# Car Price Prediction
# and
# Heart Disease Classification

Hamid Hamidi, Thet Htet Chan Nyein (ID:90150020), and Yanzhao Qian

Authors are in alphabetic order and have equal contribution.

# Contents

# Abstract

Generalized Linear Models (GLMs) are the extension of the ordinary linear regression models. GLMs enable us to use different distributions for the response with distinctive link functions [1]. Here, we use two different data sets to show the broad applications of the GLMs in real-world problems, one of which is the "Car Price Prediction" [2], and the other is "Heart Failure prediction" [3]. In our analyses, we focus on model fitting and highlighting the most important variables instead of predicting desired outcomes and their accuracy. Our study of each data set is reported in its corresponding section. In the following, we discuss why we have chosen these data sets and provide a detailed description of our analyses along with the reasons and intuitions behind them. In both of these studies, all analyses were performed using the R programming language [4].

# 1. Car Price Prediction

## 1.1 Introduction

One of the largest automotive markets in the world is the USA car market [5]. Since 1982, when Honda invested in the USA car market, many other companies have been joining and competing in the USA car market resulting in foreign investment of more than $110 billion [5]. These days, with skilled workers, local and governmental supports, a huge consumer market, and many other reasons, the USA car market is a primer market in the car industry. A new Chinese car company wants to join and compete in the USA car market. In the following, our goal is to identify significant variables affecting the car price and quantify their significance. These analyses are usually performed by a third party, such as a consulting company, or the business strategy division of the investing company. According to our findings, they can manipulate many variables, such as the car design, to have a better business strategy to enter the USA car market. These analyses can directly affect the success of billions of dollars investment. Consequently, our analyses are vital and should be detailed and valid.

We found out the car price (response) distribution is quite close to the Gamma distribution; therefore, we used the GLM with Gamma distribution and logarithmic link function to model the price of cars for distinctive variables. We also suspected that it might be possible to model the logarithm of price with Gaussian distribution and identity link function. However, the distribution of the logarithmic price is not close to the Normal distribution. Consequently, we only used the Gamma distribution with the logarithmic link function. We performed variable selection and selected the most reasonable model (details in the Statistical Analyses section).

Using these analyses, we were able to identify several significant variables contributing to the car price, such as the car manufacturer (or the so-called brand of the car), the engine location (cars with rear engines are usually sports cars with higher prices), and the engine size (the bigger the higher the price).

Our data set, and consequently, our analyses have some limitations as well. For instance, electric cars are more than 2.5% of the USA car market [6] but are not included in our data set. Additionally, luxury brands such as Rolls-Royce and Lincoln are missing. Furthermore, the majority of sports cars are missing in our data set, showing our limitation in analyzing the sports and luxury car price variables. In the following, we present a detailed description of our analysis, methods, and results.

## 1.2 Data Collection and exploration

The data were collected from the Kaggle website (kaggle.com), an online open-source community of data scientists and machine learning practitioners. One can easily access the online version of our data through [2]. The data did not contain any missing values and was ready for analysis. However, we made minor changes and corrections in the data set.

We removed the CAR ID column as it does not contain useful information for our analyses. Additionally, we changed the names of the cars into manufacturers' names. This way, the variable would represent the car brand (or manufacturer) reputation, which might have an impact on car price, instead of the model of the car, which is unique for most cars and would not impact the car price.

Additionally, we removed the some of the covariates that have colinearity issues. For instance, covariates engine type and cylinder numbers are related since engine types are usually determined by the number of cylinders and how those cylinders are arranged [7]. The same scenario is true for fuel systems and fuel type as the fuel system of a car be affected by which fuel does the car consume [8]. To solve this issue of colinearity in our analysis, we removed engine type and fuel system.

Afterward, we tried to figure out the response distribution to use the appropriate link function and family of distribution [1]. The distribution of the response resembles the Gamma distribution (Figure 1 (A)). We also visualized the logarithm of the response since it might be Gaussian (Figure 1 (B)). As one can see in Figure 1 (B), the logarithm of price does not resemble the Gaussian distribution. Therefore, we decided to only use the Gamma distribution with the log link function (See Statistical Analyses for more details).

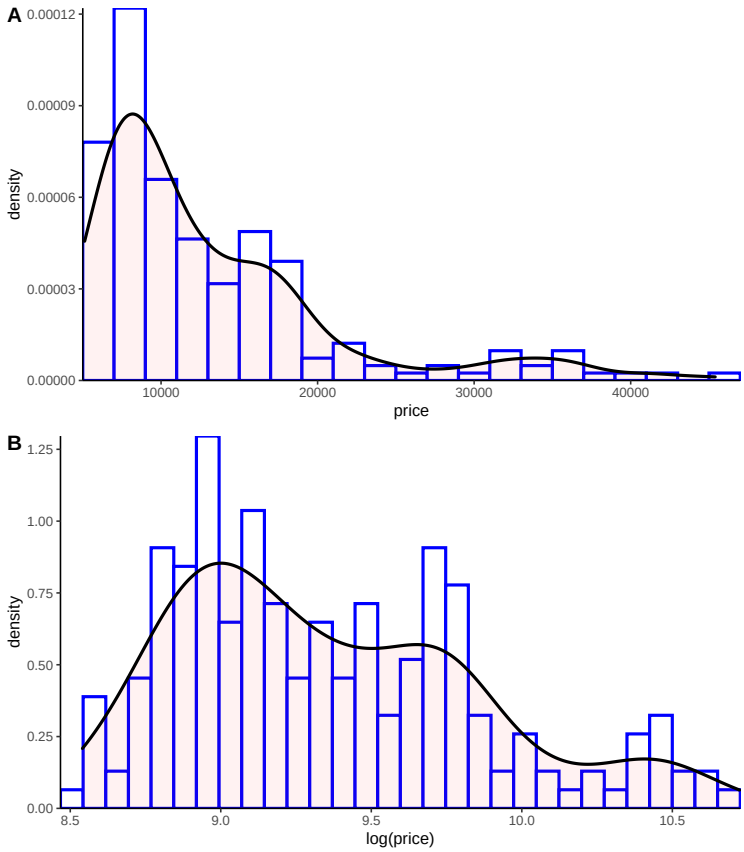



Figure 1: Distribution of the price.
(A) The distribution of the response (B) The distribution of the logarithm of the response.

Figure 2 shows an overview of the manufacturers, the range of their cars' prices, and the fuel type of their productions. As it is obvious, electric cars are missing, and diesel cars are the minority. Moreover, as shown in Figure 2, the car brands (or manufacturers) may affect the car price. For instance, cars from Porsche have a higher price compared to cars from Nissan or Mazda. Furthermore, we can see that famous sports car brands such as Ferrari and luxury manufacturers, such as Rolls-Royce and Lincoln, are missing.

In Appendix Figure 1, we can see that the car price is correlated with engine size; however, it might not be a linear correlation. Also, what stands out in this figure is the general growth of the engine size with increasing the number of Cylinders, and also, the response will rise with increasing any of them.

We also suspected that there might be some trends with the quadratic increase of numerical variables. Therefore, we investigated these patterns. For instance, in Appendix Figure 2, we divided the wheelbase[1] of cars into four groups and visualized the trend between the wheelbase and the response. As this figure shows, the car price is not growing linearly with the increase of the wheelbase. Hence, we included quadratic terms in our statistical model as well. In the next section, the details of our statistical analyses are described.

[1]In cars, the wheelbase is the distance between the front and rear wheels [9].

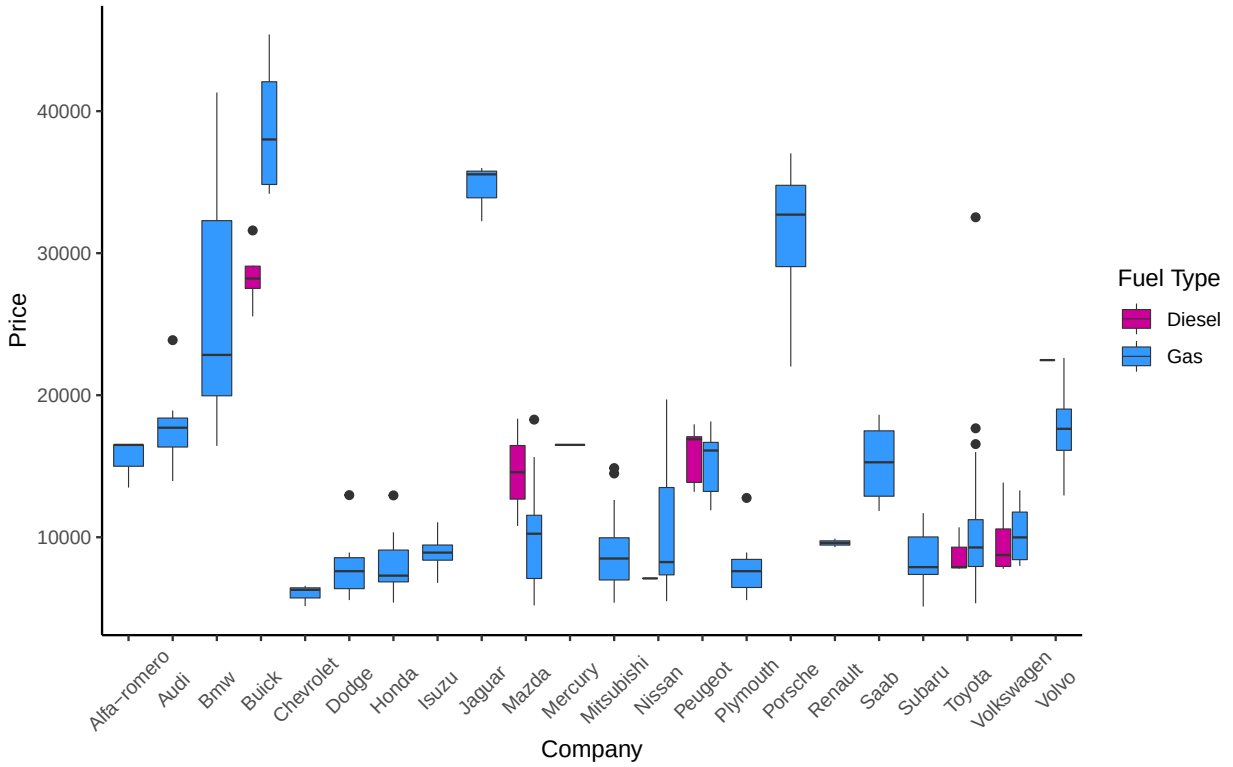



Figure 2: The range of car price in different brands and fuel types

## 1.3 Statistical Analyses

In Car Price Dataset, we have 205 observations and 27 variables including the response variable (price). Before starting the statistical analysis, let us define the data dictionary in order to better understand our dataset.

| Data Dictionary | |
|---|---|
| Variable Name | Definition |
| CarID | Unique ID of cars |
| Symboling | Its assigned insurance risk rating,(-3,-2,-1,0,1,2,3),-3= the most risky, 3=safest |
| car name | Name of car (Categorical) |
| car company | car manufacturer (Categorical) |
| fueltype | Car fuel type i.e gas or diesel (Categorical) |
| aspiration | Aspiration used in a car,i.e, std or turbo (Categorical) |
| doornumber | Number of doors in a car,i.e, four or two (Categorical) |
| carbody | body of car ,i.e, convertible, hardtop, hatchback, sedan or wagon(Categorical) |
| drivewheel | type of drive wheel,i.e, 4wd, fwd or rwd (Categorical) |
| enginelocation | Location of car engine,i.e, front or rear (Categorical) |
| wheelbase | Weelbase of car (Numeric) |
| carlength | Length of car (Numeric) |
| carwidth | Width of car (Numeric) |
| carheight | height of car (Numeric) |
| curbweight | The weight of a car without occupants or baggage. (Numeric) |
| enginetype | Type of engine. (Categorical) |
| cylindernumber | cylinder placed in the car (Categorical) |
| enginesize | Size of car (Numeric) |
| fuelsystem | Fuel system of car (Categorical) |
| boreratio | Boreratio of car (Numeric) |
| stroke | Stroke or volume inside the engine (Numeric) |
| compressionratio | compression ratio of car (Numeric) |
| horsepower | Horsepower (Numeric) |
| peakrpm | car peak rpm (Numeric) |
| citympg | Mileage in city (Numeric) |
| highwaympg | Mileage on highway (Numeric) |
| price(Dependent variable) | Price of car (Numeric) |

Table 1.1: *Data Dictionary*

We firstly decide which model to use for analysis by using the histograms of car price and log(car price). We cannot use normal linear regression for modeling car price since car price and log(car price) do not follow normal distribution. Upon further inspection, we decided to gamma regression with log-link since both car price and log(car price) follow gamma distribution. We specifically chose log-link for gamma regression instead of its canoncial link (negative-inverse link) since the car price can only be positive.

