Gaby Masak

D600 – Statistical Data Mining

Task 1: Linear Regression Analysis

10/2/2024

**Part 1: GitLab Repository URL**

The GitLab repository for this task can be found following this link:

<https://gitlab.com/wgu-gitlab-environment/student-repos/gmasak/d600-statistical-data-mining/-/blob/gmasak-main-patch-04058/D600%20Task%201%20Repository/D600Task1.py>

The repository history is extremely limited, noting that technical difficulties with both building a pipeline and generating a token from the WGU GitLab led to the majority of the coding to be developed before the initial commit could be made. Eventually the WGU technical team was able to resolve these issues, but forgot to make me aware.

**Part 2: Purpose of Linear Regression Analysis**

Data analytics has many practical applications in every industry. For the real estate industry, linear regression analysis can be used to predict housing prices. It is common knowledge that various features can influence pricing, but exactly which predictors can be very locally driven. For example, one could ask the age-old question—does size matter? Linear regression can be used to quantify this question and make future estimates, provided the right set of variables are selected, appropriate analysis steps are taken, and appropriate training and test samples are derived.

**Part 3: Summary of Data Preparation Process**

Five variables were selected to calculate how size affects housing prices: price, square footage, number of bathrooms, number of bedrooms, and backyard space, as they numerically scale to represent different areas in which a house can be larger. Further justification is lent due to numeric variables being essential for prediction and technical techniques in linear regression analysis.

Table 1: Descriptive Statistics of Variables

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Type | Count | Mean | Mode | Range | Max/Min |
| Price | Dependent | 7000 | 307281.97 | N/A | 961675.64 | 1046675.64/85000 |
| Square Footage | Independent | 7000 | 1048.95 | 550 | 2324.7 | 2824.7/550 |
| Number of Bathrooms | Independent | 7000 | 2.1314 | 1 | 4.81 | 5.80723/1 |
| Number of Bedrooms | Independent | 7000 | 3.0086 | 3 | 6 | 7/1 |
| Backyard Space | Independent | 7000 | 511.51 | 300, 418.29, 516.29 | 1630.97 | 1631.36/0.39 |

Univariate and Bivariate Distributions of the Dependent and Independent Variables:

Price

A diagram of a graph

Description automatically generated

Figure 1: Boxplot of Price

A graph of a graph

Description automatically generated with medium confidence

Figure 2: Histogram of Price

A graph of a function

Description automatically generated

Figure 3: Density Plot of Price

Square Footage

A diagram of a box plot

Description automatically generated

Figure 4: Boxplot of Square Footage

A graph of a function

Description automatically generated

Figure 5: Density Plot of Square Footage

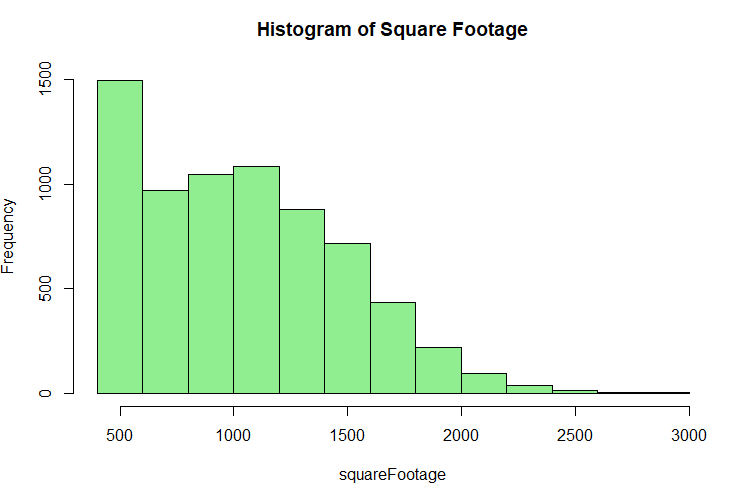


Figure 6: Histogram of Square Footage

Number of Bathrooms

A diagram of a bathroom

Description automatically generated

Figure 7: Boxplot of Number of Bathrooms

A graph of a function

Description automatically generated

Figure 8: Density Plot of Number of Bathrooms

A graph of a bathroom

Description automatically generated

Figure 9: Histogram of Number of Bathrooms

Number of Bedrooms

A diagram of a box plot

Description automatically generated

Figure 10: Boxplot of Number of Bedrooms

A graph of a function

Description automatically generated

Figure 11: Density Plot of Number of Bedrooms

A graph of a number of bedrooms

Description automatically generated

Figure 12: Histogram of Number of Bedrooms

Backyard Space

A diagram of a box plot

Description automatically generated

Figure 13: Boxplot of Backyard Space

A graph of a plot

Description automatically generated

Figure 14: Density Plot of Backyard Space

A graph of a number of green bars

Description automatically generated

Figure 15: Histogram of Backyard Space

Price vs. Square Footage

A blue and white diagram

Description automatically generated

Figure 16: Scatterplot of Price vs Square Footage

Price vs. Number of Bathrooms

A blue dot diagram with white text

Description automatically generated

Figure 17: Scatterplot of Price vs Number of Bathrooms

Price vs. Number of Bedrooms

A graph of blue lines

Description automatically generated

Figure 18: Scatterplot of Price vs. Number of Bedrooms

Price vs. Backyard Space

A blue dot diagram with white text

Description automatically generated

Figure 19: Scatterplot of Price vs. Backyard Space

**Part 4: Analysis and Results**

The 7000 rows were separated randomly into a training set with 5600 rows and a testing set with 1400 rows. These sets can respectively be found in the attached files D600TrainingSet.csv and D600TestSet.csv. After separating the data into training and testing sets, the data can be analyzed.

A screenshot of a computer

Description automatically generated

Figure 20: Screenshot of Regression Results



Figure 21: Screenshot of the Mean Square Errors

**Part 5: Summary of Data Analysis**

1. Libraries used in Python and R
   1. Python:
      1. pandas: Used to analyze and manipulate large datasets.
      2. numpy: Used to manipulate arrays and matrices.
      3. statsmodels.api: Used for classes and functions of various statistical methods for statistical tests and data exploration.
         1. ols:
      4. sklearn.model\_selection: Used for model selection, analyzing data, training and testing models.
         1. train\_test\_split: Used to randomly split dataframe into training and testing sets, specifying the percentage of test data, and grounding the results.
      5. sklearn.linear\_model: Used to predict values using a linear regression model between dependent and independent variables.
         1. LinearRegression: Used to apply ordinary least squares (OLS) linear regression, minimizing the sum of squares.
      6. sklearn.metrics: Used to calculate various prediction errors.
         1. mean\_squared\_error: Used to calculate the mean squared error between real and predicted values.
      7. statsmodels.stats.stattools:
         1. durbin\_watson:
   2. R:
      1. readxl: Used to import data from Excel and CSV files.
      2. forcats: Used to handle categorical variables.
      3. dplyr: Used to manipulate data by transforming, filtering, and summarizing.
      4. ggplot2: Used to create visualizations like boxplots, density graphs, etc.
2. Method used to Optimize Model and Justification

In order to properly use linear regression for analysis, it is important to make sure that data is not under- or overfitted. First, it is important to make sure that there are enough data points, of which the 7000 included in the dataset for this task are more than enough. Secondly, for a model with multiple variables, it is important to optimize the data by only including relevant inputs, otherwise overfitting can occur. When visually observing the variables, it is clear that a positive correlation exists between price and square footage, number of bathrooms, and number of bedrooms, however, Figure 19 displays a radial distribution of data that is seemingly random.

Optimizing the model is an effective means of weighing whether each variable should be included. This was achieved by forward stepwise selection which ultimately concluded that backyard space was not necessary to forecast price. Forward stepwise selection involves iteratively adding variables to the model one at a time based on their statistical significance. At each step, the variable that improves the model the most is added, and this process continues until no further significant improvement is observed. This method helps in identifying the most relevant predictors while avoiding overfitting by excluding variables that do not contribute meaningfully to the model’s predictive power. By using forward stepwise selection, the model becomes more robust and interpretable, ensuring that only the most impactful variables (square footage, number of bathrooms, and number of bedrooms) are included in the final analysis.

1. Verification of Assumptions Used to Create the Optimized Model

Several assumptions must be met while optimizing a linear regression model for validity and reliability: linearity, independence of residuals, homoscedasticity, normality of residuals, lack of multicollinearity, and absence of endogeneity. Tackling the assumption of linearity between the independent and dependent variables is verified visually in Figures 16, 17, and 18, where the scatterplots display a positive linear correlation between price and the respective independent variable. Secondly, the assumption that residuals are independent of each other can be verified by use of a Durbin-Watson test, which specifies a result of 1.973, suggesting that no significant autocorrelation exists as the value is close to two. Thirdly, the assumption of homoscedasticity, which means that the variance of the residuals is constant across all levels of the independent variables, can be addressed by examining residual plot in Figure 22 for any patterns that suggest non-constant variance. The seemingly random distribution that is more clustered and banded instead of being described as curved or funnel shaped.

A blue dotted diagram with numbers and a white background

Description automatically generated with medium confidence

Figure 22: Residuals vs. Fitted Values

Normality of residuals can be checked using Q-Q plots, in Figure 23, or statistical tests like the Shapiro-Wilk test. It can be noted that in Figure 23 a slight curve is observed, which indicates that the variables used are skewed, which is expected as variables number of bedrooms, number of bathrooms, square footage, and price all skewed right. Otherwise, the Q-Q plot does not notable deviate.

