Multiple Linear Regression on King County House Prices Dataset

Shreyansh Misra, Maxwell Owens, Luke Pasterczyk, Akhila Garre, and Maithili Revankar

2025-03-12

Introduction

Multiple Linear Regression (MLR) is a statistical method used to estimate the relationship between a dependent variable and multiple independent variables, assuming a linear relationship between them.

The MLR equation is given by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$$

where β_0 is the intercept, and β_1 and β_2 are the coefficients determined by minimizing the sum of the squared differences between observed and predicted values.Once fitted, the model can be used to predict Y for given X_n values.

This study focuses on predicting home prices in King County, Washington using Multiple Linear Regression. The dataset, sourced from Kaggle (www.kaggle.com/datasets/shivachandel/kc-house-data), contains 21,613 records of homes sold between May 2014 and May 2015. We specifically analyzed nine predictors to estimate house prices.

Our investigation leveraged several R packages for data preprocessing, visualization, and modeling. `ggplot2` was used for data visualization, while `tidyr` and `dplyr` helped with data transformation. The `caret` package was used machine learning workflows, and `car`, `lmtest`, and `nlme` were used for regression and linearity testings.

We find that the logarithmic function of 8 of these predictors are a very strong linear predictor of the logarithmic function of price, expressed by:

$$log(Y) = 6.3611 - 0.1871 * log(\text{bedrooms}) - 0.1406 * log(\text{bathrooms}) + 0.5216 * log(\text{sqft\_living}) - 0.0527 * log(\text{sqft\_lot})$$
$$+0.3705 * log(\text{waterfront}) + 0.1534 * log(\text{view}) + 0.3691 * log(\text{condition}) + 1.4207 * log(\text{grade})$$

Data description

```
housing.df <- read.csv("kc_house_data.csv")
num_rows <- nrow(housing.df)
sum_missingdata <- sum(is.na(housing.df))
cat("Number of Rows: ", num_rows, "   Rows with Missing Data: ", sum_missingdata)
```

```
## Number of Rows:  21613    Rows with Missing Data:  0
```

Dataset

his dataset contains house sale prices for King County, which includes Seattle.

The dataset consists of house prices from King County an area in the US State of Washington, which also covers Seattle. It includes homes sold between May 2014 and May 2015. There are 10 variables and 21613 observations, of which 9 are features for the target house sales price. From an initial analysis, there were no missing data points.

```
head(housing.df, 3)
```

	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition
	<dbl>	<int>	<dbl>	<int>	<int>	<dbl>	<int>	<int>	<int>
1	221900	3	1.00	1180	5650	1	0	0	3
2	538000	3	2.25	2570	7242	2	0	0	3
3	180000	2	1.00	770	10000	1	0	0	3