For the analysis of car price dataset, we decided to construct two models:

(1) The main effect model

(2) The main effect model plus the squared numeric covariates

We decided not to include interaction terms in our analysis since most of our covariates are categorical and those most categorical covariates have more than 2 levels.

For the first model, I firstly constructed the model with all the covariates (excluding the covariates that causes the collinearity issues) with price as response.The formula for the model is as follows.

$$log(\mu) = \beta_0 + \sum_{i=1}^{22} \beta_i x_i$$

The formula for the full main effect model is:

**Full Model 1 Forumula**

log(price)~fueltype+aspiration+doornumber+carbody+drivewheel+enginelocation+
cylindernumber+car_company+symboling+wheelbase+carlength+carwidth+carheight +curbweight+enginesize
+boreratio+stroke+compressionratio+horsepower+peakrpm+citympg+highwaympg

The following table shows the results of full main effect model.

| | Estimate | Pr(>\|t\|) | significance |
|---|---|---|---|
| (Intercept) | 7.5 | 8.47e-12 | *** |
| Fueltype gas | -0.23 | 0.54 | not significant |
| Aspiration Turbo | 0.08 | 0.10 | not significant |
| Doornumber two | -0.03 | 0.28 | not significant |
| Carbody Hardtop | -0.16 | 0.05 | * |
| Carbody Hatchback | -0.21 | 1.87e-03 | ** |
| Carbody Sedan | -0.14 | 0.06 | . |
| Carbody Wagon | -0.15 | 0.06 | . |
| Drivewheel fwd | -0.04 | 0.52 | not significant |
| Drivewheel rwd | -6.23e-03 | 0.93 | not significant |
| enginelocation rear | 0.78 | 6.63e-09 | *** |
| cylinder num5 | 0.04 | 0.75 | not significant |
| cylinder num4 | 0.20 | 0.19 | not significant |
| cylinder num6 | 0.04 | 0.75 | not significant |
| cylinder num3 | 0.51 | 0.02 | * |
| cylinder num12 | -0.13 | 0.66 | not significant |
| cylinder num2 | 0.31 | 0.12 | not significant |
| Audi | 0.07 | 0.62 | not significant |
| BMW | 0.36 | 8.25e-04 | *** |
| Buick | -0.05 | 0.72 | not significant |
| Chevrolet | -0.25 | 0.05 | . |
| Dodge | -0.35 | 1.07e-03 | ** |
| Honda | -0.18 | 0.09 | . |
| Isuzu | -0.16 | 0.30 | not significant |
| Jaguar | -0.37 | 0.02 | * |
| Mazda | -0.09 | 0.34 | not significant |
| Mercury | -0.15 | 0.32 | not significant |
| Mitsubishi | -0.40 | 2.21e-04 | *** |
| Nissan | -0.15 | 0.12 | not significant |
| Peugeot | -0.37 | 4.53e-03 | ** |
| Plymouth | -0.36 | 8.23e-04 | *** |
| Porsche | 0.03 | 0.83 | not significant |
| Renault | -0.27 | 0.05 | . |
| Saab | 0.1 | 0.39 | not significant |

| | Estimate | Pr(>\|t\|) | significance |
|---|---|---|---|
| Subaru | -0.20 | 0.10 | . |
| Toyota | -0.20 | 0.03 | * |
| Volkswagen | -0.12 | 0.26 | not significant |
| Volvo | -0.09 | 0.43 | not significant |
| symboling | 1.43e-03 | 0.93 | not significant |
| wheelbase | 0.02 | 6.54e-04 | *** |
| carlength | -7.39e-03 | 0.02 | * |
| carwidth | 0.03 | 0.02 | * |
| carheight | -0.03 | 1.06e-04 | *** |
| curbweight | 5.29e-04 | 1.21e-06 | *** |
| enginesize | 2.55e-03 | 0.07 | . |
| boreratio | -0.17 | 0.1 | . |
| stroke | -0.01 | 0.82 | not significant |
| compressionratio | -0.01 | 0.62 | not significant |
| horesepower | 0.02e-04 | 0.41 | not significant |
| peakrpm | 5.31e-05 | 0.15 | not significant |
| citympg | -0.01 | 0.08 | . |
| highwaympg | 0.01 | 0.19 | not significant |

Signif.codes: overwhelming *** strong ** moderate * borderline .

Table 1.2: *Full Main Effect Model results*

From Table 1.2, we can see that some covariates are insignificant. Hence, we conducted stepwise selection with both forward and backward direction and AIC as selection criteria [10]. AIC or Akaike information criterion is a popular method for variable selction for model building.

For choosing a model from a sequence of model candidates $M_i$, $i = 1, 2, \ldots K$. The AIC is defined as

$$\text{AIC}_i = -2 \log L_i + 2V_i$$

From the AIC stepwise selection, we obtained the stepwise model for main effects.

The formula for stepwise main effect is as follows:

**Stepwise Model 1 Formula**

log(price)~aspiration+carbody+enginelocation+cylindernumber

+car_company+wheelbase+carlength+carwidth+carheight+curbweight+enginesize

+boreratio+peakrpm+citympg+highwaympg

The following table shows the result of the stepwise selection for main effects.

| | Estimate | Pr(>\|t\|) | significance |
|---|---|---|---|
| (Intercept) | 7.12 | 3.04e-14 | *** |
| Aspiration Turbo | 0.11 | 2.93e-04 | *** |
| Carbody Hardtop | -0.11 | 0.05 | . |
| Carbody Hatchback | -0.19 | 2.73e-03 | ** |
| Carbody Sedan | -0.10 | 0.12 | not significant |
| Carbody Wagon | -0.12 | 0.12 | not significant |
| enginelocation rear | 0.82 | 9.36e-11 | *** |
| cylinder num5 | 0.02 | 0.84 | not significant |
| cylinder num4 | 0.17 | 0.2 | not significant |
| cylinder num6 | 0.01 | 0.92 | not significant |
| cylinder num3 | 0.49 | 0.02 | * |
| cylinder num12 | -0.13 | 0.46 | not significant |

| | Estimate | Pr(>\|t\|) | significance |
|---|---|---|---|
| cylinder num2 | 0.19 | 0.09 | . |
| Audi | 0.02 | 0.84 | not significant |
| BMW | 0.34 | 5.85e-04 | *** |
| Buick | -1.1 | 0.39 | not significant |
| Chevrolet | -0.29 | 0.02 | * |
| Dodge | -0.4 | 2.17e-05 | *** |
| Honda | -0.23 | -0.01 | * |
| Isuzu | -0.15 | 0.13 | not significant |
| Jaguar | -0.45 | 1.25e-03 | *** |
| Mazda | -0.12 | 0.20 | not significant |
| Mercury | -0.16 | 0.30 | not significant |
| Mitsubishi | -0.45 | 8.02e-07 | *** |
| Nissan | -0.18 | 0.03 | * |
| Peugeot | -0.39 | 4.68e-04 | *** |
| Plymouth | -0.40 | 2.5e-05 | *** |
| Porsche | 0.02 | 0.90 | not significant |
| Renault | -0.32 | 5.46e-03 | ** |
| Saab | 0.07 | 0.55 | not signicant |
| Subaru | -0.12 | 0.04 | * |
| Toyota | -0.22 | 9.97e-03 | ** |
| Volkswagen | -0.15 | 0.08 | . |
| Volvo | -0.12 | 0.25 | not significant |
| wheelbase | 0.02 | 1.83e-04 | *** |
| carlength | -0.01 | 5.63e-03 | ** |
| carwidth | 0.03 | 0.02 | * |
| carheight | -0.03 | 1.43e-05 | *** |
| curbweight | 6.01e-4 | 5.71e-12 | *** |
| enginesize | 2.76e-3 | 0.02 | * |
| boreratio | -0.16 | 0.09 | . |
| peakrpm | 6.5e-5 | 0.03 | * |
| citympg | -0.02 | 0.02 | * |
| highwaympg | 0.01 | 0.09 | . |

Signif.codes: overwhelming *** strong ** moderate * borderline .

Table 1.3: *Stepwise Main Effect Model results*

For the second model, I firstly constructed the model with all the covariates (excluding the covariates that causes the collinearity issues) plus the squared of the numeric terms.

**Full Model 2 Forumula**

log(price)~fueltype+aspiration+doornumber+carbody+drivewheel+enginelocation+

cylindernumber+car_company+symboling+wheelbase+carlength+carwidth+carheight+curbweight+enginesize

+boreratio+stroke+compressionratio+horsepower+peakrpm+citympg+highwaympg

+wheelbase^2 + carlength^2 + carwidth^2 +carheight^2 + curbweight^2

+enginesize^2 + boreratio^2 +peakrpm^2 + citympg^2 + highwaympg^2

The following table shows the results of full main effect model with squared numeric terms.

| | Estimate | Pr(>\|t\|) | significance |
|---|---|---|---|
| (Intercept) | 18.87 | 0.20 | not significant |
| Fueltype gas | 0.21 | 0.63 | not significant |
| Aspiration Turbo | 0.10 | 0.05 | . |
| Doornumber two | -0.03 | 0.29 | not significant |
| Carbody Hardtop | -0.19 | 0.02 | * |
| Carbody Hatchback | -0.22 | 2.36e-03 | *** |
| Carbody Sedan | -0.15 | 0.05 | * |
| Carbody Wagon | -0.14 | 0.11 | not significant |
| Drivewheel fwd | -0.02 | 0.67 | not significant |
| Drivewheel rwd | -0.02 | 0.74 | not significant |
| enginelocation rear | 0.75 | 4.54e-07 | *** |
| cylinder num5 | 0.12 | 0.47 | not significant |
| cylinder num4 | 0.26 | 0.19 | not significant |
| cylinder num6 | 0.10 | 0.47 | not significant |
| cylinder num3 | 0.56 | 0.04 | * |
| cylinder num12 | -0.41 | 0.21 | not significant |
| cylinder num2 | 0.16 | 0.58 | not significant |
| Audi | 0.05 | 0.78 | not significant |
| BMW | 0.42 | 1.13e-03 | ** |
| Buick | 0.18 | 0.38 | not significant |
| Chevrolet | -0.25 | 0.09 | . |
| Dodge | -0.36 | 2.85e-03 | ** |
| Honda | -0.19 | 0.11 | not significant |
| Isuzu | -0.10 | 0.43 | not significant |
| Jaguar | -0.04 | 0.87 | not significant |
| Mazda | -0.02 | 0.83 | not significant |
| Mercury | -0.09 | 0.61 | not significant |
| Mitsubishi | -0.40 | 7.56e-04 | *** |
| Nissan | -0.10 | 0.37 | not significant |
| Peugeot | -0.22 | 0.14 | not significant |
| Plymouth | -0.37 | 1.85e-03 | ** |
| Porsche | 0.06 | 0.68 | not significant |
| Renault | -0.25 | 0.08 | . |
| Saab | 0.14 | 0.30 | not significant |
| Subaru | -0.13 | 0.35 | not significant |
| Toyota | -0.18 | 0.10 | . |
| Volkswagen | -0.11 | 0.36 | not significant |
| Volvo | 1.64e-03 | 0.99 | not significant |
| symboling | 2.92e-03 | 0.87 | not significant |
| wheelbase | 0.04 | 0.65 | not significant |
| carlength | 0.02 | 0.73 | not significant |
| carwidth | -0.25 | 0.52 | not significant |
| carheight | -0.22 | 0.37 | not significant |
| curbweight | 1.59e-03 | 6.92e-03 | ** |
| enginesize | 9.74e-04 | 0.81 | not significant |
| boreratio | -0.11 | 0.94 | not significant |
| stroke | 0.22 | 0.74 | not significant |
| compressionratio | 0.02 | 0.63 | not significant |
| horesepower | 7.68e-04 | 0.53 | not significant |
| peakrpm | -5.02e-04 | 0.21 | not significant |
| citympg | -0.04 | 0.29 | not significant |
| highwaympg | 7.74e-04 | 0.98 | not significant |
| I(wheelbase^2) | -9.46e-05 | 0.84 | not significant |
| I(carlength ^2) | -7.12e-05 | 0.59 | not significant |
| I(carwidth ^2) | 2.13e-03 | 0.47 | not significant |
| I(carheight ^2) | 1.67e-03 | 0.46 | not significant |
| I(curbweight ^2) | -2.03e-07 | 0.06 | . |
| I(enginesize ^2) | 5.67e-06 | 0.47 | not significant |

| | Estimate | Pr(>\|t\|) | significance |
|---|---|---|---|
| I(boreratio ^2) | -0.02 | 0.95 | not significant |
| I(peakrpm) ^2 | 5.46e-08 | 0.16 | not significant |
| I(citympg^2) | 4.46e-04 | 0.46 | not significant |
| I(highwaympg^2) | 1.25e-04 | 0.81 | not significant |

Signif.codes: overwhelming *** strong ** moderate * borderline .

