the report is this case of Least Squares. As a result, all following discussion will center around solving this overdetermined case.

In solving Least Squares problems, several methods can be used to find the vector β that minimizes the squared sum of residuals, including directly solving the resulting Normal Equations, solving through QR Decomposition or Singular-Value Decomposition.

Using the Normal Equations

Solving this problem (minimizing the sum of squares) through geometry, or by solving the over-determined system, we arrive at the following Normal Equations, an equation for the Least Squares fit, meaning that the β that solves the Normal Equations also is the solution to the Least Squares problem.

$$X^T X\beta = X^T y$$

$$\beta = (X^T X)^{-1} X^T y$$

If A has full rank, we can utilize one of several methods, including Gaussian Elimination, LU Factorization, or Cholesky Factorization to solve the Normal Equations. Solving the Normal Equations can be computationally quick and can be performed easily using the methods listed above but can lead to difficulties with inverting $A^T A$ if it is near-singular – leading to massive inaccuracies in computation of the coefficients.

Using QR Decomposition

Avoiding the Normal Equations entirely, the Least Squares problem can also be solved with QR Decomposition. The fundamental theorem for QR Decomposition is as follows.

If X is a real $m \times n$ matrix with $m \geq n$, and full rank, there exists a unique $m \times n$ orthogonal matrix Q ($Q^T Q = I_n$), and a unique $n \times n$ upper triangular matrix R with positive diagonals such that $X = QR$.

We use QR Decomposition to solve the Least Squares problem, by applying $X = QR$ to the equation $X\beta = y$. Rewriting the equation, we now have:

$$X\beta = y \rightarrow QR\beta = y \rightarrow R\beta = Q^T y$$

Back-substitution can then be applied to find β.

L2 Regularization (Ridge Regression)

L2 Regularization, commonly referred to as Ridge Regression, adds a different regularization term to OLS. The objective of Ridge Regression is to produce lower variance models that respond to noise better and produce less prediction error when tested on new data.

$$\min_{x \in R^n} \sum (X\beta - y)^2 + \lambda \sum_{i=1}^{n} {\beta_i}^2$$

The regularization term, multiplied by a variable coefficient λ, minimizes the size of the regression coefficients, flattening their slopes compared to the OLS estimates. It can help provide model estimates for datasets that suffer from large amounts of multicollinearity, the existence of linear relationships between some of the features.

This regularization term leads to the following Ridge Regression Normal Equations:

$$\beta = (X^T X + \lambda I)^{-1} X^T y$$

3. Presentation of a Specific Case Study or Problem

Housing prices depend on several factors including the neighbourhood of the property, distance from the city center, number of bedrooms and bathrooms, building size, and lot size. In this project, a dataset consisting of housing data from Melbourne, Australia, was analyzed to determine a linear model for house price based on several factors. The dataset consisted of 13580 observations 21 features associated with a listed house, including the response variable.

Dataset Pre-Processing and Analysis

Before solving the regression problems numerically in Matlab, initial pre-processing and descriptive analysis of the dataset was completed using Python.

One of the columns in the dataset was *Method*, representing whether the property was sold or not. To focus on an analysis of only properties that were sold, unsold observations were removed from the dataset. After this, 8 features that potentially influenced home prices were selected. These selected features were the number of rooms, the number of bathrooms, the region name, the distance from the downtown core, the number of properties in the neighbourhood, the type of property, the lot size, the building area, and the specific riding/council area. Rows in the dataset with missing values were dropped because numerically solving the Normal Equations in Matlab would be impossible. After these steps were taken, the dataset had 4116 remaining rows.

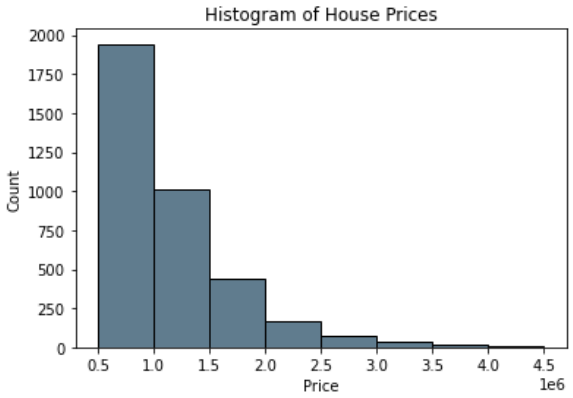

Figure 1: Histogram of the Response Variable, Price.