3 rows | 1-10 of 11 columns

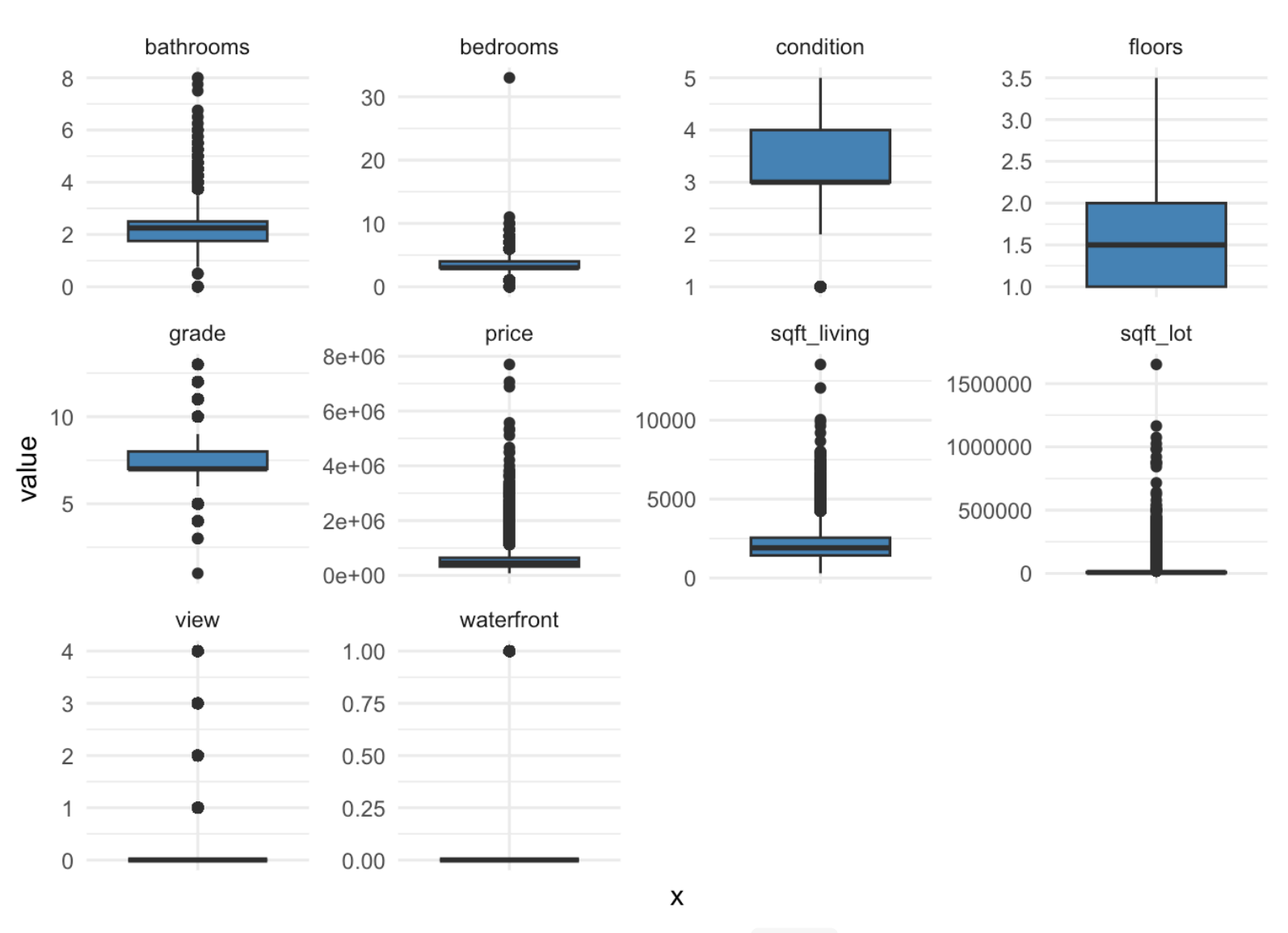
Variables

- price: Price of house sale in currency of USD
- bedrooms: Number of bedrooms
- bathrooms: Number of Bathrooms, where 0.5 represents a bathroom with a toilet but with no shower
- sqft_living: Square footage of the apartments interior living space
- sqft_lot: Square footage of the land space
- floors: Number of floors
- waterfront: An index to indicate if the house was overlooking the waterfront or not. 0 represents no waterfront, 1 represents with waterfront.
- view: An index from 0 to 4 of how good the view of the property was. 0 represents no good view, 4 represents very good view.
- condition: An index from 1 to 5 on the condition of the house. 1 represents poorer condition, and 5 represents superb condition.

We can identify `price` as our dependent variable as the median value of homes in the neighborhood is what we are predicting. The remaining 9 variables are our independent variables.

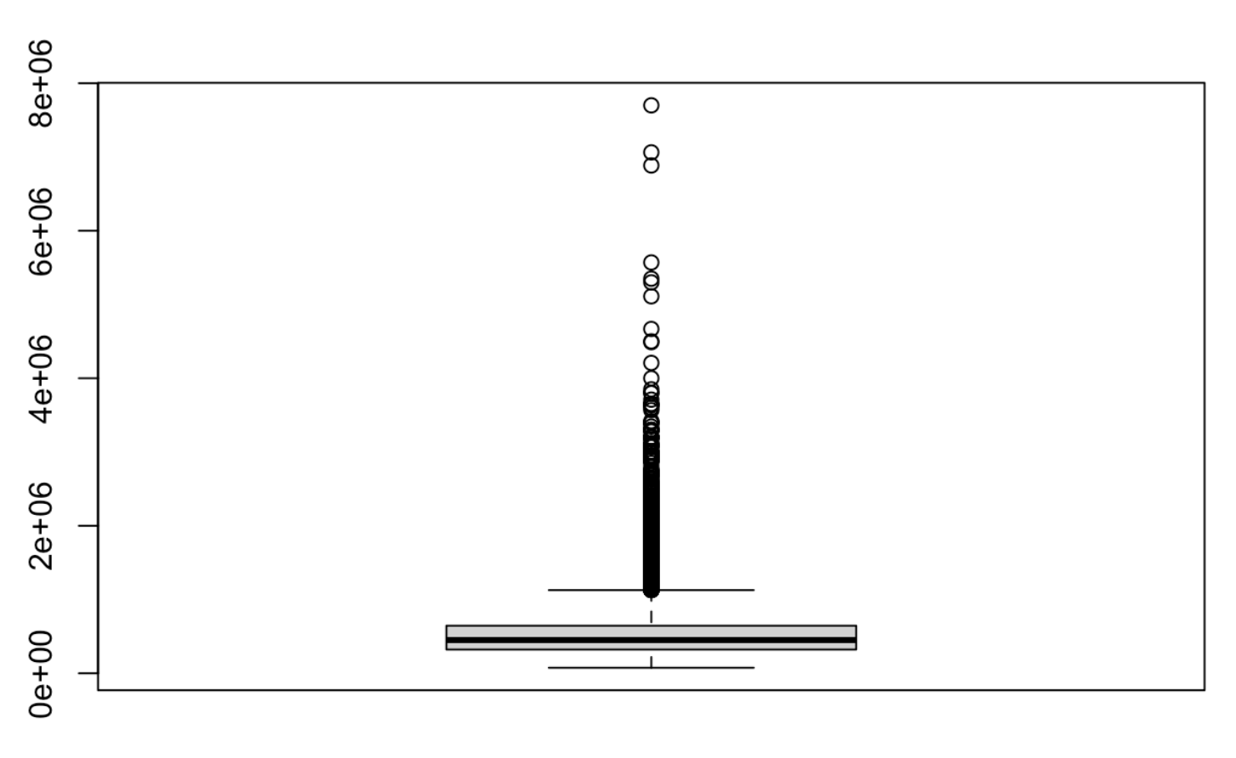
Outlier Detection

```
housing_box <- housing.df %>%
  pivot_longer(cols = everything(), names_to = "variable", values_to = "value")
ggplot(housing_box, aes(x = "", y = value)) +
  geom_boxplot(fill = "steelblue") +
  theme_minimal() +
  facet_wrap(~ variable, scales = "free_y") +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())
```



