## 1. Introduction

The dataset that I am exploring is for an analysis of statistical value for housing data. The different variables such as number of bedrooms and age of the house will influence sale price for the homes. There will be three types of analysis being performed. The first analysis will be the first order regression using quantitative and qualitative variables. The second analysis will be the regression models using quantitative variables and the last analysis will be using a nested model of the second order regression.

## 2. Data Preparation

The variables that are deemed of importance are age of home; price; square footage; number of bathrooms; view; the house grade; the age of appliances; and the crime rate within its location. There are a total of 22 columns and 2,692 rows within the dataset.

## 3. Model #1 - First Order Regression Model with Quantitative and Qualitative Variables

### Correlation Analysis

The scatterplot of price vs living area:



The scatterplot of price vs age of the homes:



The first model shows a trend that is positive for living space and price of the home. In general, if the living is of the house increases than the price of the house will increase as well. The second model shows no trend for price of home vs the age of the home.

From the data provided it shows that the correlation coefficient for the living is 0.6895 while the age of the house is -0.0746.

The strength between the living area and the age of the house correlation coefficients shows a modest positive correlation for the living area since it is between 0.40 and 0.80. The correlation for age of house is negative since it is between 0.0 and 0.40.

### Reporting Results

The equations for multiple regression model while using the price as the response variable and the grade of the house, living area, number of bathrooms and views as the predictors are as followed:

The will be the intercept. The values are regression terms for living area, grade of the home, number of bathrooms, view1 and view2. The terms are the values for the variables living area, grade of the home, number of bathrooms, dummy variable for view = mountains and dummy variable for view = forest.

The model equation for the variables is as followed:

The is and is 0.6469. It shows that 65% of the variation in price can be expressed by using a model that uses living area, grade of the home, number of bathrooms, and dummy variables for view1 being mountains and view2 being forest as predictor variables.

The beta estimate for living area is . The beta estimate for view2 is .



**

The plot model that contains the fitted values shows homoscedasticity as well as linearity.

### Evaluating Significance of Model

The equation to shows a significant 5% level of significance is to perform an F-test and to use the null and alternative hypothesis as followed:

The null hypothesis means that the beta estimates are equal to zero and the model is not significant. The alternative hypothesis means that the beta estimate does not equal to zero and the model is significant.

The P-value for the model is . It shows that it is lower than the level of significance of 5% or 0.05. The null hypothesis will be rejected, and the alternative hypothesis will be used.

One of the predictors has a slope coefficient that is different than 0 and therefore a relationship exists between at least one of the predictor variables and the response variable in this model.

The equation to shows a significant 5% level of significance is to perform a T-test and to use the null and alternative hypothesis as followed:

|  |
| --- |
| **The P-value for sqft\_living is**   |
|  |
|  |
|  |
| **The P-value for grade is**  |
|  |
|  |
|  |
| **The P-value for bathrooms is**  |
|  |
|  |
|  |
|  |
| **The P-value for view1 is**  |
|  |
|  |
|  |
| **The P-value for view2 is**   |
|  |

The null hypothesis means that the beta estimates are equal to zero and the model is not significant. The alternative hypothesis means that the beta estimate does not equal to zero and the model is significant.

The P-value is lower than the level of significance of 0.05. It shows that it is lower than the level of significance of 5% or 0.05. The null hypothesis will be rejected, and the alternative hypothesis will be used.

### Making Predictions Using Model

**Prediction scenario 1 (Back out to the lake):**

The equation for the hypothetical values of 2,150sq ft, grade 7, and has 3 bathrooms as followed:

The predicted price will be $630,785.70 using the hypothetical values.

The 90% interval predictions are (422684.5, 838887) for the prices of the homes. The prediction interval shows us that at least 90% of the housing price falls between the numbers of 422684.5 and 838887.

The 90% confidence intervals are (610013.7, 651557.7) for the prices of the homes. The prediction interval shows us that at least 90% of the housing price falls between the numbers of 610013.7 and 651557.7

**Prediction Scenario 2 (Back out to the road):**

The equation for the hypothetical values of 2,150sq ft, grade 7, and has 3 bathrooms as followed:

The predicted price will be $402,121.90 using the hypothetical values.