Table 1.4: *Full Main Effect Model with Squared Terms results*

From Table 1.4, we can see that some covariates are insignificant. Hence, we conducted stepwise selection with both forward and backward direction and AIC as selection criteria. From the AIC stepwise selection, we obtained the stepwise model for main effects plus squared numeric terms.

The formula for stepwise main effect plus squared numeric model is as follows:

**Stepwise Model 2 Formula**

log(price) ~ aspiration + carbody + enginelocation

+car_company + carheight + curbweight + peakrpm + citympg

+wheelbase^2 + carlength^2 + carwidth^2 + curbweight^2

+enginesize^2 + peakrpm^2 + citympg^2 + highwaympg^2

The following table shows the result of the stepwise selection for main effects plus squared numeric terms.

| | Estimate | Pr(>\|t\|) | significance |
|---|---|---|---|
| (Intercept) | 9.07 | 6.38e-13 | *** |
| Aspiration Turbo | 0.11 | 7.62e-05 | *** |
| Carbody Hardtop | -0.17 | 0.01 | * |
| Carbody Hatchback | -0.23 | 1.82e-04 | *** |
| Carbody Sedan | -0.16 | 0.01 | * |
| Carbody Wagon | -0.16 | 0.02 | * |
| enginelocation rear | 0.68 | 5.36e-11 | *** |
| Audi | -0.03 | 0.78 | not significant |
| BMW | 0.31 | 5.16e-04 | *** |
| Buick | 0.04 | 0.73 | not significant |
| Chevrolet | -0.19 | 0.07 | . |
| Dodge | -0.35 | 1.1e-03 | ** |
| Honda | -0.2 | 0.02 | * |
| Isuzu | -0.13 | 0.18 | not significant |
| Jaguar | -0.12 | 0.46 | not significant |
| Mazda | -0.10 | 0.20 | not significant |
| Mercury | -0.17 | 0.23 | not significant |
| Mitsubishi | -0.39 | 3.90e-06 | *** |
| Nissan | -0.16 | 0.05 | * |
| Peugeot | -0.33 | 7.12e-04 | *** |
| Plymouth | -0.35 | 8.51e-05 | *** |
| Porsche | 0.02 | 0.86 | not significant |
| Renault | -0.29 | 0.01 | ** |
| Saab | 0.08 | 0.43 | not significant |

| | Estimate | Pr(>\|t\|) | significance |
|---|---|---|---|
| Subaru | -0.25 | 2.34e-03 | ** |
| Toyota | -0.22 | 4.98e-03 | ** |
| Volkswagen | -0.15 | 0.07 | . |
| Volvo | -0.1 | 0.27 | not significant |
| carheight | -0.03 | 2.61e-05 | *** |
| curbweight | 1.43e-03 | 2.19e-05 | *** |
| peakrpm | -4.85e-04 | 0.14 | not significant |
| citympg | -0.04 | 1.13e-03 | ** |
| I(wheelbase^2) | 1.06e-04 | 2.49e-05 | *** |
| I(carlength ^2) | -2.17e-05 | 7.87e-03 | ** |
| I(carwidth ^2) | 1.98e-04 | 0.02 | * |
| I(curbweight ^2) | -1.71e-07 | 5.98e-03 | ** |
| I(enginesize ^2) | 3.96-06 | 0.01 | * |
| I(peakrpm ^2) | 5.27e-08 | 0.09 | . |
| I(citympg^2) | 3.5e-04 | 0.08 | . |
| I(highwaympg^2) | 1.84e-04 | 0.06 | . |

Signif.codes: overwhelming *** strong ** moderate * borderline .

Table 1.5: *Stepwise Main Effect Model with Squared Terms results*

For each stepwise model, we decided to conduct likelihood ratio test to compare with their respective full models. The following are the results of the likelihood ratio tests.

$H_0$ : Our stepwise model is the same as its respective full model.

$H_a$ : Our stepwise model different from its respective full model.

| Tests | Model | LogLikelihood value | Test Statistics | $P(\chi^2 \geq \chi^2_{test})$ |
|---|---|---|---|---|
| Full Model 1 vs Stepwise Model 1 | Full Model 1 | -1733 (df=53) | 1.93(df=8) | 0.983 |
| | Stepwise Model 1 | -1735 (df=45) | | |
| Full Model 2 vs Stepwise Model 2 | Full Model 2 | -1721 (df=63) | 9.42(df=22) | 0.991 |
| | Stepwise Model 2 | -1731 (df=41) | | |

Table 1.6: *Likelihood Ratio Tests for Full Models vs Stepwise AIC Models*

Since p-values for both likelihood ratio tests are large, we fail to reject the null hypothesis. Also, since the p-values are close to 1, there is a siginificant evidence to support that our Stepwise Models are the same as their respective full model counterparts.

Next, we are going to conduct deviance tests for both Best Models in order to check model adequacy.

$H_0$ : Our stepwise model is adequate.

$H_a$ : Our stepwise model not adequate.

| Model | Residual Deviance | $P(\chi^2 \geq \chi^2_{test})$ |
|---|---|---|
| Stepwise Model 1 | 2.01 (161 df) | 1 |
| Stepwise Model 2 | 1.83(165 df) | 1 |

Table 1.7: *Devaince Tests for Stepwise AIC Models*

Since p-values for both deviance tests are large we fail to reject the null hypothesis. Also, since the p-values are 1, there is a siginificant evidence to support that our Stepwise Models are adequate.

We then decided to choose the best model out of those two Stepwise models after the model adequacy testing. Since Stepwise Model 1 and Stepwise Model 2 are not nested model for each other, we cannot use Likelihood

Ratio Test for model comparison. Instead, we decided to use AIC score, Pseudo $R^2$ and deviance values as the selection criteria.

**Stepwise Model 1:** AIC= 3560, Pseudo-$R^2$ = 0.965, deviance= 2.01 (161 df)

**Stepwise Model 2:** AIC= 3543, Pseudo-$R^2$ = 0.966, deviance= 1.83 (165 df)

Since Stepwise Model 2 has the lower AIC, higher Pseudo $R^2$ and lower deviance score than Stepwise Model 1, we decided to choose Stepwise Model 2 as our best model.

Next, we constucted 95% t-confidence interval for our parameter estimates. We obtained the confidence interval of paramter estimates by

$$\hat{\beta}_i \pm SE_{\hat{\beta}_i} t_{\alpha/2,165}$$

| | Estimate | SE | Lower Limit | Upper Limit |
|---|---|---|---|---|
| (Intercept) | 9.07 | 1.16 | 6.77 | 11.36 |
| Aspiration Turbo | 0.11 | 0.03 | 0.06 | 0.17 |
| Carbody Hardtop | -0.17 | 0.07 | -0.31 | -0.04 |
| Carbody Hatchback | -0.23 | 0.06 | -0.34 | -0.11 |
| Carbody Sedan | -0.16 | 0.06 | -0.28 | -0.03 |
| Carbody Wagon | -0.16 | 0.07 | -0.29 | -0.02 |
| enginelocation rear | 0.68 | 0.1 | 0.49 | 0.87 |
| Audi | -0.03 | 0.09 | -0.21 | 0.16 |
| BMW | 0.31 | 0.09 | 0.14 | 0.48 |
| Buick | 0.04 | 0.12 | -0.19 | 0.27 |
| Chevrolet | -0.19 | 0.11 | -0.40 | 0.02 |
| Dodge | -0.35 | 0.09 | -0.52 | -0.17 |
| Honda | -0.2 | 0.09 | -0.37 | -0.03 |
| Isuzu | -0.13 | 0.09 | -0.31 | 0.06 |
| Jaguar | -0.12 | 0.16 | -0.42 | 0.19 |
| Mazda | -0.10 | 0.08 | -0.26 | 0.05 |
| Mercury | -0.17 | 0.13 | -0.44 | 0.10 |
| Mitsubishi | -0.39 | 0.08 | -0.55 | -0.23 |
| Nissan | -0.16 | 0.08 | -0.32 | -2.64e-03 |
| Peugeot | -0.33 | 0.10 | -0.52 | -0.14 |
| Plymouth | -0.35 | 0.09 | -0.53 | -0.18 |
| Porsche | 0.02 | 0.11 | -0.19 | 0.23 |
| Renault | -0.29 | 0.11 | -0.55 | -0.08 |
| Saab | 0.08 | 0.1 | -0.12 | 0.26 |
| Subaru | -0.25 | 0.08 | -0.42 | -0.09 |
| Toyota | -0.22 | 0.08 | -0.37 | -0.07 |
| Volkswagen | -0.15 | 0.08 | -0.31 | 0.01 |
| Volvo | -0.1 | 0.09 | -0.28 | 0.08 |
| carheight | -0.03 | 6.81e-03 | -0.04 | -0.02 |
| curbweight | 1.43e-03 | 3.26e-04 | 7.82e-04 | 2.07e-03 |
| peakrpm | -4.85e-04 | 3.25e-04 | -1.13e-03 | 1.57e-04 |
| citympg | -0.04 | 0.01 | -0.06 | -0.02 |
| I(wheelbase^2) | 1.06e-04 | 2.43e-05 | 5.75e-05 | 1.54e-04 |
| I(carlength ^2) | -2.17e-05 | 8.07e-06 | -3.77e-05 | -5.78e-06 |
| I(carwidth ^2) | 1.98e-04 | 8.67e-05 | 2.67e-05 | 3.69e-04 |
| I(curbweight ^2) | -1.71e-07 | 6.14e-08 | -2.92e-07 | -4.98e-08 |
| I(enginesize ^2) | 3.96-06 | 1.55e-06 | 8.94e-07 | 7.01e-06 |
| I(peakrpm ^2) | 5.27e-08 | 3.08e-08 | -8.19e-09 | 1.14e-07 |
| I(citympg^2) | 3.5e-04 | 2.01e-04 | -4.71e-05 | 7.47e-04 |
| I(highwaympg^2) | 1.84e-04 | 9.84e-05 | -10e-06 | 3.79e-04 |

Table 1.8: *95 % t-confidence interval for paramter estimates*

Also, since our response is log(car price), we also constructed the 95% t-confidence interval for multiplicative effects. We obtained the confidence interval of multiplicative effects by

$$exp(\hat{\beta}_i \pm SE_{\hat{\beta}_i} t_{\alpha/2,165})$$

| | $exp(\hat{\beta}_i)$ | Lower Limit | Upper Limit |
|---|---|---|---|
| (Intercept) | 8468.52 | 873.67 | 85612.46 |
| Aspiration Turbo | 1.12 | 1.06 | 1.18 |
| Carbody Hardtop | 0.84 | 0.74 | 0.96 |
| Carbody Hatchback | 0.80 | 0.71 | 0.90 |
| Carbody Sedan | 0.86 | 0.76 | 0.97 |
| Carbody Wagon | 0.86 | 0.75 | 0.98 |
| enginelocation rear | 1.97 | 1.63 | 2.38 |
| Audi | 0.97 | 0.81 | 1.17 |
| BMW | 1.36 | 1.15 | 1.62 |
| Buick | 1.04 | 0.83 | 1.31 |
| Chevrolet | 0.83 | 0.67 | 1.02 |
| Dodge | 0.71 | 0.60 | 0.84 |
| Honda | 0.82 | 0.69 | 0.97 |
| Isuzu | 0.88 | 0.73 | 1.06 |
| Jaguar | 0.89 | 0.65 | 1.21 |
| Mazda | 0.90 | 0.77 | 1.06 |
| Mercury | 0.85 | 0.64 | 1.11 |
| Mitsubishi | 0.68 | 0.58 | 0.80 |
| Nissan | 0.85 | 0.73 | 1 |
| Peugeot | 0.72 | 0.59 | 0.87 |
| Plymouth | 0.70 | 0.59 | 0.84 |
| Porsche | 1.02 | 0.83 | 1.26 |
| Renault | 0.75 | 0.60 | 0.93 |
| Saab | 1.08 | 0.89 | 1.30 |
| Subaru | 0.78 | 0.66 | 0.91 |
| Toyota | 0.80 | 0.69 | 0.94 |
| Volkswagen | 0.86 | 0.73 | 1.02 |
| Volvo | 0.91 | 0.76 | 1.08 |
| carheight | 0.97 | 0.96 | 0.98 |
| curbweight | 1.00 | 1.00 | 1.00 |
| peakrpm | 1.00 | 1.00 | 1.00 |
| citympg | 0.96 | 0.94 | 0.99 |
| I(wheelbase^2) | 1.00 | 1.00 | 1.00 |
| I(carlength ^2) | 1.00 | 1.00 | 1.00 |
| I(carwidth ^2) | 1.00 | 1.00 | 1.00 |
| I(curbweight ^2) | 1.00 | 1.00 | 1.00 |
| I(enginesize ^2) | 1.00 | 1.00 | 1.00 |
| I(peakrpm ^2) | 1.00 | 1.00 | 1.00 |
| I(citympg^2) | 1.00 | 1.00 | 1.00 |
| I(highwaympg^2) 1.00 | 1.00 | 1.00 | |

Table 1.9: *95 % t-Confidence Interval for multiplicative effect*

Finally, we did residual analysis to check whether our residuals follow normality and constant variance assumptions. For our analysis, we decided to use standardized residuals instead of residuals for easier interpretations. We obtained standardzied residuals by finding the difference between observed and fitted values and dividing with fitted values. Afterwards, we drew the residuals plots for further analysis.