There are evidently outliers in the dataset, and we will primarily concern ourselves with outliers in `price` as that is out dependent variable.

```
boxplot(housing.df$price)
```



Outlier detection and removal was performed utilizing the IQR method.

```
q1 <- quantile(housing.df$price, 0.25)
q3 <- quantile(housing.df$price, 0.75)
IQR = IQR(housing.df$price)
outliers <- subset(housing.df, housing.df$price < (q1 - (1.5 * IQR)) | housing.df$price > (q3 + (1.5 * IQR)))
num_outliers <- nrow(outliers)
housing.df <- subset(housing.df, !(rownames(housing.df) %in% rownames(outliers)))
new_ds <- nrow(housing.df)
cat("The dataset with outliers removed now has ", new_ds, "rows.")
```

```
## The dataset with outliers removed now has  20467 rows.
```

The final transformation applied to the dataset involved passing it through a logarithmic function. This transformation was chosen to address skewness in the data and stabilize variance, making it more suitable for MLR modeling in the later stages of this investigation.

```
housing.df <- log(housing.df+1)
```

Analysis

Train and Test Split

Our dataset was split into a training and testing set, where approximately 75% of the dataset is used for training and the remaining 30% is used for testing. Specifically, 15352 were used for training and were reserved for testing 5115.

```
set.seed(123)
split <- 0.75
trainIndex <- createDataPartition(housing.df$price, p = split)
trainIndex <- unlist(trainIndex)
train <- housing.df[trainIndex, ]
test <- housing.df[-trainIndex, ]
num_row_train <- nrow(train)
num_row_test <- nrow(test)
cat("Number of Rows in Train Set: ", num_row_train, "   Number of Rows in Test Set: ", num_row_test)
```

```
## Number of Rows in Train Set:  15352    Number of Rows in Test Set:  5115
```

```
train_control <- trainControl(method="cv", number=10)
model <- train(price ~ . , method="lm", data = train, trControl=train_control)
print(model$results)
```

```
##   intercept      RMSE  Rsquared       MAE      RMSESD RsquaredSD       MAESD
## 1      TRUE 0.3229572 0.4767439 0.2614558 0.006095231 0.01972628 0.005047197
```

10-fold cross-validation was employed in this investigation. It is significantly faster than LOOCV and still gives a good idea of how well the model works, so it's the best choice for saving time without losing accuracy.

```
summary(model)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.17948 -0.23694  0.02882  0.22244  1.81338
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.351632   0.075344  84.302  < 2e-16 ***
## bedrooms    -0.185340   0.016742 -11.070  < 2e-16 ***
## bathrooms   -0.147126   0.017764  -8.282  < 2e-16 ***
## sqft_living  0.520377   0.013408  38.811  < 2e-16 ***
## sqft_lot    -0.051084   0.003388 -15.078  < 2e-16 ***
## floors       0.023463   0.016302   1.439     0.15
## waterfront   0.366014   0.072113   5.076 3.91e-07 ***
## view         0.154339   0.008812  17.516  < 2e-16 ***
## condition    0.374862   0.019485  19.239  < 2e-16 ***
## grade        1.411413   0.032339  43.644  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3229 on 15342 degrees of freedom
## Multiple R-squared:  0.477,  Adjusted R-squared:  0.4767
## F-statistic:  1555 on 9 and 15342 DF,  p-value: < 2.2e-16
```

All the variables except `floor` have p-values > 0.05 making them signficiant predictors of `price`. As floor does not have a p-value > 0.05 it can be removed from our final model as it does not influence the price of homes in Kings County to an extent that can deem it signigicant. The remaining 8 predictors were included in the final mode;=l.

```
model <- train(price ~ bedrooms + bathrooms + sqft_living + sqft_lot + waterfront + view + condition + grade, met
hod="lm", data = train, trControl=train_control)
summary(model)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.18268 -0.23737  0.02886  0.22170  1.82620
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.361078   0.075060  84.747  < 2e-16 ***
## bedrooms    -0.187107   0.016698 -11.206  < 2e-16 ***
## bathrooms   -0.140575   0.017171  -8.187 2.89e-16 ***
## sqft_living  0.521637   0.013380  38.986  < 2e-16 ***
## sqft_lot    -0.052654   0.003208 -16.414  < 2e-16 ***
## waterfront   0.370458   0.072049   5.142 2.76e-07 ***
## view         0.153408   0.008788  17.456  < 2e-16 ***
## condition    0.369082   0.019067  19.357  < 2e-16 ***
## grade        1.420733   0.031685  44.839  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3229 on 15343 degrees of freedom
## Multiple R-squared:  0.4769, Adjusted R-squared:  0.4766
## F-statistic:  1748 on 8 and 15343 DF,  p-value: < 2.2e-16
```

