STA 302 PROJECT PROPOSAL

Analysis of a dataset for Heart Attack patients

Manjima Banerjee
Student Number - 1006926230

Introduction of the Dataset

- The data for our project is taken from all the patients who have suffered from a heart attack.
- It gives us detailed information on the counts for their integral health factors such as Cholesterol, Resting Blood Pressure, Heart Rate, Fasting Blood Sugar, Chest Pain Type and many others.
- These factors are strongly linked to an individual's pulmonary system and therefore help us check tendency of an individual having a heart attack, find probable causes for it and also focus on the necessary solutions or ways to reduce vulnerability to it.

Research Question

‘Does the level of Cholesterol increase with age of patients who have had a heart attack?’

- In US, about 805,000 people have a heart attack every year and globally 1 in 14 people are living with a heart or circulatory disease.
- This question can help the population and medical professionals to understand the relation of cholesterol with age and how it indirectly affects the tendency of having a heart attack.
- We can find what age groups are more vulnerable to having an attack and need to be extra cautious for their health (focus on diet and lifestyle)
- As this health issue is so common and concerning at the same time, it is integral to focus on keeping the health factors in the dataset under the right levels.
- The variables in focus for this question are **age of patient** (in years) and the **level of cholesterol** (mg/dl).

Background Research and Keyword Search

- For my research to provide more foundation, I focused on peer reviewed articles specifying studies of age groups that have a higher or lower levels of cholesterol and were seen to be at greater risk for a cardiovascular disease.
- Some keywords to be used were 'peer reviewed articles', 'HDL'(High-density lipoprotein) and 'LDL'(low-density lipoprotein) cholesterol, 'serum cholesterol with age', 'heart attack and age', 'coronary heart disease and age', 'coronary heart disease and cholesterol' etc.
- The total search hits from these keywords were 3,410,000 and the databases that were searched were around 10 to 12 articles or papers.

Background Research and Summary of Literature

‘Determinants of the Increase of Serum Cholesterol with Age: A Longitudinal study’

- This article investigates the increase of serum cholesterol with age over a period of 6-10 years by keeping habitual food consumption and body mass index across participants constant and similar.
- It suggested that as people grow older, they become more sensitive to cholesterol elevating factors in the diet.
- From age 20 onwards, serum cholesterol creeps up and along with this, we see that the accumulation of fat adipose tissue is larger in older people, increasing their chances of having a higher level of cholesterol.

Background Research and Summary of Literature

‘Total Serum Cholesterol Levels and Mortality Risk as a Function of Age’

- This article aims to evaluate relationship between serum cholesterol level and all-cause, coronary heart disease (CHD) as a function of age.
- The relationship with CHD mortality was significantly positive at ages 40, 50, and 60 years but reduced in effect with age.
- This relationship was not quite significant, at age 70 and 80 years and negative.
- It recommends physicians to be cautious about initiating cholesterol-lowering treatment in men and women above 65 to 70 years of age as our relationship was not quite significant for those ages.

Background Research and Summary of Literature

‘Cholesterol and Cardiovascular Disease in the Elderly. Facts and Gaps’

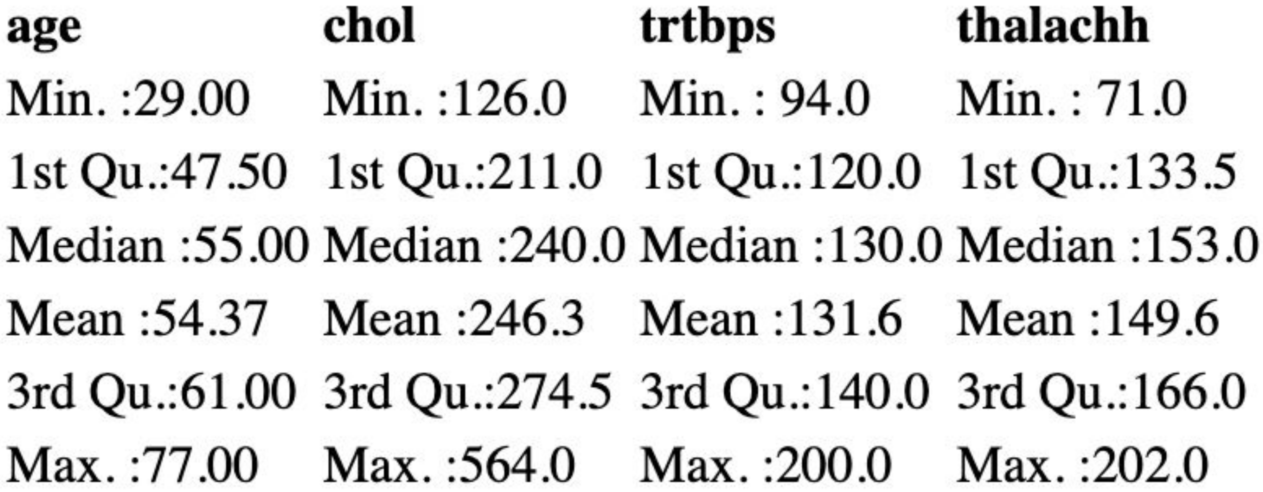
- Hypercholesterolemia (condition where the body is not able to get rid of the type of cholesterol that builds in the arteries causing heart diseases) is a major cardiovascular risk factor that increases the incidence of atherosclerotic diseases (heart attack, stroke) in adults.
- Majority of cardiovascular disease cases and deaths occur in the elderly (>65 years) and very elderly (> 80 years).
- An increase in aging population, combines to produce higher population levels of cholesterol, therefore increasing the risks of cardiovascular disease and death