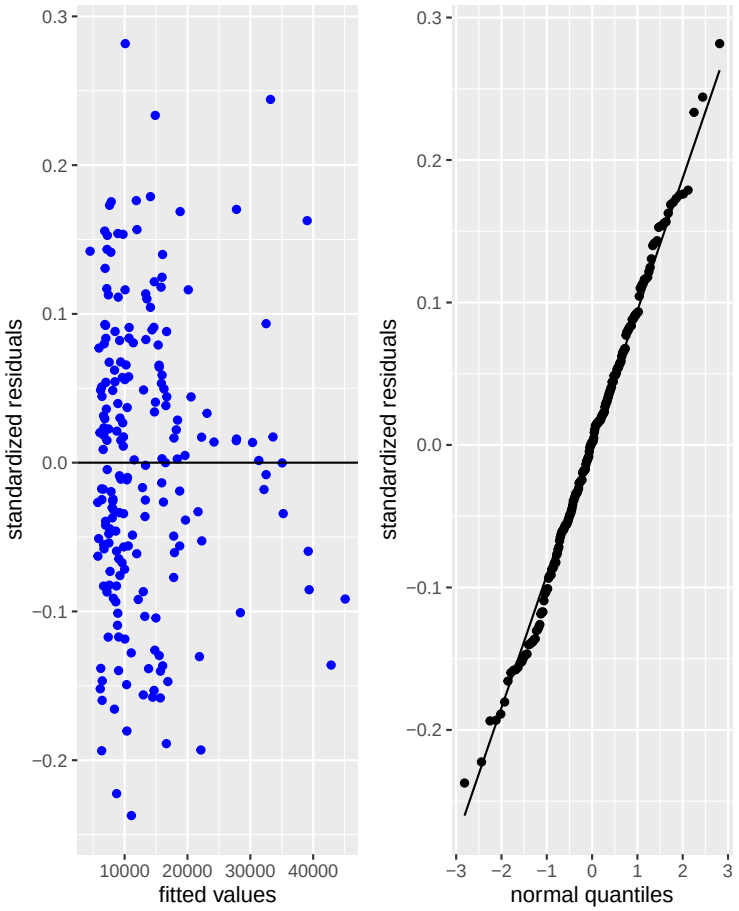



Figure 3: Distribution of the price.
(Left) Std. Residual vs Fitted (Right) Nomral QQ Plot of Std. Residuals .

From the plot of standardized residual vs fitted, we can see that there is no obvious pattern within the standardized residuals. From the Normal QQ plot of standardized residuals, we can see that most of the standardized residuals revolve around the normality line and there is no residual that greatly deviates from that line. Hence, results from the standardized residuals indicate that our model is a good fit.

## 1.4 Conclusion

Our best model include covariates: aspiration, enginetype, carbody, carcompany, carheight, curbweight, peakrpm, citympg, wheelbase, carlength, carwidth, enginesize and highwaympg. Before we proceed with the conclusion of this analysis, the following are some of the highlights of our parameter interpretations.

(1) Car price is increased by approximately 2 folds when a car has a rear engine.

(2) Car bodies other than convertible (reference level) causes car price to go down.

(3) Among car manufacturers, BMW has the best reputation as car price increases by 1.4 times when the brand is BMW. On the other hand, the opposite is true for Plymouth and Dodge having the worst reputation (car price decreses by 29% when those companies are manufacturers).

(4) Most numeric covariates in our model has the quadratic relationship with log(price).

(5) Our Pseudo-$R^2$ value is exceptionally high (approximately around 96%)/

Even though our model shows the promising prospects on studying car prices in the United States, there are certain limitations that we need to address.

Firstly, electric cars that are made up of 10% of current car market are not included in our dataset. Secondly, we can see the absence of luxuy cars and sports cars. Finally, the data is limited for certain brands such as Mercury only having one sample in our entire dataset.

Therefore, it can be concluded that although our model shows promising results for the study of car prices,it may not reflect the current car prices in the United States due to the limitations of our data.

# 2. Heart Disease Classification

## 2.1 Introduction

Heart Failure, also known as congestive heart failure, can broadly be defined as a condition that happens when the heart cannot supply the body's need for Oxygen and blood [11]. According to the latest annual statistical report from the American Heart Association and the National Institutes of Health, about 6.2 million adults in the United States have heart failure [12]. Furthermore, in 2018, heart failure was mentioned on 379,800 death certificates (13.4%) [12] and cost about $30 billion annually [13].

This suggests that identifying the core health behaviors and risk factors influencing heart failure is critical not only for our community health but also for our economy. Therefore, we decided to analyze the "Heart Failure prediction" data set [3] to find variables playing a key role in heart failure. As the response in our data set is binary (0 or 1), we used logistic regression to model the probability of having heart failure. Moreover, we performed variable selection to select the best model and determine major factors in heart failure (details in the Statistical Analyses section).

This study has generally revealed causal factors in heart failure such as sex, exercise angina (a type of chest pain during performing exercises), distinctive types of chest pain, and squared Cholesterol level.

The generalisability of our results is subject to certain limitations. For instance, our data set does not cover younger generations (less than 28 years old), which will cause the analyses to be biased toward older ages. Another issue that was not addressed in this study was the mortality of the patients. This might not seem arguable at first look. However, many patients with asymptomatic chest pain might have lived without any critical problems throughout their lives, and patients with other types of chest pain might have faced devastating situations. This might cause our findings to be questionable from different perspectives. Overall, our study concluded significant risk factors in heart failure. The following chapters are a detailed description of our analyses, methods, and results.

## 2.2 Data Collection and exploration

The data were collected from the Kaggle website (kaggle.com), an online open-source community of data scientists and machine learning practitioners. One can easily access the online version of our data through [3]. The data contain 918 subjects with 11 covariates and one binary response (Heart Failure or not). The data do not have any missing values and are ready for analysis.

Figure 4 represents the distribution of age in samples for their sexuality and heart condition. As it is obvious, the number of male samples is dramatically higher than female samples indicating whether our data set is biased or males have more heart failure than females. According to two distinctive independent studies, men have more incidents of heart failure, which is consistent with our data set [14], [15]. Another observation in Figure 4 is that the range of the age starts from 28, showing that our data set does not contain younger generations, and our analysis is not valid for younger ages.

As shown in Appendix Figure 3 (A), asymptomatic chest pain type is more frequent in men and women, and also samples with asymptomatic chest pain type are more exposed to heart failure. This observation is not surprising because patients with heart failure may show symptoms [16]. Appendix Figure 3 (B) presents that samples with exercise-induced angina[2] are more exposed to heart failure in both sex groups.

In Appendix Figure 4, we can see that the patients with heart problems have higher oldpeak[3], and as one might suspect, the heart failure might be affected by the quadratic of the oldpeak as well. Therefore, we decided to include quadratic forms of numerical variables in our modeling as well. A detailed description of the main findings, together with our conclusion, is provided in the next chapters.

## 2.3 Statistical Analyses

[2]Angina is a type of chest pain caused by reduced blood flow to the heart [17].
[3]ST depression induced by exercise relative to rest [18].

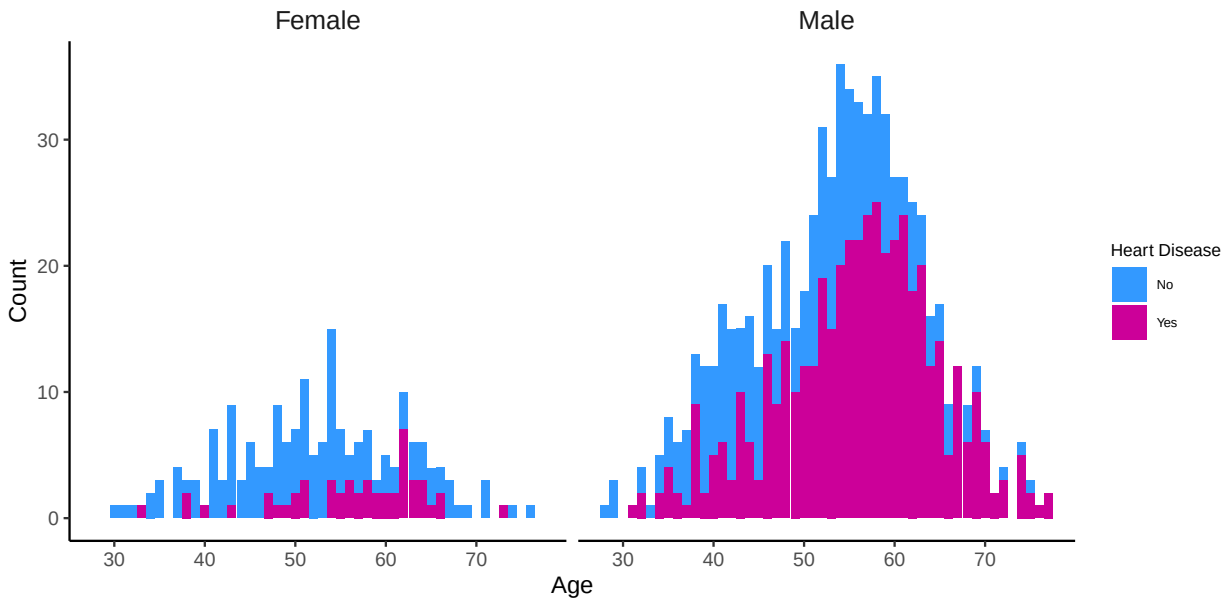



Figure 4: The distribution of the age of the samples with respect to their sex and heart condition

### 2.3.1 Three GLM models and stepwise AIC selection

To begin with, let's assign each feature with a variable name .

| Data Dictionary | | | |
|---|---|---|---|
| Variable Name | Definition | Explanation | Variable name |
| Age | age of the patient | years | $x_1$ |
| Sex | sex of the patient | M: Male, F: Female | $x_2$ |
| ChestPainType | chest pain type | TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic | $x_3$ |
| RestingBP | resting blood pressure | mm Hg | $x_4$ |
| Cholesterol | serum cholesterol | 1: if FastingBS $>$ 120 mg/dl, 0: otherwise | $x_5$ |
| FastingBS | fasting blood sugar | fasting blood sugar [1: if FastingBS $>$ 120 mg/dl, 0: otherwise] | $x_6$ |
| RestingECG | resting electrocardiogram results | Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of $>$ 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria | $x_7$ |
| MaxHR | maximum heart rate achieved | Numeric value between 60 and 202 | $x_8$ |
| ExerciseAngina | exercise-induced angina | Y: Yes, N: No | $x_9$ |
| Oldpeak | oldpeak = ST | Numeric value measured in depression | $x_{10}$ |
| ST_Slope | the slope of the peak exercise ST segment | Up: upsloping, Flat: flat, Down: downsloping | $x_{11}$ |
| HeartDisease | output class | 1: heart disease, 0: Normal | $y$ |

Table 2.1: *Data Dictionary*

**Modeling All the Main Covariates** Since the response is having heart disease or not, it is a binomial distributed response. So it is suggested we could use a logistic regression model. At first, we created a logistic model of all the main effects, and the formula of the model is

$$\log \frac{p}{1-p} = \beta_0 + \sum_{i=1}^{11} \beta_i x_i$$

where $p$ is the probability to get the heart disease, and $\beta_i, i = 1, 2, \ldots, 11$ is the coefficients of the parameter, $\beta_0$ is the intercept. So we fitted the model, and got the estimated parameters shown in the table .

| | Estimate | Pr(>\|z\|) | significance |
|---|---|---|---|
| (Intercept) | -1.16 | 0.411 | not significant |
| Age | 0.0166 | 0.21 | not significant |
| SexM | 1.47 | 1.6e-07 | *** |
| ChestPainTypeATA | -1.83 | 2.03e-08 | *** |
| ChestPainTypeNAP | -1.69 | 2.34e-10 | *** |
| ChestPainTypeTA | -1.49 | 0.00058 | ** |
| RestingBP | 0.00419 | 0.485 | not significant |
| Cholesterol | -0.00411 | 0.000154 | ** |
| FastingBS | 1.14 | 3.59e-05 | *** |
| RestingECGNormal | -0.177 | 0.515 | not significant |
| RestingECGST | -0.269 | 0.443 | not significant |
| MaxHR | -0.00429 | 0.393 | not significant |
| ExerciseAnginaY | 0.9 | 0.000231 | ** |
| Oldpeak | 0.381 | 0.00131 | * |
| ST_SlopeFlat | 1.45 | 0.000703 | ** |
| ST_SlopeUp | -0.994 | 0.0272 | . |

Signif.codes: overwhelming *** strong ** moderate * borderline .