We can further discuss the predictive abilities of this model after validating the assumptions of linearity for multiple linear regression (MLR). However, quickly analyzing the model reveals that it explains approximately 47.69% of the variance in the response variable, as indicated by the R-squared value of 0.4769. The formula for the fitted regression model is:

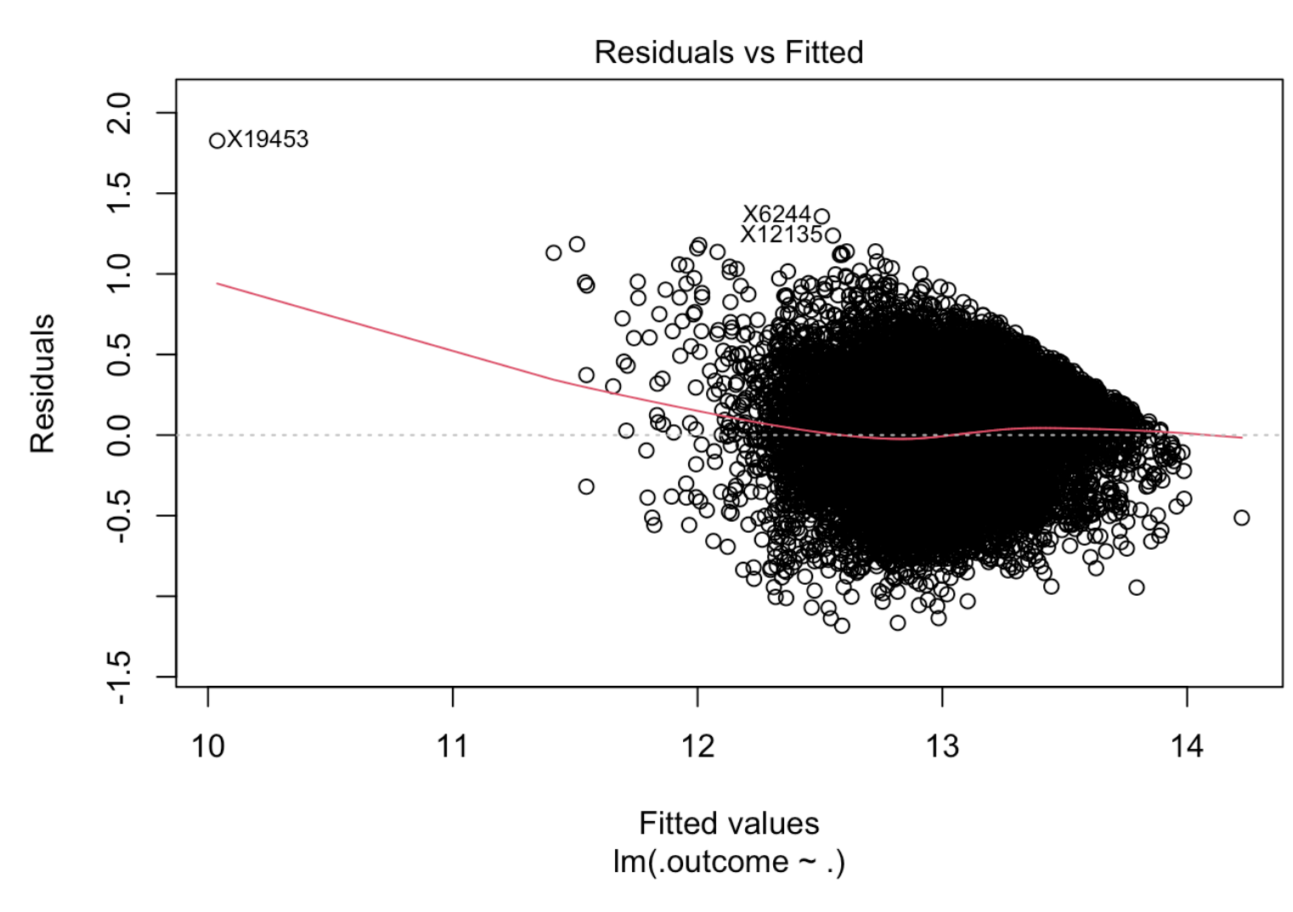
$$log(Y) = 6.3611 - 0.1871 * log(\text{bedrooms}) - 0.1406 * log(\text{bathrooms}) + 0.5216 * log(\text{sqft\_living}) - 0.0527 * log(\text{sqft\_lot})$$
$$+0.3705 * log(\text{waterfront}) + 0.1534 * log(\text{view}) + 0.3691 * log(\text{condition}) + 1.4207 * log(\text{grade})$$