A histogram of the housing prices showed that of the 4116 observations, almost half were listings sold for between 0.5-1M AUD. The average value of a house in the dataset is 1.08M.

The Region Names and Council Areas were both categorical variables, with 8 regions and 30 council areas. To allow for creation of a linear model, these categorical features were transformed into new columns, adding 36 new features to represent these categories.

After this step, the dataset was split into X, the 4116×48 matrix, and y, the 4116×1 column vector representing the prices of the different observations. The matrix was normalized, as Ridge Regression requires features to have similar values and constant variance. The data was then split into two groups, with 80% of the data put into a "training" set and the remaining 20% held back to test the regression models after development.

4. Mathematical Solution and Selection of Appropriate Numerical Technique

The following pre-processed datasets were loaded into Matlab using the readmatrix() function.

X_{Train}: 3292×48, y_{Train}: 3292×1

X_{Test}: 824×48, y_{Test}: 824×1

A column vector of ones was added to the front of the X_{Train} and X_{Test} matrices, to add an intercept to the linear models, representing the first coefficient $\hat{\beta}_0$ in the coefficient vector.

Ordinary Least Squares Regression Implementation

We can attempt to use the Normal Equations to solve for the 49×1 vector of estimated model coefficients, $\hat{\beta}_{OLS}$, by implementing the following equation in Matlab.

$$\hat{\beta}_{OLS} = \left(X_{Train}^T X_{Train}\right)^{-1} X_{Train}^T y_{Train}$$

When solving for $\hat{\beta}_{OLS}$ with the Normal Equations, Matlab provides a warning that the matrix is close to singular or badly scaled. As seen in the closed-form solution for the coefficients using the Normal Equations, $X^T X$ needs to be inverted. If this matrix is ill-conditioned and near singular the inversion can result in large inaccuracies in the computation of $\hat{\beta}_{OLS}$.

To see if this is a problem, we can calculate β_{OLS} using QR Decomposition and compare the results of each model on the testing data. Using QR Decomposition:

$$X_{Train} = QR$$

$$R\hat{\beta}_{OLS} = Q^T y$$

Which can be solved in Matlab with back-substitution to find $\hat{\beta}_{OLS}$.

As the goal of Ordinary Least Squares is to minimize the sum of squared residuals in the model training data, the sum of squared residuals for the Training Set using $\hat{\beta}_{OLS}$ calculated with the Normal Equations against $\hat{\beta}_{OLS}$ computed with QR Decomposition were compared.

$$SS_{Res} = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

Where $\hat{y}_i$ represents the prediction made for each observation in the Training Set, based on the equation $X\hat{\beta} = \hat{y}$.

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j X_{ij}$$

Using these equations in Matlab:

$$Normal\ Equations: \ SS_{Res} = 5.7 \cdot 10^{14}$$

$$QR\ Decomposition: \ SS_{Res} = 4.713 \cdot 10^{14}$$

The Normal Equations clearly fail to minimize the sum of the squared residuals. Since QR Decomposition does not depend upon inverting $X^T X$, it is more accurate.

Ridge Regression Implementation

Several Ridge Regression models were run in Python with a grid search to find a numerical value for the regularization term λ that would minimize the Mean-Squared Error of Ridge Regression. A value of $\lambda = dd$ was selected for numerical computation in Matlab.

To compute the Ridge Regression coefficient estimates, we can use the Normal Equations to solve for the 49×1 vector of model coefficients, $\hat{\beta}_{Ridge}$, by implementing the following equation in Matlab.

$$\hat{\beta}_{Ridge} = \left(X_{Train}^T X_{Train}\right)^{-1} X_{Train}^T y_{Train}$$

The matrix is not close to singular, and as a result the Ridge Regression coefficient vector does not need to be found with QR Decomposition.

RMSE Computation

A common form of error measurement for Linear Regression is the Root-Mean-Square Error (RMSE) of the model. RMSE is calculated using the following equation.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2}$$

Where $\hat{y}_i$ represents each prediction from the Test Set, based on the equation $X\hat{\beta} = \hat{y}$.

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j X_{ij}$$

Using the above equations, the RMSE for predictions made with $\hat{\beta}_{OLS}$ and $\hat{\beta}_{Ridge}$ was computed in Matlab.

$$RMSE_{OLS} = 432{,}590$$

$$RMSE_{Ridge} = \ 432{,}370$$

5. Analysis and Interpretation of the Results in the Context of the Application

The RMSE values calculated show the performance of the models on new data – the models were created using only the X_{Train} matrix and y_{Train} vector, and the RMSE values were calculated with the X_{Test} matrix and y_{Test} vector.