The 90% interval predictions are (194826, 609417.9) for the prices of the homes. The prediction interval shows us that at least 90% of the housing price falls between the numbers of 194826 and 609417.9

The 90% confidence intervals are (392274.8, 411969.1) for the prices of the homes. The prediction interval shows us that at least 90% of the housing price falls between the numbers of 392274.8 and 411969.1

## 4. Model #2 - Complete Second Order Regression Model with Quantitative Variables

### Correlation Analysis

Price vs age of appliances:



Price vs crime rate:



Both models show no linear pattern, and both models have curves. Although, it would be wise to use the second order model for the variables.

### Reporting Results

The equations for multiple regression model while using the price as the response variable and the grade of the house, living area, number of bathrooms and views as the predictors are as followed:

The will be the intercept. The values are regression terms for appliance, crime, age, interaction of appliances and age with crime, appliances age squared, and the crime rate squared. The are the values for the hypothetical values for crime rate squared and appliances age squared.

The model equation for the variables is as followed:

The is and is 0.8084. It shows that 81% of the variation in price can be expressed by using a model that uses appliances age and crime rate as the predictor variables.





The plot model that contains the fitted values shows homoscedasticity as well as linearity.

### Evaluating Significance of Model

The null and alternative hypothesis equation to perform an F-test are as followed:

The null hypothesis means that the beta estimates are equal to zero and the model is not significant. The alternative hypothesis means that the beta estimate does not equal to zero and the model is significant.

The P-value for the model is . It shows that it is lower than the level of significance of 5% or 0.05. The null hypothesis will be rejected, and the alternative hypothesis will be used. Therefore, a relationship exists between at least one of the predictor variables and the response variable in this model.

The equation to shows a significant 5% level of significance is to perform a T-test and to use the null and alternative hypothesis as followed:

The null hypothesis means that the beta estimates are equal to zero and the model is not significant. The alternative hypothesis means that the beta estimate does not equal to zero and the model is significant.

The P-value for appliance is . The P-value for crime is while the P-value for the interactions of appliance and crime is 0.284. The P-value for appliance age squared is . The P-value for crime rate squared is. The P-values mentioned shows that it is lower than the level of significance of 5% or 0.05. While the P-value for the interaction of appliance age and crime rate are not below the level of significance.

As mentioned before. It shows that there is a relationship between price of the home and the predictor variables of crime rate; appliances age; appliance age squared, and crime rate squared. While it also shows that there is not relationship between the interactions of crime rate and appliance age against the price of the home.

### Making Predictions Using Model

**Prediction Scenario 1(One-Year-Old Appliances):**

The equation for the price of the home of 1 year old appliances while in the area that has a crime rate of 81.02 for every 100,000 are as followed:

The predicted price will be $864,423.40 using 1 year old appliances and the area of crime rate for every 100,000 peoples

The 90% interval predictions are (711566.6, 1017280). The prediction interval shows us that at least 90% of the housing price falls between the numbers of 711566.6 and 1017280.

The 90% confidence intervals are (854109.1, 874737.7). The prediction interval shows us that at least 90% of the housing price falls between the numbers of 854109.1 and 874737.7

**Prediction Scenario 2 (15-Year-Old-Appliances):**

The equation for the price of the home of 1 year old appliances while in the area that has a crime rate of 81.02 for every 100,000 are as followed:

The predicted price will be $271,051.60 using 15-year-old appliances and the area of crime rate of 200.50 for every 100,000 peoples

The 90% interval predictions are (118454.4, 423648.8). The prediction interval shows us that at least 90% of the housing price falls between the numbers of 118454.4 and 423648.8

The 90% confidence intervals are (265846, 276257.2). The prediction interval shows us that at least 90% of the housing price falls between the numbers of 265846 and 276257.2

## 5. Nested Models F-Test

### Reporting Results

The equation for the first order model of price of appliances age and crime rate are as followed:

The is the slope. are the beta estimates for crime, appliance age, and the interaction between appliance age and crime. are the hypothetical values for appliance age and crime.

The equation for beta estimates from the model is as followed:

### Evaluating Significance of Model

The equation to perform the F-test for the null and alternative hypothesis are as followed:

The null hypothesis means that the beta estimates are equal to zero and the model is not significant. The alternative hypothesis means that the beta estimate does not equal to zero and the model is significant.

The P-value for the model is . It shows that it is lower than the level of significance of 5% or 0.05. The null hypothesis will be rejected, and the alternative hypothesis will be used. Therefore, a relationship exists between at least one of the predictor variables and the response variable in this model.