Table 2.2: *Estimation and Significance of Full Effect Model*

As we can see from the table, there are some variables that are not significant. We should drop some variables to make the model simpler. We choose backward step selection to do so. And we get our estimated parameters shown in the table .

| | Estimate | Pr(>\|z\|) | significance |
|---|---|---|---|
| (Intercept) | -1.72 | 0.0436 | . |
| Age | 0.0231 | 0.0518 | not significant |
| SexM | 1.47 | 1.36e-07 | *** |
| ChestPainTypeATA | -1.86 | 8.89e-09 | *** |
| ChestPainTypeNAP | -1.72 | 6.13e-11 | *** |
| ChestPainTypeTA | -1.49 | 0.000494 | ** |
| Cholesterol | -0.00398 | 0.000106 | ** |
| FastingBS | 1.13 | 3.41e-05 | *** |
| ExerciseAnginaY | 0.936 | 8.21e-05 | *** |
| Oldpeak | 0.377 | 0.00121 | * |
| ST_SlopeFlat | 1.46 | 0.000654 | ** |
| ST_SlopeUp | -1.03 | 0.0211 | . |

Signif.codes: overwhelming *** strong ** moderate * borderline .

Table 2.3: *Estimation and Significance of Reduced Effect Model*

From the table, we can see all the variables are significant now. here is the **model 1** formula.

$$\begin{aligned}\log\frac{\hat{p}}{1-\hat{p}} &= -1.7 + 0.023\text{Age} + 1.5\text{SexM} - 1.9\text{ChestPainTypeATA} \\ &\quad - 1.7\text{ChestPainTypeNAP} - 1.5\text{ChestPainTypeTA} - 0.0040\text{Cholesterol} \\ &\quad - 1.1\text{FastingBS} \ + 0.94\text{ExerciseAnginaY} \ + 0.38\text{Oldpeak} \\ &\quad + 1.5\text{ST\_slopeFlat} - 1.0\text{ST\_slopeUp}\end{aligned}$$

We need to test if the selected model is good enough to represent the origin mode. We did log-likelihood ratio test for the step-wise selected variables to see if the drop out is good. The null hypothesis is

$$H_0 : \beta_{\text{RestingBP}} = \beta_{\text{RestingECG}} = \beta_{\text{MaxHR}} = 0 \text{ v.s } H_1 : \text{At least one of these parameters not 0}$$

We used the formula to get LLR statistic as

$$LLR = 2(\ell(\text{full model}) - \ell(\text{reduced mode})) = 2 \times (-297.0925 + 297.9042) = 0.804$$

with degrees of freedom of 4. So we can calculate the p-value is 0.8046016, which is very high. So we cannot reject $H_0$. So we can accept the reduced model.

**Modeling the Square of Numerical Variables** Now we investigated the square of the numerical variables. We did this approach because response may have some quadratic effect of the numerical variables, and square of categorical variables do not make any difference. The odds model is

$$\log\frac{p}{1-p} = \beta_0 + \sum_{i=1}^{11}\beta_i x_i + \beta_{12}x_1^2 + \beta_{42}x_4^2 + \beta_{72}x_7^2 + \beta_{102}x_{10}^2$$

The model is fitted by R. And all the parameters and significance of them are demonstrated in the table .

| | Estimate | Pr(>\|z\|) | significance |
|---|---|---|---|
| Age | -0.029 | 0.786 | not significant |
| SexM | 1.49 | 2.28e-07 | *** |
| ChestPainTypeATA | -1.71 | 2.21e-07 | *** |
| ChestPainTypeNAP | -1.68 | 8.06e-10 | *** |
| ChestPainTypeTA | -1.38 | 0.00231 | * |
| RestingBP | -0.0286 | 0.547 | not significant |
| Cholesterol | -0.0116 | 2.87e-05 | *** |
| FastingBS | 1.08 | 0.000173 | ** |
| RestingECGNormal | -0.165 | 0.553 | not significant |
| RestingECGST | -0.296 | 0.41 | not significant |
| MaxHR | -0.0229 | 0.581 | not significant |
| ExerciseAnginaY | 1.08 | 2.09e-05 | *** |
| Oldpeak | -0.304 | 0.333 | not significant |
| ST_SlopeFlat | 1.77 | 0.000112 | ** |
| ST_SlopeUp | -7.29e-01 | 1.26e-01 | not significant |
| I(MaxHR^2) | 7.45e-05 | 6.24e-01 | not significant |
| I(Age^2) | 4.72e-04 | 6.37e-01 | not significant |
| I(Oldpeak^2) | 2.67e-01 | 1.96e-02 | . |
| I(RestingBP^2) | 1.20e-04 | 4.95e-01 | not significant |
| I(Cholesterol^2) | 2.17e-05 | 2.22e-03 | * |

Signif.codes: overwhelming *** strong ** moderate * borderline .

Table 2.4: *Estimation and Significance of Squared Model*

But there are still some useless covariates in the model. To simplify the model, a backward step-wise selection was used to the square of numerical model. The parameters and their significance of selected model are shown in the table .

| | Estimate | Pr(>\|z\|) | significance |
|---|---|---|---|
| (Intercept) | -1.05e+00 | 1.25e-01 | not significant |
| SexM | 1.50e+00 | 1.00e-07 | *** |
| ChestPainTypeATA | -1.72e+00 | 1.00e-07 | *** |
| ChestPainTypeNAP | -1.68e+00 | 0.00e+00 | *** |
| ChestPainTypeTA | -1.37e+00 | 1.98e-03 | * |
| Cholesterol | -1.18e-02 | 9.90e-06 | *** |
| FastingBS | 1.05e+00 | 2.10e-04 | ** |
| ExerciseAnginaY | 1.02e+00 | 1.92e-05 | *** |
| ST_SlopeFlat | 1.74e+00 | 1.20e-04 | ** |
| ST_SlopeUp | -7.23e-01 | 1.22e-01 | not significant |
| I(Age^2) | 2.22e-04 | 4.77e-02 | . |
| I(Oldpeak^2) | 1.70e-01 | 2.21e-04 | ** |
| I(Cholesterol^2) | 2.24e-05 | 1.44e-03 | * |

Signif.codes: overwhelming '***' strong '**' moderate '*' boderline '.'

Table 2.5: *Estimation and Significance of Reduced Squared Model*

We now apply Log-likelihood ratio test to get

$$LLR = 2(\ell(\text{full model}) - \ell(\text{reduced mode})) = 2 \times (-286.9175 + 288.5404) = 3.245983$$

with 4 degrees of freedom. So the p-value is 0.9179859, which means the reduction is highly possible. The formula for **model 2** is

$$\begin{aligned}\log \frac{\hat{p}}{1-\hat{p}} = &- 1.1 + 1.5\text{SexM} - 1.7\text{ChestPainTypeATA} \\ &- 1.7\text{ChestPainTypeNAP} - 1.4\text{ChestPainTypeTA} - 0.0012\text{Cholesterol} \\ &+ 1.0\text{FastingBS} + 1.0\text{ExerciseAnginaY} \\ &+ 1.7\text{ST\_slopeFlat} - 7.2\text{ST\_slopeUp} \\ &+ 2.2 \times 10^{-4}\text{Age}^2 + 1.7\text{Oldpeak}^2 + 2.2 \times 10^{-5}\text{Cholesterol}^2\end{aligned}$$

**Modeling the Interaction of Numerical** We also modeled the interaction between numerical terms, and the model is

$$\log \frac{p}{1-p} = \beta_0 + \sum_{i=1}^{11} \beta_i x_i + \sum_{i \in \mathcal{N}} \beta_{i2} x_i^2 + \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N}} \beta_{ij} x_i x_j$$

where $\mathcal{N} = \{1, 4, 7, 10\}$, which means the set of numerical variables. We also did backward stepwise selection to reduce the model, ane got all the coefficients demonstrated in the table

| | Estimate | Pr(>\|z\|) | significance |
|---|---|---|---|
| (Intercept) | -1.95 | 0.129 | not significant |
| Age | 0.0213 | 0.0917 | . |
| SexM | 1.42 | 4.16e-07 | *** |
| ChestPainTypeATA | -1.81 | 2.24e-08 | *** |
| ChestPainTypeNAP | -1.66 | 6.4e-10 | *** |
| ChestPainTypeTA | -1.45 | 0.000997 | *** |
| Cholesterol | -0.00386 | 0.000271 | *** |
| FastingBS | 1.2 | 1.44e-05 | *** |
| MaxHR | 0.000751 | 0.897 | not significant |
| ExerciseAnginaY | 0.993 | 6.6e-05 | *** |
| Oldpeak | 0.668 | 0.373 | not significant |
| ST_SlopeFlat | 1.810000 | 7.59e-05 | *** |
| ST_SlopeUp | -0.732000 | 1.20e-01 | not significant |
| I(Oldpeak^2) | 0.324000 | 6.27e-03 | ** |
| MaxHR:Oldpeak | -0.007750 | 1.42e-01 | not significant |

Signif.codes: overwhelming *** strong ** moderate * borderline .

Table 2.6: *Estimation and Significance of Reduced Interaction Model*

Similarly, we did a Log-likelihood ratio test to the reduction, and get

$$LLR = 2(\ell(\text{full model}) - \ell(\text{reduced mode})) = 2 \times (-288.2246 + 292.3055) = 8.161785$$

with 11 degrees of freedom. So the p-value is 0.699, which means the reduction is acceptable. And the formula for **model 3** is

$$\begin{aligned}
\log \frac{\hat{p}}{1-\hat{p}} = &- 1.9 + 0.02\text{Age} + 1.4\text{SexM} - 1.8\text{ChestPainTypeATA} \\
&- 1.7\text{ChestPainTypeNAP} - 1.4\text{ChestPainTypeTA} - 0.0039\text{Cholesterol} \\
&+ 1.19\text{FastingBS} + 7.5 \times 10^{-4}\text{MaxHR} + 1.0\text{ExerciseAnginaY} \\
&+ 1.8\text{ST_slopeFlat} - 0.73\text{ST_slopeUp} \\
&+ 0.32\text{Oldpeak}^2 \\
&- 7.7 \times 10^{-3}\text{MaxHR} \times \text{Oldpeak}
\end{aligned}$$

### 2.3.2 Analysis of the Three Models

**Receiver operating characteristics (ROC)** graphs are useful for organizing classifiers and visualizing their performance [19]. ROC curves demonstrate true positive rate on the Y axis and false positive rate on the X axis with different decision criterion. If the ROC curve can reach left-up corner in the graph, false positive rate reach zero and true positive rate of 1, which means the model has an perfect classification [20]. Therefore, the more for CROC curve lean to the right-up corner, the better performance of the model has. The ROC curves for three models are shown in the plot5.

In the ROC graph 5, three models have almost coincided into the same curve, which means three models have similar performance. To further investigate, AUC, the area under the ROC curve was also calculated to compare three models. AUC for model 1 is 0.932, for model 2 and 3 are 0.9368, 0.936. Three AUC are very similar, as we expected.