Assumptions of Multiple Linear Regression

1. Linearity - the relationship between the independent variables and dependent variable should be linear
2. Independence Of Errors - each data point's error should be independent of other points' errors (no observation should influence another)
3. Homoscedasticity - variance of the errors remains consistent across all values of the independent variable
4. Normality Of Errors - errors are normally distributed
5. Multicollinearity - whether independent variables in a linear regression equation are correlated

Assumption 1: Linearity

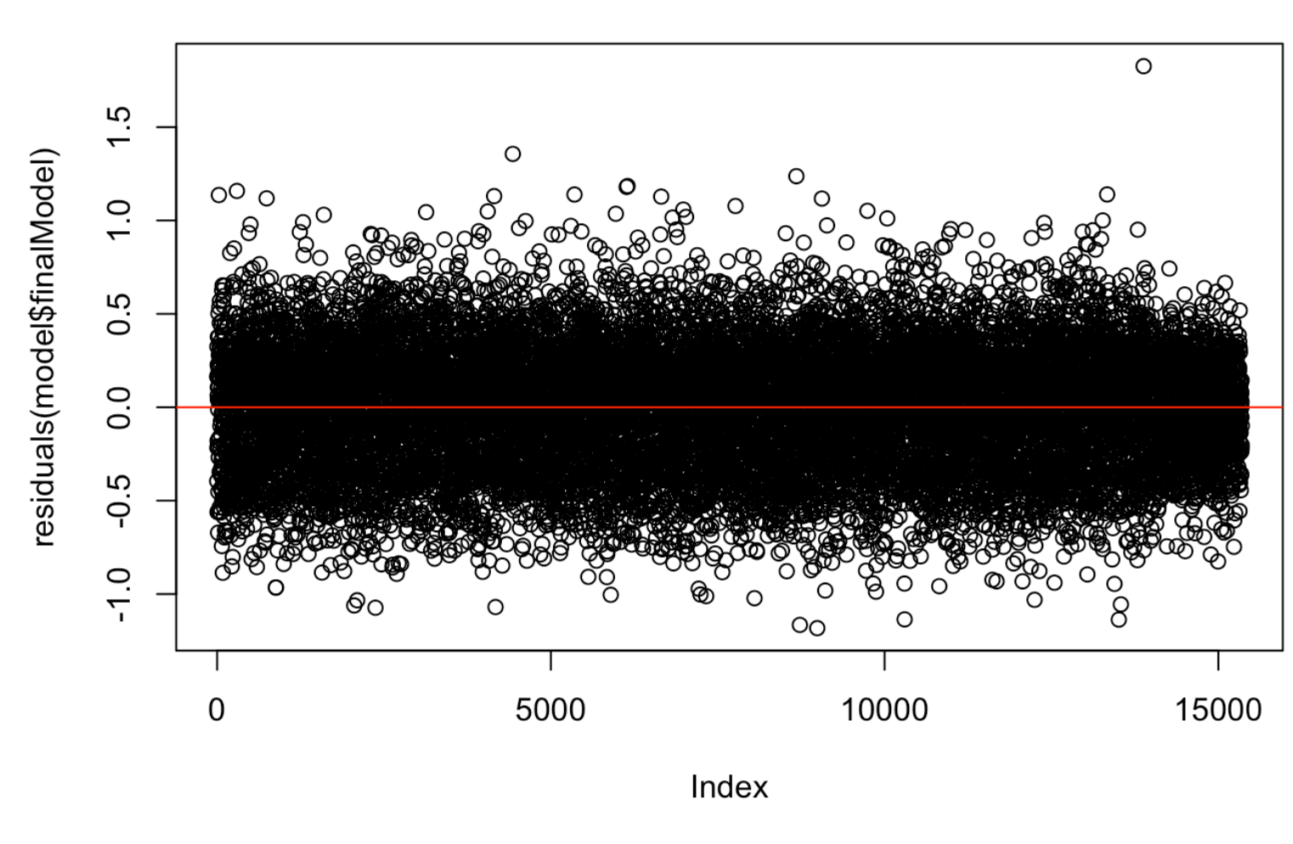
```
plot(model$finalModel, 1)
```



The Residual vs Fitted graph displays a random pattern with red line at 0 (given the residual range is less than [-1,1]). This indicates linearity.