The equation to shows a significant 5% level of significance is to perform a T-test and to use the null and alternative hypothesis as followed:

The null hypothesis means that the beta estimates are equal to zero and the model is not significant. The alternative hypothesis means that the beta estimate does not equal to zero and the model is significant.

The P-value for appliance is  . The P-value for crime is . The P-value for interaction between appliance and crime is It shows that it is lower than the level of significance of 5% or 0.05. The null hypothesis will be rejected, and the alternative hypothesis will be used. Therefore, a relationship exists between the predictor variables and the response variables of the home price.

### Model Comparison

In comparison of this reduced model with the complete second order model for price using appliance age and crime. We must perform a nested model F-test.

The reduced model is a model without the terms (variables) being tested. The complete model is a model that includes terms (variables) being tested. The nested model F-test is used to test whether we can use the variables being tested and only use the reduced model.

The equation for complete second order model is as followed:

The is the slope. are beta estimates for appliance age, crime, the interaction between appliance age and crime, appliance age squared, and crime rate squared. are values for appliance age and crime to predict price based on those values. are values for appliance age squared and crime rate squared.

The equation for reduced model is as followed:

The is the slope. are beta estimates for appliance age, crime, and the interaction between appliance age and crime. are values for appliance age and crime.

To perform the nested model F-test to shows a 5% level of significance and to determine the null and alternative hypotheses are as followed:

The null hypothesis means that the beta estimates are equal to zero and the model is not significant. The alternative hypothesis means that the beta estimate does not equal to zero and the model is significant.

The P-value for nested F-test is . It shows that it is lower than the level of significance of 5% or 0.05. The null hypothesis will be rejected, and the alternative hypothesis will be used. Therefore, the second order model should be used and not the reduce model.

## 6. Conclusion

The analysis performed was the complete second order model; the first order regression model; the reduce model test; and the nested F-test for the reduced model and the second order model. I would use the second order model with the predictors of crime rate; appliance age; interaction appliance age and crime rate; appliance age squared, and crime rate squared. Since the second order model has a highvalue it can show that different variations of percentages can be shown with this model. The importance of this analysis was to help predict the home’s worth using the models within this dataset. It as well shown us that the variables used have a relationship with house prices and can be used to price a house correctly or if you are looking to get your house apprised.

## 1. Introduction

The dataset that I will be using is about heart disease dataset. The result from this analysis is expected to be used to predict the chances of having heart disease using a variety of factors. The analysis that will be used is regression a random forest and a random forest regression model

## 2. Data Preparation

The important variables being used are gender, cp (chest pain), cholesterol, resting ecg, exercise induced angina, peak of exercise, resting bp, target. There is 14 columns and 303 rows within the data set.

## 3. Model #1 - First Logistic Regression Model

### Reporting Results

The equation for the multiple regression model is:

The for the intercept. The for popular regression terms for age, resting blood pressure and maximum heart rate achieved.

The equation using these variables:

The equation used to form a model that is linear for beta terms:

The equation used for terms of the natural log of odds:

 is for probability in this case the event is developing heart disease.

 is the chances or odds of developing heart disease.

The equation for regression model:

The equation in this model for natural log of odds:

### Evaluating Model Significance

The Hosmer-Lemeshow goodness of fit (GOF) test assesses whether the model observes values of Y which are either 0 or 1. For this model it will be used to assess the model fitting the data or not.

Null and Alternative hypotheses:

( is 41.978 while the P-value is 0.7168. Showing a level of significance is 5%.

The P-value of 0.7168 is higher than level of significance of 0.05 or 5%. The null hypothesis should not be rejected.

Null and Alternative hypotheses calculating age based on Wald’s test with a 5% level of significance:

 for age.

Null and Alternative hypotheses calculating if resting blood pressure based on Wald’s test with a 5% level of significance:

 for resting blood pressure.

Null and Alternative hypotheses calculating if maximum heart rate achieved on Wald’s test with a 5% level of significance:

 for maximum heart rate achieved.

P-value of age is 0.5578. P-value for resting blood pressure is 0.0392. P-value for maximum heart rate achieved is 8.06e-10.

For resting blood pressure and maximum heart rate achieved are significant at a 5% level of significance. For age is not significant.