**Pseudo-R square** Pseudo-R square sometimes is used to measure the explanatory of the model. The McFadden Pseudo-R square is stated below [21].

$$R^2_{\text{MF}} = 1 - \frac{\text{LL( Full )}}{\text{LL( Null )}}$$

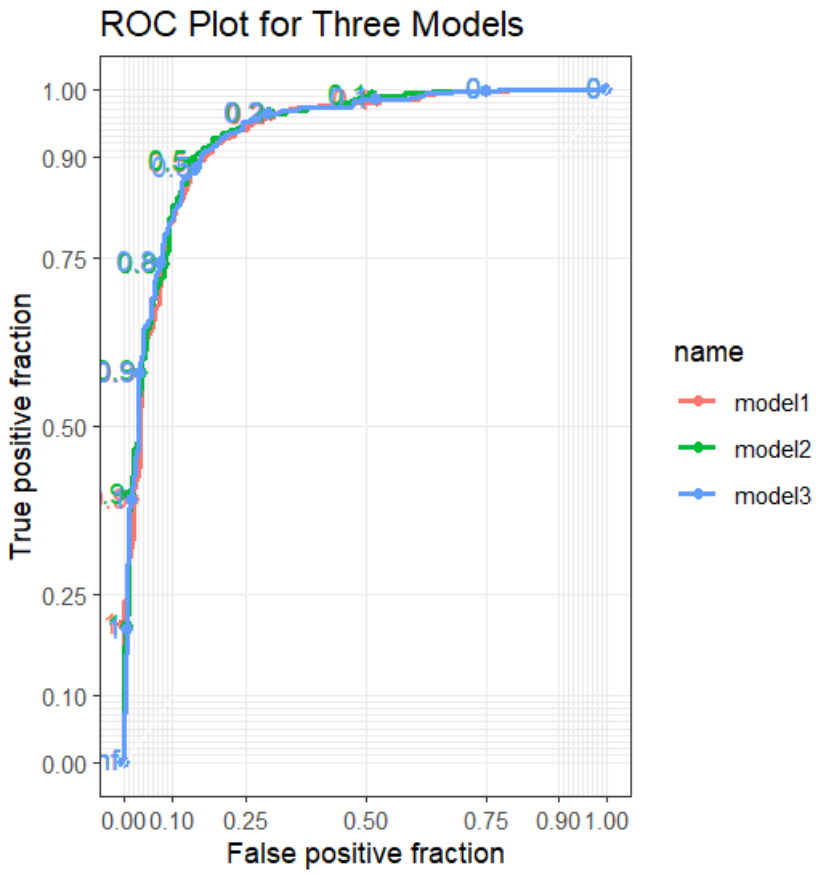



Figure 5: The ROC curves for Three Models

The McFadden Pseudo-R square is between 0 and 1. The higher value Pseudo-R square is, the higher explanatory the model have. McFadden Pseudo-R square of the three models are calculated as 0.537, 0.543 and 0.528, among which model 2 has the highest value.

**Deviance Goodness of Fit** Finally, we calculated the residual deviance for the three models. We have the null hypothesis that is $H_0$: the model fits well versus $H_1$: the model fits poorly.

Model 1 has residual deviance of 595.81 on 906 degrees of freedom. Thus the p-value is near to 1. That means model 1 is a good fit. Model 2 has the deviance of 577.08 on 905 degrees of freedom. Thus p-value is approximate 1. Model 3 has 584.61 on 903 degrees of freedom with p-value 1. Among the deviance of three models, model 2 has the least value.

**The Akaike information criterion (AIC)** AIC is a popular method for comparing the adequacy of multiple, possibly non-nested models.The objective of AIC model selection is to estimate the information loss when the probability distribution $f$ associated with the true (generating) model is approximated by probability distribution $g$, associated with the model that is to be evaluated [22]. For choosing a model from a sequence of model candidates $M_i$, $i = 1, 2, \ldots K$. The AIC is defined as

$$\mathrm{AIC}_i = -2 \log L_i + 2V_i$$

where $L_i$, the maximum likelihood for the candidate model i, is determined by adjusting the $V_i$ free parameters in such a way as to maximize the probability that the candidate model has generated the observed data [22]. Akaike showed that the model with lowest information loss has the lowest AIC [23].

The AIC for three models are 619.81, 603.08 and 614.61, respectly, among which model 2 has the lowest value.

**Binned Residual Plots** Residual plots, residual-versus-fitted values, is commonly used in linear regression models to diagnosis the model, such as to assess the validity of assumptions, to identify features not captured by the model, and to find outliers. But it is not useful for a binomial outcome model, because the response is discrete. Binned residual plots can be used to assess the binomial outcome model like logistic regression.

To construct a binned residual plot, Data are split into bins containing equal numbers of observations, and the average residual is plotted against the average predicted probability for each bin. For each bin, approximate 95% confidence limits are $\pm 2\sqrt{p(1-p)/n}$, estimated by using standard deviation of residuals in each bin. If a model is good, most(95%) of the points are expected to lie between two confidence limits [24].

The binned residual of model 2 is shown in following graph6, the plots for other two models are shown in the appendix. Thus, We can conclude that three models are good, since most of the residuals lie within the confidence limits.

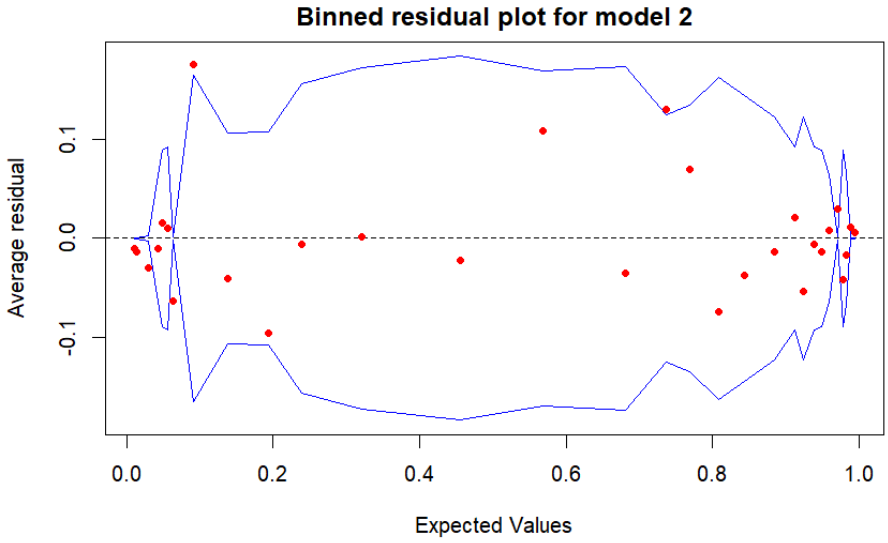



Figure 6: Binned Residuals Plots for Model 2

## 2.4 Conclusion

The best model to predict heart disease is model 2, which contains SexM, ChestPainTypeATA ChestPainTypeNAP, ChestPainTypeTA, Cholesterol, FastingBS, ExerciseAnginaY, ST_slopeFlat, ST_slopeUp, Age^2, Oldpeak ^2, Cholesterol ^2. The formula of the model 2 is displayed below.

$$\begin{aligned}\log\frac{\hat{p}}{1-\hat{p}} = &-1.1 + 1.5\text{SexM} \\ &- 1.7\text{ChestPainTypeATA} - 1.7\text{ChestPainTypeNAP} - 1.4\text{ChestPainTypeTA} \\ &- 0.0012\text{Cholesterol} \\ &+ 1.0\text{FastingBS} + 1.0\text{ExerciseAnginaY} \\ &+ 1.7\text{ST\_slopeFlat} - 7.2\text{ST\_slopeUp} \\ &+ 2.2\times 10^{-4}\text{Age}^2 + 1.7\text{Oldpeak}^2 + 2.2\times 10^{-5}\text{Cholesterol}^2\end{aligned}$$

We can have some insights about these parameters.

(1) The odds ratio of having a heart disease for males compared to females is 4.5, with a 95% confidence interval (2.6,7.8), which means male has a higher risk to get heart disease than female.

(2) The odds ratio of having heart disease for someone with exercise induced Angina compared to someone without is 2.8, with a 95% confidence interval (1.7,4.4). This fact indicates that exercise induced angina is an indication of heart disease.

(3) Someone with chest pain type of non-anginal pain, typical angina and atypical angina is less likely having heart disease comparing to someone with asymptomatic chest pain type.

(4) The odds ratio of having a heart disease is positively related to the fasting blood sugar. The odds ratio will increase by one when the sugar levels increase by one unit.

(5) If the slope of the peak exercise ST segment is up-sloping, it is least likely to have a heart disease, followed by down-sloping. Flat slope is most related to having a heart disease.

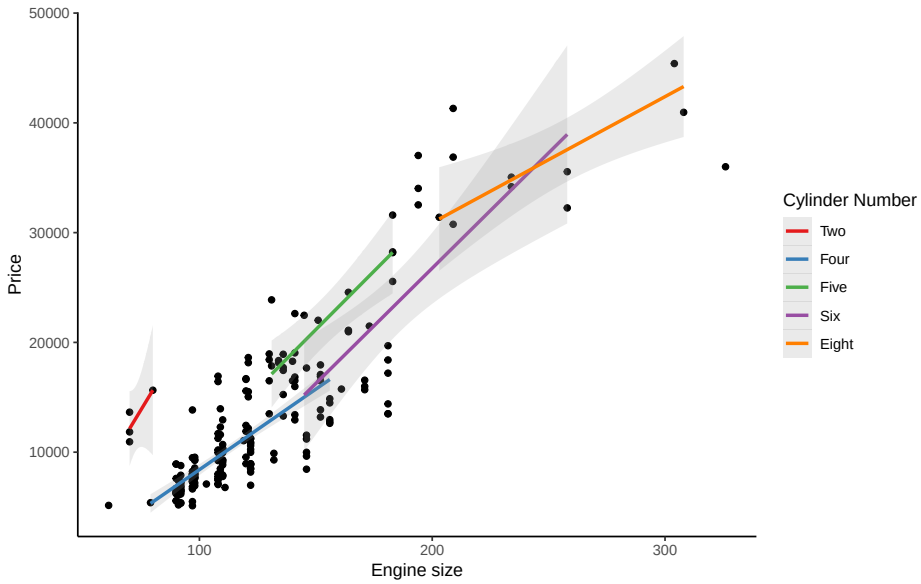
(6) The log-odds of having heart disease is quadratic related to Cholesterol level. If Cholesterol level is less than 27.2, then the log odds ratio of having heart disease is negatively related to it. Otherwise, it is positively related. This result indicates that either too high or too low of cholesterol level may related to heart disease.

(7) The odds ratio of having a heart disease is positively related to the quadratic of age. The odds ratio will increase by $2.2 \times 10^{-4}$ when the squared age increase by one unit.

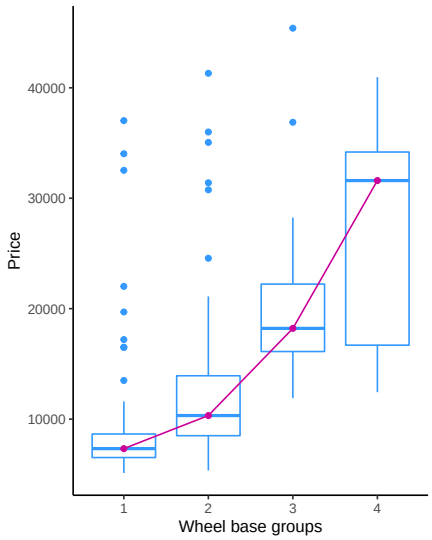
There are still some limitations in our model. In the parameter interpretation (3), someone with asymptomatic chest pain type is the least likely to have a heart disease, which is counter intuitive. How come someone without chest pain is even less likely to have a heart disease than those with chest pain? This should be further investigated.