Assumption 2: Independence Of Errors

```
plot(residuals(model$finalModel))
abline(h = 0, col='Red')
```



Our plot indicates the residuals are randomly scattered and centered around the horizontal line, indicating that the residuals are approximately equal to zero. We can also more formally verify that our errors are independent with a Durbin-Watson test. Given the large p-value, we fail to reject the null hypothesis. The autocorrelation is 0, or the errors are independent.

```
dwtest(model$finalModel)
```

```
##
##  Durbin-Watson test
##
## data:  model$finalModel
## DW = 1.996, p-value = 0.4008
## alternative hypothesis: true autocorrelation is greater than 0
```

Assumption 3: Homoscedasticity

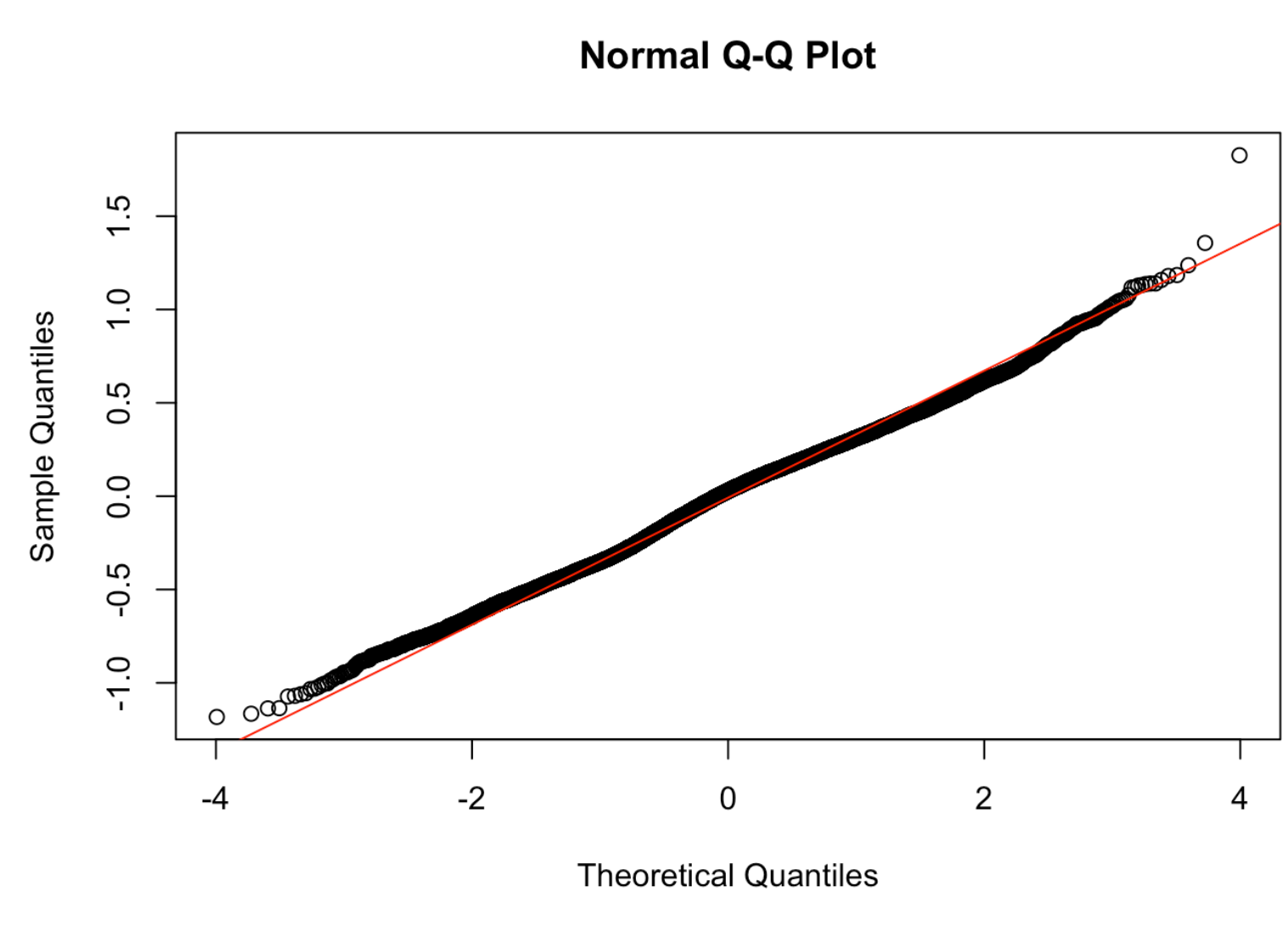
```
bptest(model$finalModel)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model$finalModel
## BP = 644.58, df = 8, p-value < 2.2e-16
```

