PSTAT 126 Project Step 2

Minu Pabbathi, Ziqian Zhao, Paul Zhang

2024-05-08

Introduction

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains information about patients and risk factors for heart disease as well as whether or not they eventually end up developing heart disease. For the sake of convenience, we have randomly chosen 500 observations from the original data set. Here, we are going to investigate the relationship between the variable "age" and "maximum heart rate".

Assumptions

To fit the data into a simple linear model, we will need to check that whether the residuals meet following assumptions: i) The residuals are normally distributed ($\varepsilon_i \sim N(0, \sigma^2)$); ii) the residuals have constant variance ($\text{Var}(\varepsilon_i) = \sigma^2$); iii) the residuals are independent from each other ($\varepsilon_i \perp \varepsilon_j$).

Scatter Plot of Age vs. Maximum Heart Rate

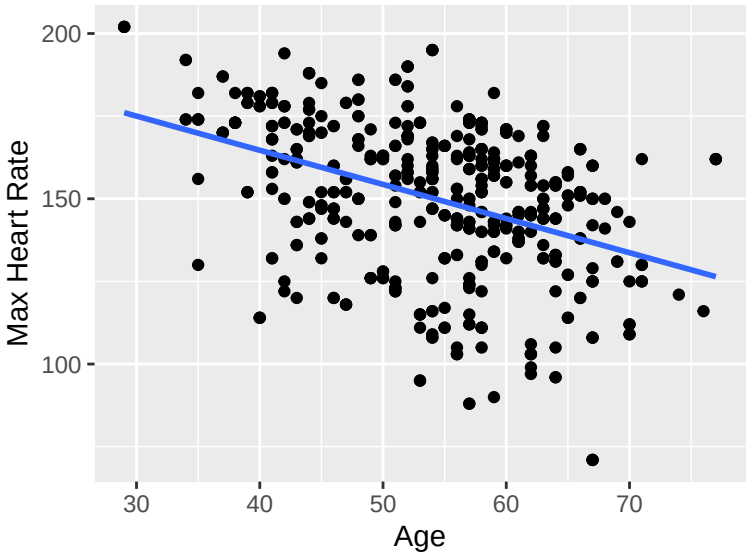

Figure 1: Relation between age and maximum heart rate achieved

Hypothesis Test

To see if the slope is significant, we are going to do a two-tailed hypothesis test. The line of best fit should be

$$y_i = \beta_0 + \beta_1 x_i$$

where β_0 is the intercept and β_1 is the slope. If the slope is zero, it means that every observation has the same response which would be of no use.

Therefore, we are going to test if the slope β_1 is zero, and conclude the significance of the slope.

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Analysis of p-value and confidence interval

p-value

```
##
## Call:
## lm(formula = thalach ~ age, data = heart_disease_data_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.773 -11.544   4.093  15.389  44.791
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 206.0176     5.6674  36.351   <2e-16 ***
## age          -1.0335     0.1034  -9.993   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.29 on 498 degrees of freedom
## Multiple R-squared:  0.167,  Adjusted R-squared:  0.1654
## F-statistic: 99.86 on 1 and 498 DF,  p-value: < 2.2e-16
```

From the summary table, we summarized the model as

$$\text{Heart Rate}_i = 206 - 1.03 \cdot \text{Age}_i$$

The p-value for $\beta_1 < 2 \times 10^{-16}$, so we reject null hypothesis and conclude that there is association between age and maximum heart rate achieved.

Construct Confidence interval

```
## [1] -1.2366529 -0.8303449
```

The confidence interval for β_1 is (-1.24,-0.83) at a significant level of 95%. This is to say that repeatedly estimating β_1 with same procedure, 95% of the estimation lies between this interval. Since 0 is not included in this interval, it is most likely that $\beta_1 \neq 0$. Therefore, we conclude that there is some association between age and maximum heart rate achieved.

Residual Analysis and R^2

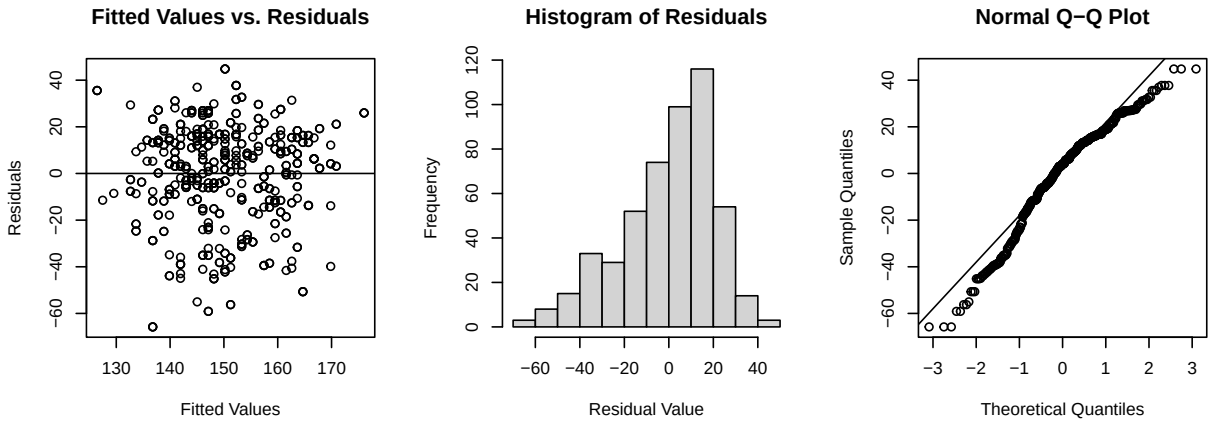

Figure 2: Fitted Values vs. Residuals, Histogram of Residuals

When examining a plot of the residuals against the fitted values, the points appear to be randomly scattered, indicating that they have mostly constant variance and linearity. The histogram of the residuals appears slightly skewed to the left. However, it does not strongly deviate from a normal distribution enough to violate the assumptions of a linear model. Overall, the residuals indicate that the model is good fit and meets the four assumptions for inference.

The R^2 value is 0.167, meaning that approximately 17 percent of the variance in maximum heart rate is explained by age.

Conclusion

The results were mostly as expected. The relationship between age and maximum heart rate has been studied for a while, and our results are consistent with the literature. The residuals met the assumptions for inference - that is, they were normally distributed with $\mathbb{E} = 0$, had a constant variance of σ^2, and were independent. $\hat{\beta}_1$ was found to be -1.0335 with a 95% confidence interval (-1.24, -0.83), meaning our results were statistically significant. Ultimately, 17% of the variance in maximum heart rate was explained by age. Going forward, we would like to examine if adding another predictor changes the accuracy of the model.