Both linear models have an RMSE of ~432,000. This is an extremely high error, considering the average price of a home in the dataset was 1.08M.

Effect of Different Features on Price

Due to the scaling taken to guarantee the performance of Ridge Regression, the coefficients do not clearly show the effect of each feature on the price of a unit. However, by looking at graphs showing price plotted against different features, as well as looking at the sign of different coefficient values, we see their impact.

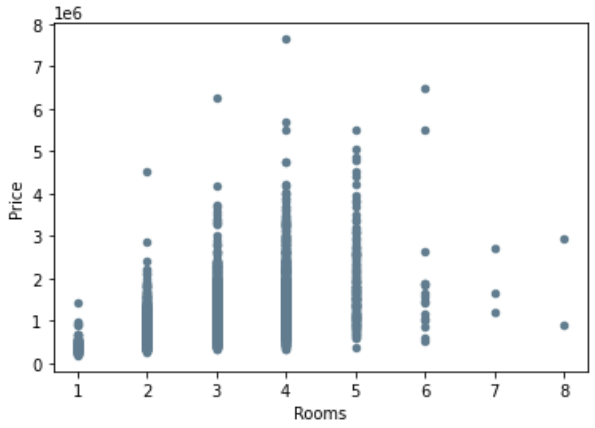

Figure 2: Scatter Plot of Number of Rooms and Price

From a scatter plot of the number of rooms versus the price of a property, we see a positive relationship, however the relationship may not be strictly linear. This qualitative observation is confirmed by the coefficient for the number of rooms being positive.

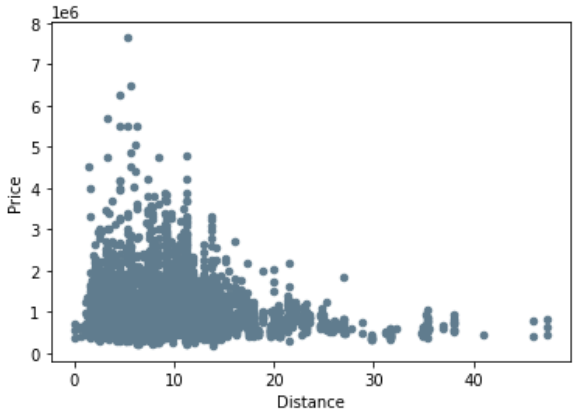

Figure 3: Scatter Plot of Number of Rooms and Price

From a scatter plot of the distance from downtown Melbourne versus the price of a property, we see a negative relationship. This is confirmed by the coefficient for the number of rooms being negative.

OLS Versus Ridge Regression

Ridge Regression works by adding a regularization term, representing a degree of bias. This helps deal with multi-collinearity, a situation when some features, represented by columns in X, are linear combinations of other features. By minimizing the effect of multi- collinearity, Ridge Regression can provide coefficient estimates that can better predict new data.

The results calculated in the previous section show that the Ridge Regression model provided a lower RMSE value than the OLS model – meaning Ridge Regression did a better job of predicting the prices for observations in the test set.

6. Conclusions

Both linear models have an extremely high RMSE of ~432,000, showing low prediction accuracy. As a result, the models are not useful or reliable for predicting the price of different units.

Linear Regression should be used to develop models for processes and datasets that can be well-described by a linear model – considering the magnitude of the error, it is possible that a linear model cannot accurately predict house prices in Melbourne. Even if this is the case, there are several steps that could be taken to potentially improve the performance of the mode.

Further Pre-Processing

Multicollinearity can lead to inaccurate predictions. To improve the model, further pre-processing could be completed to measure the multicollinearity of certain features that could then be removed.

Outliers can also be troublesome – observations that are not in line with most of the data, either in terms of response value (price) or feature values, can have a large effect on a linear model. Outliers with feature or response values over a certain threshold could be removed, to remove their effect on the model's coefficients.

Applying Search Strategies

Search strategies like adding different features in a stepwise manner and evaluating their effect on the RMSE of the model could be attempted to construct a better model.

Citations

[1] Stigler, S.M. (1978) "Mathematical Statistics in the Early States," *The Annals of Statistics*, Vol. 6, pp. 239-265.