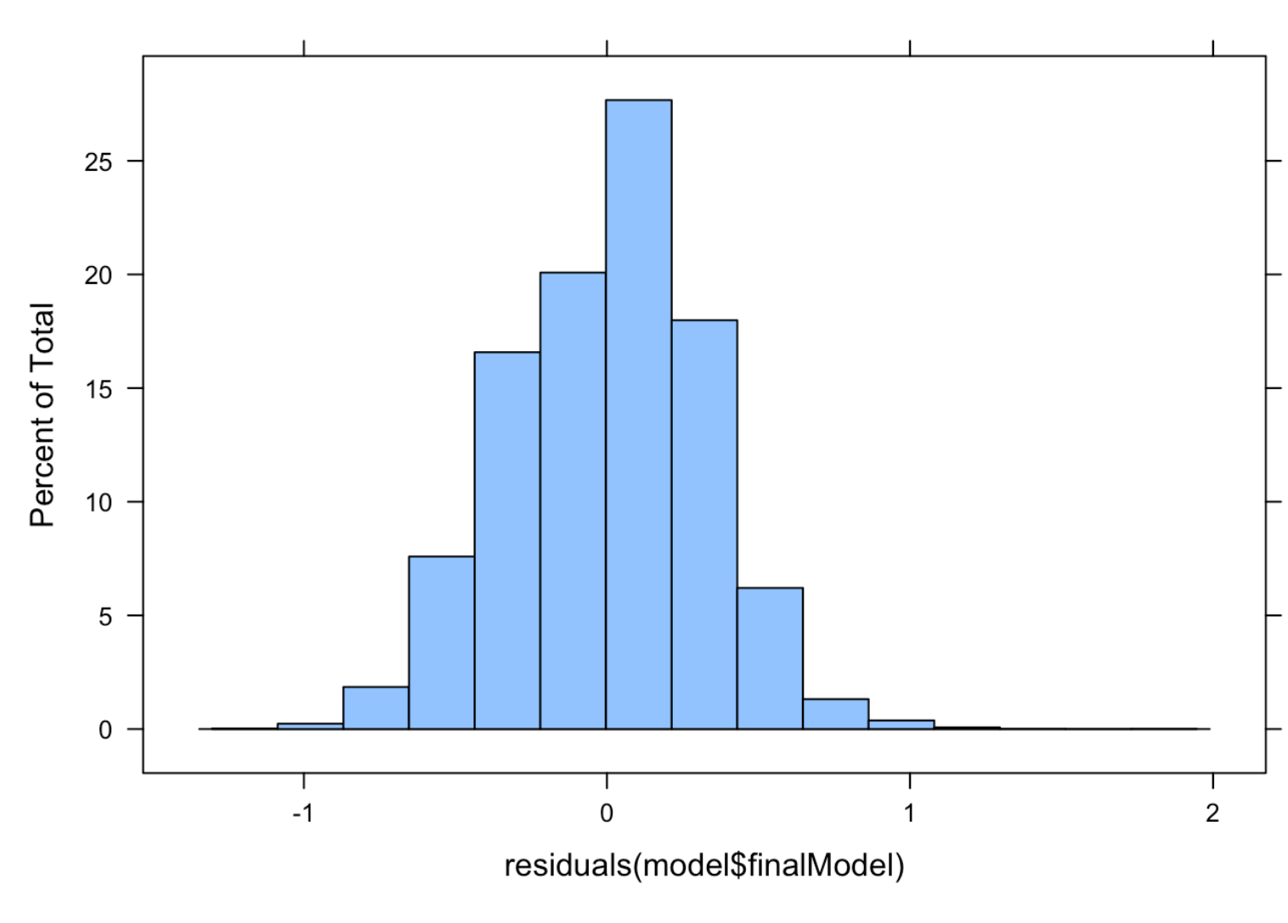
The Breusch-Pagan test gave a BP value of 644.58 and a p-value < 2.2e-16, meaning we fail the test for homoscedasticity. This happens because housing prices naturally have more variation at higher values, which is normal for pricing data. Since we already applied a log transformation, this isn't a big issue.