A graph with a line and a red line

Description automatically generated

Figure 23: Q-Q Plot

Lack of multicollinearity can be assessed using Variance Inflation Factor (VIF) values to ensure that the independent variables are not highly correlated with each other. Values larger than 5 are said to be highly correlated. As seen in Figure 24, ignoring the constant value calculated by the regression equation, the variables square footage, number of bathrooms, and number of bedrooms are moderately correlated, but do not surpass the value 5.

A screenshot of a computer

Description automatically generated

Figure 24: Matrix of VIF Values

Finally, ensuring no endogeneity involves making sure that there is no correlation between the independent variables and the error term, which can be addressed by including all relevant variables in the model. For the purposes of determining whether and how size affects housing prices, the variables selected was sufficient. However, if the goal is to find which variables out of the 21 variables provided affect the price, all variables would need to be presented initially, with optimization sorting which to include or exclude.

1. Provide the Regression Equation and Discuss the Coefficient Estimates

The Regression Equation can be viewed in Figure 25 below:



Figure 25: Screenshot of Regression Equation

Price is predicted in the regression equation by using 3 predictor variables. The weight of each predictor variable on price is expressed by its coefficient estimate. This can be explained by examining square footage, where the coefficient variable is 159.50, meaning that for every extra square foot of living space a property has, the price increases by $159.50. Notably, the number of bedrooms and bathrooms largely affect the price, leading to respective increases of $56,031.80 and $58,172.55 for each additional bedroom and bathroom.

1. Discuss the Model Metrics:
   1. R2 and adjusted R2 of the training set

As seen in Figure 20, the R² and adjusted R² of the training set are 0.596 and 0.595 respectively. It can be inferred that the model explains approximately 59.6% of the variance in the dependent variable, indicating a moderate level of explanatory power. The small difference between R² and adjusted R² suggests that the model is not overly complex and that the predictors included are relevant and contribute meaningfully to the model’s performance. However, there is still a significant portion of the variance that is unexplained, indicating that there may be other factors influencing the dependent variable that are not captured by the current model. As previously mentioned, in answering the specific organizational goal of the linear regression model, this suggests that while size does matter, but other factors should ultimately be considered too for more accurate analysis.

* 1. Comparison of the MSE for the training set to the MSE of test set

As seen in Figure 21, the MSE values of the training and the test sets are 9257650915.837856 and 9296492933.70988 respectively. This means that the model performs consistently on both the training and test datasets, indicating that it has not overfitted to the training data. The close similarity in MSE values suggests that the model generalizes well to new, unseen data, providing reliable predictions. However, the relatively high MSE values also indicate that there is room for improvement in the model’s accuracy, possibly by incorporating additional relevant features or by fine-tuning the existing model parameters. One possible measure is normalizing price to numbers between 0 and 1 before applying linear regression models and optimization, as the values in the hundreds of thousands of dollars contrast those of ones or thousands of the independent variables.

1. Discuss Results and Implications for Prediction Analysis

The goal of the linear regression analysis was to determine whether size matters. After optimizing the model by removing backyard space as a predictor variable and applying ordinary least squares regression analysis to find a regression equation that could predict home prices, it can be determined that size (square footage, number of bathrooms, number of bedrooms) does matter. It should also be noted that size is not the only metric that matters in terms of predicting price, as discussed when observing the R² values. If the goal is to accurately predict the price wholistically, all variables should be presented, with optimization determining which to ultimately keep.

1. Recommend Course of Action for Real-World Organization

In Economics, value is observed as being the highest price a buyer is willing to pay for an item. In real estate, this concept is especially important as current prices are not always easy to estimate as the pool of demand is limited, and previous data provides a better opportunity for analysis. Listing a house at an appropriate price is imperative as listing too high will result in poor demand, with the listing growing stale and affecting future potential, or listing too low, where the seller risks leaving money on the table and not realizing the full potential value of their property. Therefore, accurate pricing strategies are crucial to balance demand and maximize returns. That said, this knowledge can also be applied to more than simply estimating price. For example, the analysis completed above would be useful for a developer that is planning a new development. The analysis would show the developer that backyard space is not highly valued and does not affect price leading the developer to design higher-density homes with smaller lot sizes. The developer may also choose to install more bedrooms and bathrooms, which are sized smaller to accommodate more within the same square footage, in order to maximize their potential selling prices. Ultimately, the real estate data is invaluable for making informed decisions that align with market demands and optimize profitability. By leveraging such insights, developers and agents can strategically plan their projects and listings to meet buyer preferences and maximize returns on investment.

**Part 6: Panopto Audiovisual Presentation**

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=7074ba17-7ca2-436b-8c88-b20b014f60a0