Form table output for confusion matrix:



Confusion Matrix for this model:



The confusion matrix results are:

* True positives: 127
* True negatives: 83
* False positives: 55
* False negatives: 38

**Accuracy:** Ratio of the number of correct predictions to the total number of observations.

**Precision:** Ratio of correct positive predictions to the total predicted positives.

**Recall**: Ratio of correct positive predictions to the total positive’s examples.

Receiver Operating Characteristic Curve:



Area under the curve is 0.7575 or 75.75%. This indicates of how well the model distinguishes between Y = 0 and Y = 1.

### Making Predictions Using Model

**Prediction1:**

Probability of individuals 50 years of age having a resting blood pressure of 122 and a maximum heart rate of 140 of contracting heart disease is 0.4939. With these probability in place the individual has a 49.39% of having heart disease.

Odds are calculated by dividing the probability by 1 minus the probability. In this case . It is 0.4939 to 0.5061 or 49.39% to 50.61%. Showing that it is 50/50 of developing heart disease for these individuals.

**Prediction 2:**

Probability of individuals 50 years old having a resting blood pressure of 140 and ha maximum heart rate of 170 of contracting heart disease is 0.7248. With these probability in place the individual has a 72.48% of having heart disease.

Odds are calculated by dividing the probability by 1 minus the probability. In this case . It is 0.7248 to 0.2752 or 72.48% to 27.52%. Showing that it is 75/25 of developing heart disease for these individuals

These predictions show a percentage of chance individuals can develop heart disease. The first prediction shows the odds are 50/50. The second prediction the odds are 75/75. The first prediction shows that at least one person out of every two will develop heart disease. The second prediction shows that at least 3 individuals out of every 4 individuals will develop heart disease.

## 4. Model #2 - Second Logistic Regression Model

### Reporting Results

The equation for multiple regression model:

are regression terms for maximum heart rate achieved, age of the individuals, dummy term for sex1 and the dummy term for exercise induced angina. are dummy terms for cp1, cp2 and cp3. are regression terms for age squared and the interaction of thalach against age.

For this equation it is:

The equation in terms of log of odds:

### Evaluating Model Significance

The Hosmer-Lemeshow goodness of fit (GOF) test assesses whether the model observes values of Y which are either 0 or 1. For this model it will be used to assess the model fitting the data or not.

Null and Alternative hypotheses:

( is 60.596. P-value is 0.1048. Showing a level of significance is 5%.

P-value of 0.1048 is higher than level of significance of 0.05 or 5%. Null hypothesis should not be rejected.

Null and Alternative hypotheses calculating if maximum heart rate is significant based Wald’s test with a 5% level of significance:

 for maximum heart rate.

P-value for maximum heart rate 0.014760. It is significant at 5% level of significance.

Null and Alternative hypotheses calculating if age is significant with Wald’s test with a 5% level of significance:

 for age.

P-value for age is 0.510325. It is not significant at 5% level of significance.

Null and Alternative hypotheses calculating if the dummy variable for sex is significant with Wald’s test with a 5% level of significance:

 for dummy variable for sex.

P-value for sex is 1.91e-06. It is significant at 5% level of significance.

Null and Alternative hypotheses calculating if exercise induced angina is significant with Wald’s test with a 5% level of significance:

 for exercise induced angina.

P-value for exercise induced angina is 0.009133. It is significant at 5% level of significance.

Null and Alternative hypotheses calculating if the dummy variable for chest pain is significant with Wald’s test with a 5% level of significance:

 for dummy variable for chest pain.

P-value for chest pain is 0.000249. It is significant at 5% level of significance.

Null and Alternative hypotheses calculating if the dummy variable for chest pain (cp2) is significant with Wald’s test with a 5% level of significance:

 for dummy variable for chest pain (cp2).

P-value for chest pain is 2.21e-06. It is significant at 5% level of significance.

Null and Alternative hypotheses calculating if the dummy variable for chest pain (cp3) is significant based on Wald’s test with a 5% level of significance:

 for dummy variable for chest pain (cp3).

P-value for chest pain is 0.003684. It is significant at 5% level of significance.

Null and Alternative hypotheses for calculating if age2 is significant with Wald’s test with a 5% level of significance:

 for age2.

P-value for age2 is 0.810599. It is not significant at 5% level of significance.

Null and Alternative hypotheses calculating if the interaction between maximum heart rate against age is significant with Wald’s test with a 5% level of significance:

 for interaction between maximum heart rate against age.