# 3. Appendix

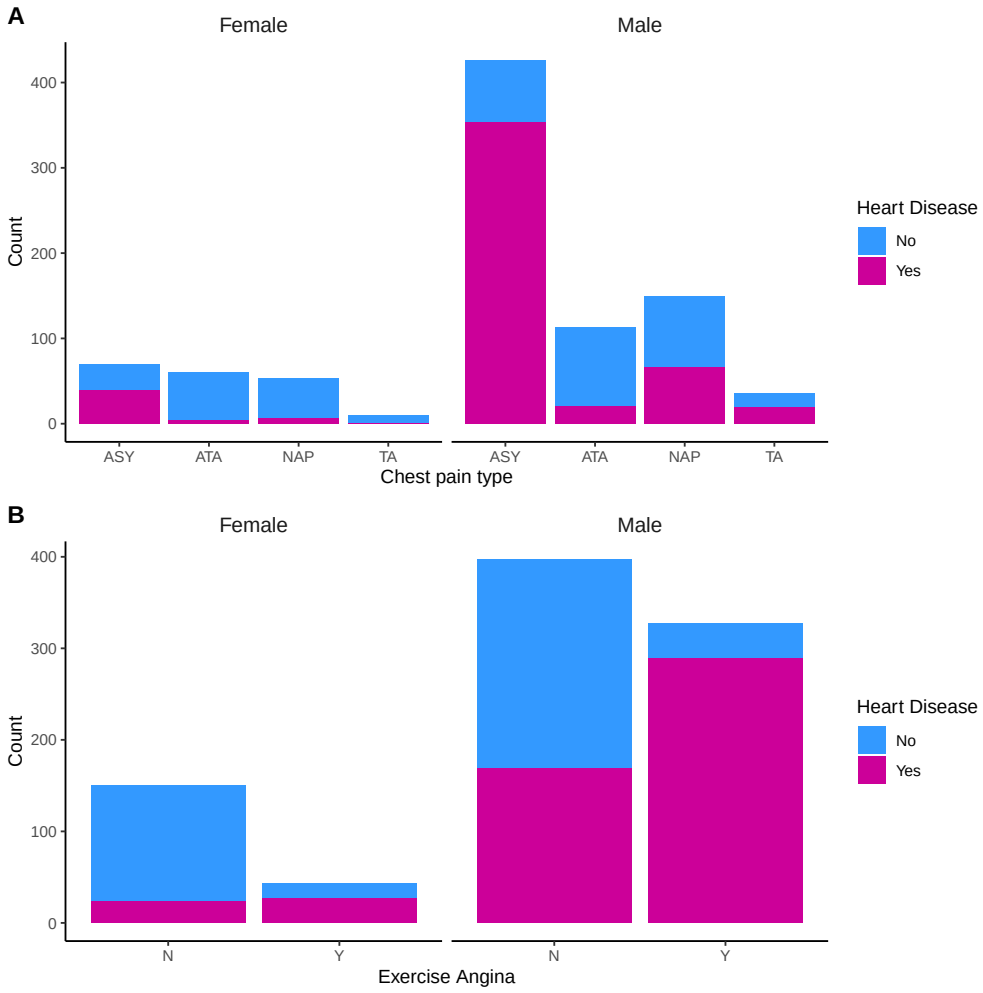
## Appendix Figures




Appendix Figure 1: The correlation between engine size and price for distinctive number of Cylinders



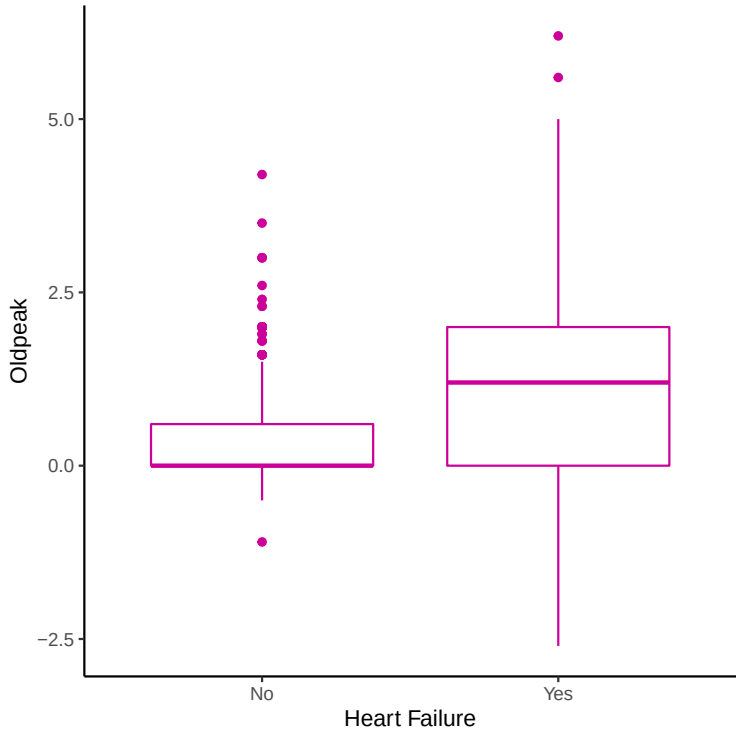

Appendix Figure 2: The quadratic increase of price with different groups of wheelbase.




Appendix Figure 3: The number of samples with different chest pain type and exercise angina. (A) The distinctive chest pain types in different genders. TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic (B) Number of samples with exercise angina with respect to their sex. Y: Yes, with exercise angina and N: No, without exercise angina.

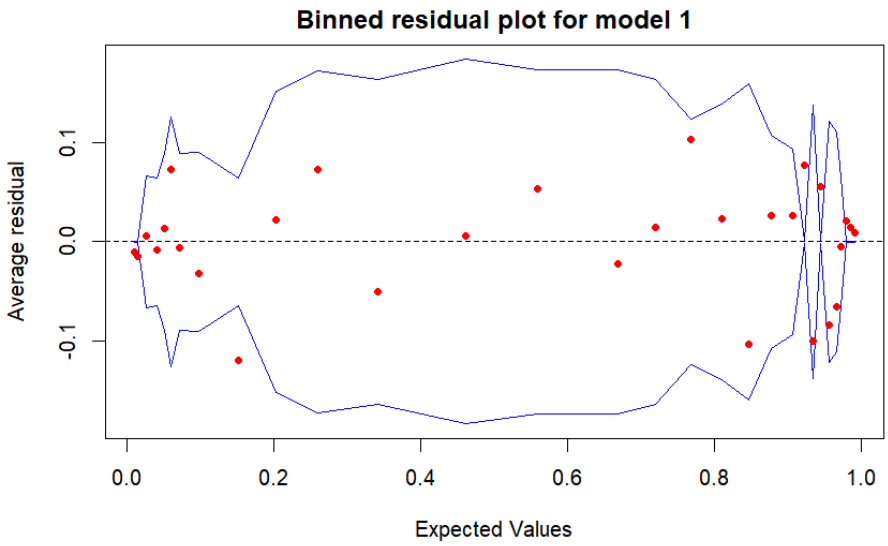
## Appendix Code

```
#car price analysis
library(readr)
car_dat <- read_csv("./price_data/CarPrice_Assignment.csv")
test_string<-toString(car_dat$CarName[1])
strsplit(test_string,"-")
test_string2<-toString(car_dat$CarName[4])
test_vec<-c(strsplit(test_string2," "))


test_car_comp<-c(rep(NA,nrow(car_dat)))
test_car_list<-vector(mode = "list", length = nrow(car_dat))

for(i in 1:nrow(car_dat)){
  test_string<-toString(car_dat$CarName[i])
  test_car_list[[i]]<-c(strsplit(test_string," "))
```



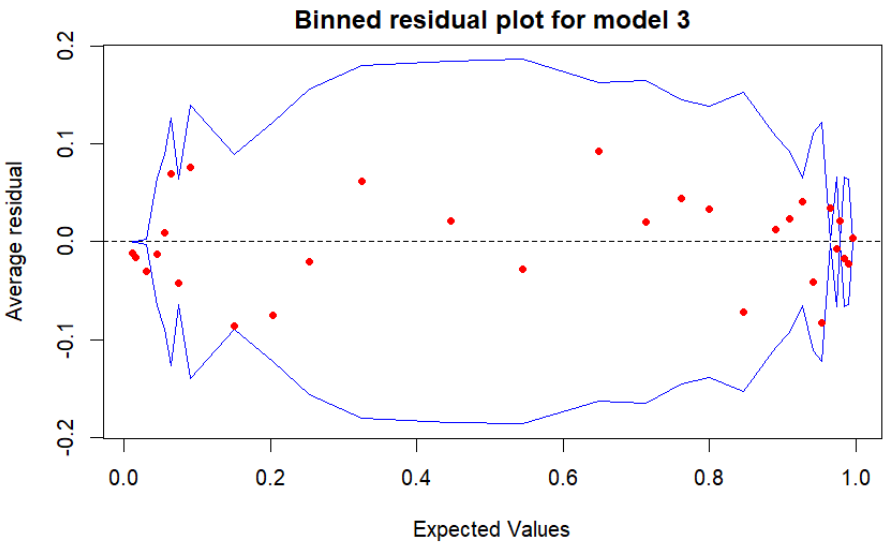

Appendix Figure 4: The relationship between having heart disease and the oldpeak.

```
  test_car_pre<-unlist(test_car_list[[i]][1])
  test_car_comp[i]<-test_car_pre[1]
}


#change the typo
car_comp_final<-c(rep(NA,nrow(car_dat)))

for(i in 1:nrow(car_dat)){
  if(test_car_comp[i]=="maxda"){
    car_comp_final[i]<-"mazda"
  }
  else if(test_car_comp[i]=="Nissan"){
    car_comp_final[i]<-"nissan"
  }
  else if(test_car_comp[i]=="porcshce"){
    car_comp_final[i]<-"toyota"
  }
  else if(test_car_comp[i]=="vokswagen"){
    car_comp_final[i]<-"volkswagen"
  }
  else if(test_car_comp[i]=="vw"){
    car_comp_final[i]<-"volkswagen"
  }
  else if(test_car_comp[i]=="toyouta"){
    car_comp_final[i]<-"toyota"
  }
```




Appendix Figure 5: Binned Residuals Plots for Model 1

```
  else{
    car_comp_final[i]<-test_car_comp[i]
  }
}

firstup <- function(x) {
  substr(x, 1, 1) <- toupper(substr(x, 1, 1))
  x
}

car_comp_final<-firstup(car_comp_final)


car_dat$car_company<-car_comp_final

View(car_dat[c(1:5),c(3,27)])




hist(car_dat$price)
#can use exponential regression or log linear model

#remove the car id and car name because they are not useful in analysis
#car name overlaps with car company variables
car_dat2<-car_dat[,-c(1,3)]

#extract character variables
character_var<-car_dat2[, sapply(car_dat2, class) == 'character']
character_var[sapply(character_var, is.character)] <- lapply(character_var[sapply(character_var, is.cha
                                                              as.factor)
```