Assumption 4: Normality Of Errors

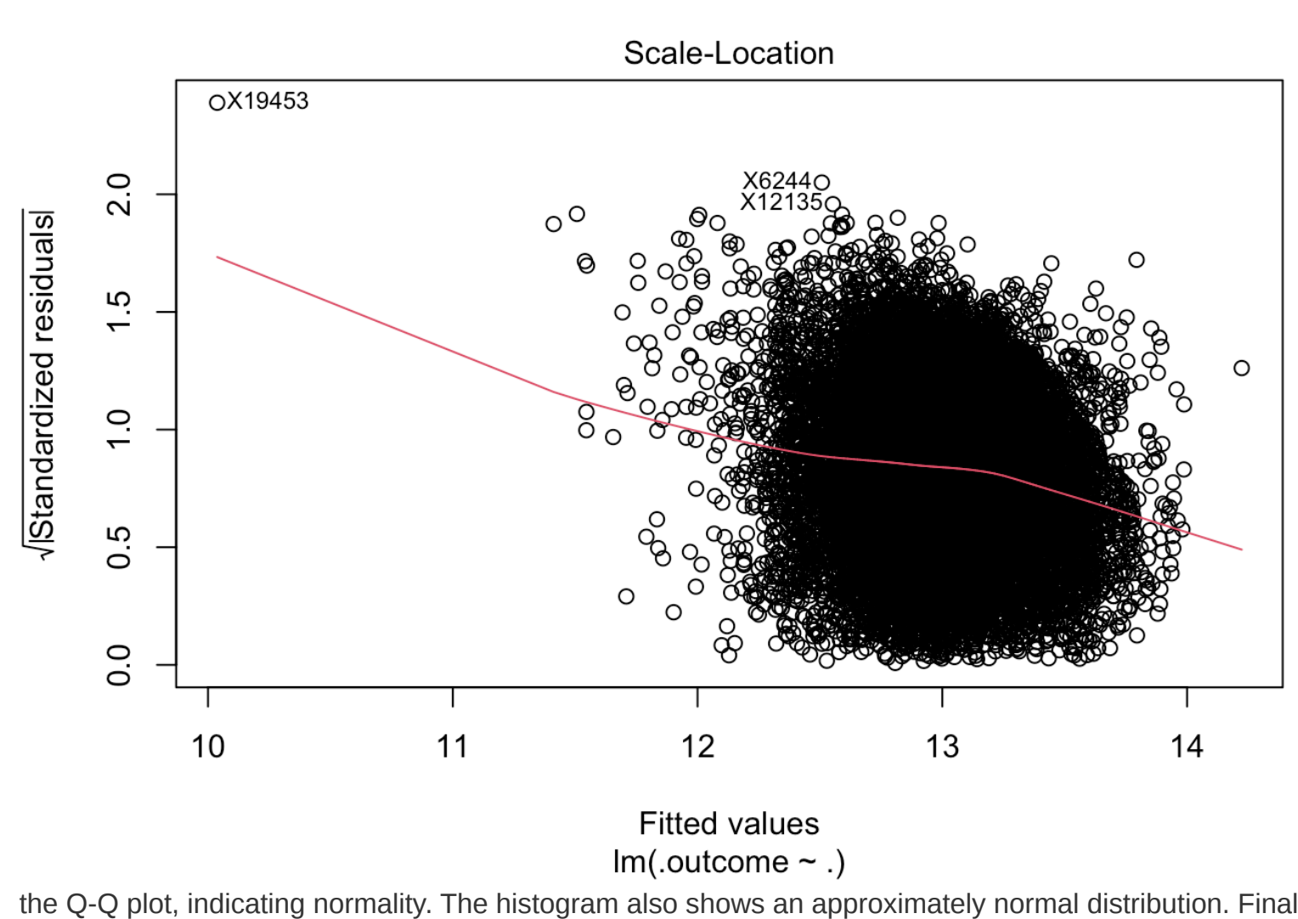
```
qqnorm(residuals(model$finalModel))
qqline(residuals(model$finalModel), col='Red')
```



```
histogram(residuals(model$finalModel), prob = TRUE)
```



```
plot(model$finalModel, 3)
```



The residuals closely follow the line in the Q-Q plot, indicating normality. The histogram also shows an approximately normal distribution. Finally, the Residuals vs. Fitted plot shows the residuals are evenly scattered around the red line, suggesting constant variance.

Assumption 5 - Multicollinearity

```
vif(model$finalModel)
```

```
##    bedrooms   bathrooms sqft_living    sqft_lot  waterfront        view
##    1.813609    2.552534    4.163795    1.196894    1.073380    1.109652
##   condition       grade
##    1.048170    2.180115
```

All VIF values for the parameters are below 5, which means there is no significant multicollinearity. Variables such as waterfront, view, and condition are close to 1 measning they are not strongly correlated with each other.

Validity

The model largely passed the 5 assumptions of linearity for MLR. Which Assumption #3 did not formally pass the testing, we can still consider our model linear as housing prices naturally have more variation at higher values. This does not significantly affect our modelling or prediction.

Model Evaluation and Prediction

To evaluate the performance of our model, we assessed it based on the five assumptions for linearity in MLR. Our model passed all tests for linearity, confirming that the relationship between predictors and the outcome variable is appropriately modeled using a linear approach. The Breusch-Pagan test indicated the presence of heteroscedasticity, which is common in pricing datasets, but the log transformation applied earlier mitigates its impact.

```
predictions <- predict(model, newdata = test)
actual <- test$price

mae <- mean(abs(predictions - actual)) # MAE
mse <- mean((predictions - actual)^2) # MSE
rmse <- sqrt(mse) # RMSE

cat("MAE: ", mae, "   MSE: ", mse, "   RMSE: ", rmse)
```

```
## MAE:  0.2572857    MSE:  0.1005026    RMSE:  0.3170215
```

In terms of model accuracy, we evaluated the residual statistics and key error metrics. These values indicate a reasonable predictive performance, though there is some variability in the residuals. The multiple R-squared value of 0.4769 suggests that approximately 47.7% of the variance in the outcome variable is explained by our predictors.

For prediction, we applied the model to the training dataset and obtained reliable estimates. However, additional validation using a test dataset or cross-validation could help provide a better understanding of whether the model is generalizable.

Conclusion

Our multiple linear regression model effectively predicts the outcome variable based on eight key predictors. The model passed linearity tests and demonstrated moderate explanatory power with an R-squared of 0.4769. The log transformation improved the linearity and distribution of residuals, ensuring a more accurate fit.

The model provided meaningful insights into how different factors influence the outcome variable. Key predictors such as `sqft_living`, `grade`, and `view` have significant positive impacts on price. However, the presence of heteroscedasticity suggests variability in residuals, which could impact predictive consistency. Furthermore, an R-squared of 0.4769 means over 50% of the variance remains unexplained.

Overall, while our model performs well within its scope, some future refinements can improve accuracy and generalizability.

References

- https://www.kaggle.com/datasets/shivachandel/kc-house-data
- https://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/