# Car Price Prediction
# and
# Heart Disease Classification

Hamid Hamidi, Thet Nyein, 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].

# Car Price Prediction

## 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 dollars [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 sport 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 sport cars are missing in our data set, showing our limitation in analyzing the sport and luxury car price variables. In the following, we present a detailed description of our analysis, methods, and results.

## Data Collection and exploration

## Statistical Analyses

## Conclusion

# Heart Disease Classification

## Introduction

Heart Failure, also know as cognitive heart failure, can broadly be defined as a condition that happens when heart is unable to supply body's need for Oxygen and blood [7]. 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 [8]. Furthermore, in 2018, heart failure was mentioned on 379,800 death certificates (13.4%) [8] and cost about $30 billion annually [9].

This suggests that identifying the core health behaviors and risk factors influencing the 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 the logistic regression to model the probability of having a heart failue. Afterward, we used

## Data Collection and exploration

## Statistical Analyses

### 3 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 | 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 | $x_6$ |
| RestingECG | resting electrocardiogram results | Numeric value between 60 and 202 | $x_7$ |
| MaxHR | maximum heart rate achieved | Y: Yes, N: No | $x_8$ |
| ExerciseAngina | exercise-induced angina | Numeric value measured in depression | $x_9$ |
| 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 | $y$ |

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 shown in the table below.

**********a model

**Analysis of the 3 Models**

1. ROC, Deviance Test, Pseudo-R square,

## Conclusion

# Appendix

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

8. 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)

9. 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)