Appendix Figure 6: Binned Residuals Plots for Model 3

```
#extract non-character variables
no_chara<-car_dat2[, sapply(car_dat2, class) != 'character']
#summary(no_chara)
car_dat3<-cbind(character_var,no_chara)

#gamma model with all explanatory variables
gamma_model_price1<-glm(price~., data= car_dat3[, -c(7,9)],family= Gamma(link = "log"))
sum_mod_gamma<-summary(gamma_model_price1)

#engine type and cylinders, fuel systems and fuel types are independent
glm(price~., data= car_dat3[, -c(7,1)],family= Gamma(link = "log"))


#using stepAIC to perform stepwise selection
library(MASS)
gamma_step_model <- stepAIC(gamma_model_price1, direction = "both",
                      trace = TRUE)
gamma_step_sum<-summary(gamma_step_model)


#glm(formula = price ~ aspiration + carbody + enginelocation +
#cylindernumber + car_company + wheelbase + carlength + carwidth +
#  carheight + curbweight + enginesize + boreratio + peakrpm +
#  citympg + highwaympg, family = Gamma(link = "log"), data = car_dat3[,-c(7, 9)])


gamma_square<-glm(price~.+ I(wheelbase^2)+I(carlength^2)+
                    I(carwidth^2)+I(carheight^2)+I(curbweight^2)+I(enginesize^2)+
                    I(boreratio^2)+I(peakrpm^2)+I(citympg^2)+I(highwaympg^2)
                , family = Gamma(link = "log"), data = car_dat3[,-c(7, 9)])
```

```
summary(gamma_square)

gamma_square_step<-stepAIC(gamma_square,direction = "both",trace=FALSE)

summary(gamma_square_step)
#lrttest for full model vs stepwise model (main effect)

(loglik_full<-logLik(gamma_model_price1))
(loglik_step<-logLik(gamma_step_model))
(test_stat1<-loglik_full-loglik_step)
(p_val1<-1-pchisq(test_stat1,8))
#Our model is the same as full model (main effect)

#lrttest for square(full model) vs stepwise square model
(loglik_sq_full<-logLik(gamma_square))
(loglik_sq_step<-logLik(gamma_square_step))
(test_stat2<-loglik_sq_full-loglik_sq_step)
(pval2<-1-pchisq(test_stat2,22))

#Our model is the same as full model(square model)

#Comparison of main effect model vs square model
AIC(gamma_step_model)
AIC(gamma_square_step)

#The square model has the lower AIC than main effects model. Hence, the square model is the best.


#model adequacy main effect mod
deviance(gamma_step_model)
df.residual(gamma_step_model)
pchisq(deviance(gamma_step_model),df.residual(gamma_step_model),lower.tail = FALSE)


#model adequnacy square mod
deviance(gamma_square_step)
df.residual(gamma_square_step)
pchisq(deviance(gamma_square_step),df.residual(gamma_square_step),lower.tail = FALSE)
#model is adequate


sum_square_step<-summary(gamma_square_step)


#95% t family Confidence Interval for estimates
coefficient_dat<-sum_square_step$coefficients[,c(1,2)]
LL<-coefficient_dat[,1]-qt(0.975,df.residual(gamma_square_step))*coefficient_dat[,2]
UL<-coefficient_dat[,1]+qt(0.975,df.residual(gamma_square_step))*coefficient_dat[,2]
conf_int_dat<-data.frame(coefficient_dat,LL,UL)
names(conf_int_dat)[3]<-"Lower Limit"
names(conf_int_dat)[4]<-"Upper Limit"
summary(conf_int_dat$Estimate[-1])
```

```
pseudo_R2_gamma_step<-1-(gamma_step_sum$deviance/gamma_step_sum$null.deviance)

pseudo_R2_square_step<-1-(sum_square_step$deviance/sum_square_step$null.deviance)



#testing with standardized residual
predict_car<-fitted(gamma_square_step)
price_actual<-car_dat3$price
std_res_car<-(price_actual-predict_car)/predict_car



#standardized residuals constant variance plot
plot(predict_car,std_res_car,xlab='fitted values',ylab='standardized residuals')
abline(h=0, col="blue")
#analysis: no obvious pattern in the standardized residuals indicating constant
#variance


#standardized residuals normality plot
qqnorm(std_res_car,ylab="sample quantiles",xlab="Theoretical Quantiles")
qqline(std_res_car)


library(ggplot2)

plot1<-ggplot(data.frame(predict_car,std_res_car),aes(x=predict_car,y=std_res_car))+
  geom_point(aes(),col="blue")+
  labs(y="standardized residuals", x = "fitted values")+
  geom_hline(yintercept = 0)



plot2<-ggplot(data.frame(std_res_car),aes(sample=std_res_car))+
  labs(y="standardized residuals",x="normal quantiles")+
  stat_qq()+stat_qq_line()



library(gridExtra)
grid.arrange(plot1,plot2,ncol=2)

#analysis: all of the standardised residuals lies on the qqline, indicating residuals
#follow approximately normal distribution



library(ggplot2)
p <- ggplot(car_dat3) + aes(x =price) +
  geom_histogram(aes(y=..density..), size=1,color="blue", fill="white", binwidth = 2000)+
  geom_density(alpha=.05, fill="red", size = 1) +
  ggtitle("Histogram of Price")+
  scale_x_continuous(expand = c(0, 0)) +
```

```
  scale_y_continuous(expand = c(0, 0)) +
  theme_classic() +
  theme(line = element_line(size = 0.5))




p2 <- ggplot(car_dat3, aes(x=car_company, y=price, fill=fueltype)) +
  geom_boxplot() +
  xlab('Company') + ylab('Price')+
  ggtitle('Car Price and Company by Fuel Type')+
  scale_fill_manual(values=c("#CC0099", "#3399FF"),
                    name = "Fuel Type",
                    labels = c("Diesel", "Gas"))+
  theme_classic() +
  theme(axis.text.x = element_text(angle = 45,vjust = 0.5))



library("dplyr")
ordered_names <- car_dat3 %>% group_by(car_company) %>% summarize(count=n()) %>% arrange(-count)
ordered_names <- ordered_names$car_company
ggplot(car_dat3) + aes(x =car_company) +
  geom_bar(size=1,color="black", fill="white")+
  scale_y_continuous(expand = c(0, 0)) +
  scale_x_discrete(limits=ordered_names)+
  theme_classic() +
  theme(line = element_line(size = 0.5))+
  ggtitle("Distribution of Car Company by Count")

#heart disease analysis
heart_disease<-read.csv('./heart.csv')
heart_disease$HeartDisease<-as.factor(heart_disease$HeartDisease)
heart_disease$ChestPainType<-as.factor(heart_disease$ChestPainType)
heart_disease$ExerciseAngina<-as.factor(heart_disease$ExerciseAngina)
heart_disease$ST_Slope<-as.factor(heart_disease$ST_Slope)
heart_disease$Sex<-as.factor(heart_disease$Sex)
summary(heart_disease)
library(pROC)
library(ggplot2)
ggplot(data = heart_disease,  aes(x = Age,  y = RestingBP, color =
                 Sex)) +
     geom_point()

library(GGally)
ggpairs(heart_disease)

heart_logit<-glm(HeartDisease~., data = heart_disease, family = binomial)
summary(heart_logit)

heart_logit_reduced = step(heart_logit)
summary(heart_logit_reduced)

LLR = 2 * (logLik(heart_logit) - logLik(heart_logit_reduced))
```

```
df = heart_logit_reduced$df.residual - heart_logit$df.residual
p_value = 1 - pchisq(LLR, df)
cat('The loglikelihood ratio test statistic is', LLR, 'and the p-value is ', p_value)

heart_logit_all_square = glm(HeartDisease~.  +I(MaxHR^2) +I(Age^2) + I(Oldpeak^2)+ I(RestingBP^2) +I(Cho
summary(heart_logit_all_square)

heart_logit_square_reduced = step(heart_logit_all_square)
summary(heart_logit_square_reduced)

LLR = 2 * (logLik(heart_logit_all_square) - logLik(heart_logit_square_reduced))
df = heart_logit_square_reduced$df.residual - heart_logit_all_square$df.residual
p_value = 1 - pchisq(LLR, df)
cat('The loglikelihood ratio test statistic is', LLR, 'and the p-value is ', p_value)
heart_logit_all_interaction = glm(HeartDisease~. +(MaxHR + Age + Oldpeak + RestingBP)^2 +I(MaxHR^2) +I(A
summary(heart_logit_all_interaction)

heart_logit_interaction_reduced = step(heart_logit_all_interaction)
summary(heart_logit_interaction_reduced)

heart_logit_interaction_reduced = glm(formula = HeartDisease ~ Age + Sex + ChestPainType + Cholesterol +
    FastingBS + MaxHR + ExerciseAngina + Oldpeak + ST_Slope +
    I(Oldpeak^2) + MaxHR:Oldpeak, family = binomial, data = heart_disease)
summary(heart_logit_interaction_reduced)

LLR = 2 * (logLik(heart_logit_all_interaction) - logLik(heart_logit_interaction_reduced))
df = heart_logit_interaction_reduced$df.residual - heart_logit_all_interaction$df.residual
p_value = 1 - pchisq(LLR, df)
cat('The loglikelihood ratio test statistic is', LLR, 'and the p-value is ', p_value)

null_model = glm(HeartDisease~ 1, data = heart_disease, family = binomial)

Pse_R2 = (deviance(null_model) - deviance(heart_logit_square_reduced)) /deviance(null_model)
cat('Pseudo-R square:', Pse_R2, '\n')
null_model = glm(HeartDisease~ 1, data = heart_disease, family = binomial)

Pse_R2 = (deviance(null_model) - deviance(heart_logit_interaction_reduced)) /deviance(null_model)
cat('Pseudo-R square:', Pse_R2, '\n')

null_model = glm(HeartDisease~ 1, data = heart_disease, family = binomial)

Pse_R2 = (deviance(null_model) - deviance(heart_logit_reduced)) /deviance(null_model)
cat('Pseudo-R square:', Pse_R2, '\n')

invisible(plot(roc(heart_disease$HeartDisease,
                   fitted(heart_logit_reduced)),
               col = "red",
               main = "ROC curves:3 models",
               legend = 'Logistic model of pure covariates'))

invisible(plot(roc(heart_disease$HeartDisease,
                   fitted(heart_logit_square_reduced)),
               print.auc = T,
               col = "blue",
```

```
               add = T))
invisible(plot(roc(heart_disease$HeartDisease,
                   fitted(heart_logit_interaction_reduced)),
               col = "yellow",
               add = T))

#Deviance Test
s = summary(heart_logit_square_reduced)
dev = deviance(heart_logit_square_reduced)

p_value = 1-pchisq(dev, s$df.residual)
cat('The deviance of the model is : ', dev,'with degree freedom of ', s$df.residual, '\n')
cat('p_value is :', p_value)

s = summary(heart_logit_interaction_reduced)
dev = deviance(s)

p_value = 1-pchisq(dev, s$df.residual)
cat('The deviance of the model is : ', dev,'with degree freedom of ', s$df.residual, '\n')
cat('p_value is :', p_value)

s = summary(heart_logit_reduced)
dev = deviance(s)

p_value = 1-pchisq(dev, s$df.residual)
cat('The deviance of the model is : ', dev,'with degree freedom ofs ', s$df.residual, '\n')
cat('p_value is :', p_value)

library(arm)
binnedplot(fitted(heart_logit_square_reduced),
        residuals(heart_logit_square_reduced, type = "response"),
        nclass = NULL,
        xlab = "Expected Values",
        ylab = "Average residual",
        main = "Binned residual plot",
        cex.pts = 0.8,
        col.pts = 1,
        col.int = "gray")

plot(heart_logit_square_reduced)

heart_logit_square_reduced$coefficients
confint.default(heart_logit_square_reduced)
```

# 4. References


1. Dobson, A.J., Barnett, A.G.: An introduction to generalized linear models. CRC press (2018)

2. Kumar, M.: Car price prediction, https://www.kaggle.com/hellbuoy/car-price-prediction

3. Palacios, F.S.: Heart failure prediction, https://www.kaggle.com/fedesoriano/heart-failure-prediction

4. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2016)

5. Analysis Unit (I&A), I.T.A.I. &: The automotive industry in the united states, https://www.selectusa.gov/automotive-industry-united-states

6. Kane, M.: US: All-electric car market share expands to 2.5, https://insideevs.com/news/526699/us-electric-car-registrations-2021h1/

7. Hasbollah, R.: Different types of car engines, https://carpart.com.au/blog/technical/different-types-of-car-engines

8. Bahadori, A.: Essentials of oil and gas utilities : Process design, equipment, and operations. Amsterdam, Netherlands : Gulf Professional Publishing (2016)

9. Wikipedia: Wheelbase, https://en.wikipedia.org/wiki/Wheelbase

10. Language, R.A., Statistical Computing, E. for: stepAIC: Choose a model by AIC in a stepwise algorithm. R Foundation for Statistical Computing, Vienna, Austria

11. National Heart, Lung, Blood Institute, N.I. of H.(NIH).: Heart failure, https://www.nhlbi.nih.gov/health-topics/heart-failure

12. Virani, S.S., Alonso, A., Benjamin, E.J., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Chang, A.R., Cheng, S., Delling, F.N., others: Heart disease and stroke statistics—2020 update: A report from the american heart association. Circulation. 141, e139–e596 (2020)

13. Benjamin, E.J., Muntner, P., Alonso, A., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Chang, A.R., Cheng, S., Das, S.R., others: Heart disease and stroke statistics—2019 update: A report from the american heart association. Circulation. 139, e56–e528 (2019)

14. Strömberg, A., Mårtensson, J.: Gender differences in patients with heart failure. European Journal of Cardiovascular Nursing. 2, 7–18 (2003)

15. Mehta, P., Cowie, M.R.: Gender and heart failure: A population perspective. Heart. 92, iii14–iii18 (2006)

16. Canto, J.G., Shlipak, M.G., Rogers, W.J., Malmgren, J.A., Frederick, P.D., Lambrew, C.T., Ornato, J.P., Barron, H.V., Kiefe, C.I.: Prevalence, Clinical Characteristics, and Mortality Among Patients With Myocardial Infarction Presenting Without Chest Pain. JAMA. 283, 3223–3229 (2000). https://doi.org/10.1001/jama.283.24.3223

17. Harvard Health Publishing, H.M.S.: Angina: Symptoms, diagnosis and treatments, https://www.health.harvard.edu/heart-health/angina-symptoms-diagnosis-and-treatments

18. Palaniappan, S., Awang, R.: Intelligent heart disease prediction system using data mining techniques. In: 2008 IEEE/ACS international conference on computer systems and applications. pp. 108–115. IEEE (2008)

19. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters. 27, 861–874 (2006). https://doi.org/https://doi.org/10.1016/j.patrec.2005.10.010

20. Receiver operating characteristic (ROC), https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

21. Walker, D.A., Smith, T.J.: Nine pseudo R2 indices for binary logistic regression models. Journal of Modern Applied Statistical Methods. 15, 848–854 (2016)

22. Wagenmakers, E.-J., Farrell, S.: AIC model selection using akaike weights. Psychonomic bulletin & review. 11, 192–196 (2004)

23. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Parzen, E., Tanabe, K., and Kitagawa, G. (eds.) Selected papers of hirotugu akaike. pp. 199–213. Springer New York, New York, NY (1998)

24. Kasza, J.: Stata tip 125: Binned residual plots for assessing the fit of regression models for binary outcomes. The Stata Journal. 15, 599–604 (2015)