P-value for the interaction between maximum heart rate against age is 0.043666. It is significant at 5% level of significance.

The form table output for confusion matrix:



Confusion Matrix for this model:



The confusion matrix results are:

* True positives: 138
* True negatives: 103
* False positives: 35
* False negatives: 27

**Accuracy:** Ratio of the number of correct predictions to the total number of observations.

**Precision**: Ratio of correct positive predictions to the total predicted positives.

**Recall:**  Ratio of correct positive predictions to the total positives examples.

Receiver Operating Characteristic curve:



Area under the curve is 0.8777 or 87.77%. This indicates how well the model distinguishes between Y = 0 and Y = 1.

### Making Predictions Using Model

**Prediction 1:**

Probability of male individual contracting heart disease who is 30 years of age have a maximum heart rate of 140 and exercise while not experiencing chest pain is 0.2645 or 26.54%

**Prediction 2:**

Probability of male individual contracting heart disease who is 30 years of age have a maximum heart rate of 145 and not experience exercise related pain is 0.8432 or 84.32%

## 5. Random Forest Classification Model

### Reporting Results

The heart disease data set being split into 80% and 20% split. Using set.seed(511038) gives 303 rows in the data set, 242 rows in training set and 61 rows in testing set.

Training and testing error for number of trees using classification random forest model for the presence of heart disease using the following variables of age; gender; chest pain type; resting blood pressure; cholesterol measurement; resting electrocardiographic measurement; exercise angina; slope of peak exercise; and number of major vessels. Shows a maximum of 200 trees and using set.seed(511038) shows the following graph below:



Optimal number of trees is 20 trees. It shows where the curve flattens out.

### Evaluating the Utility of the model

Appropriate usage of number of trees. We can create classification random forest model for the presence of heart disease using variables of age, gender, chest pain type, resting blood pressure, cholesterol measurement, resting electrocardiographic measurement, exercise angina, slope of peak exercise, and number of major vessels.

Form table output of confusion matrix:



Confusion Matrix for this model’s training set:



The confusion matrix results for the training set are:

* True positives: 130
* True negatives: 111
* False positives: 1
* False negatives: 0

**Accuracy:**  Ratio of the number of correct predictions to the total number of observations.

**Precision:** Ratio of correct positive predictions to the total predicted positives.

**Recall**: Ratio of correct positive predictions to the total positive’s examples.

Confusion Matrix for this model’s testing set:



The confusion matrix results for the testing set are:

* True positives: 28
* True negatives: 18
* False positives: 8
* False negatives: 7

**Accuracy**: Ratio of the number of correct predictions to the total number of observations.

**Precision**: Ratio of correct positive predictions to the total predicted positives.

**Recall**: Ratio of correct positive predictions to the total positive’s examples.

## 6. Random Forest Regression Model

### Reporting Results

The heart disease data set being split into 80% and 20% split. Using set.seed(511038) gives 303 rows in the data set, 242 rows in training set and 61 rows in testing set.

Training and testing error for number of trees using classification random forest model for the presence of heart disease using the following variables of age; gender; chest pain type; resting blood pressure; cholesterol measurement; resting electrocardiographic measurement; exercise angina; slope of peak exercise; and number of major vessels. Shows a maximum of 200 trees and using set.seed(511038) shows the following graph below:



Optimal number of trees for this random forest model is 10 trees. It shows where the curve flattens out.

### Evaluating the Utility of the Random Forest Regression Model

Appropriate usage of number of trees. We can create a random forest regression model for maximum heart rate achieved using age; gender; chest pain type; resting blood pressure; cholesterol measurement; resting electrocardiographic measurement; exercise angina; slope of peak exercise; and number of major vessels.

Root mean squared error for training set is 9.9028.

Root mean squared for testing set is 17.387.

## 7. Conclusion

Comparing the two logistical models analyzed. I would use the second model. It shows more variables as well as having a higher accuracy for precision and recall for the confusion matrix. It shows that it is a genuine better predictor for individuals’ risk of contracting heart disease. I also would suggest using the random forest classification model over the logistic regression model since it has higher accuracy and precision with recall.

The practical importance of these analysis was for doctors to use individuals’ health information and have knowledge of their lifestyle to determine the risk of that individuals’ chances of contracting heart disease while using all the variables to determine if the individuals chances of contracting heart disease increases or decreases.