Data Description and Variables Chosen

- This open source dataset was obtained from Kaggle and contains data for 303 patients (observations) and 14 health variables to study.
- The variables in the dataset are as follows -
  - **Age : Age of the patient**
  - Sex : Sex of the patient
  - exang: Exercise induced angina (1 = yes; 0 = no) - pain in chest coming from stress or exercise
  - ca: Number of major vessels (0-3)
  - cp : Chest Pain type (1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic)
  - trtbps : Resting blood pressure (in mm Hg)
  - **chol : Cholesterol in mg/dl fetched via BMI sensor**
  - fbs : Fasting blood sugar > 120 (mg/dl) (1 = true; 0 = false)
  - rest_ecg : resting electrocardiographic results (0: Normal, 1; Abnormality, 2: Probable or definite hypertrophy)
  - thalach : maximum heart rate achieved
  - target : 0= less chance of heart attack 1= more chance of heart attack
- **Age of Patient and Cholesterol level (mg/dl) are the variables that will be used in our study**

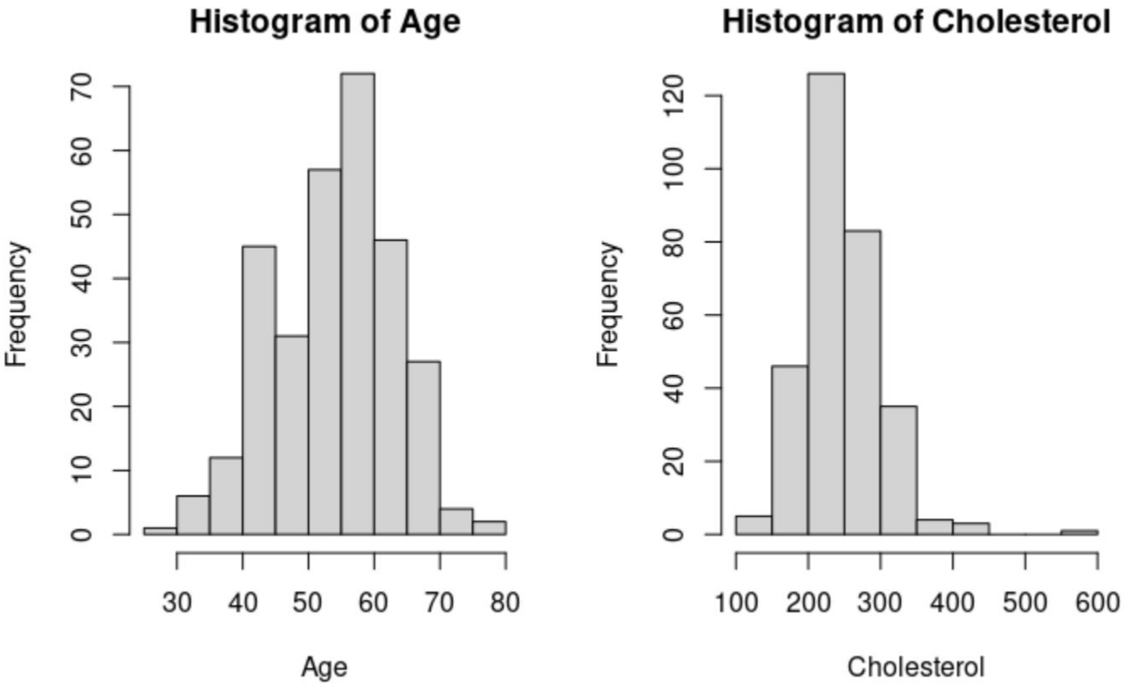
Exploratory Data Analysis (EDA)

Numerical summary of necessary variables

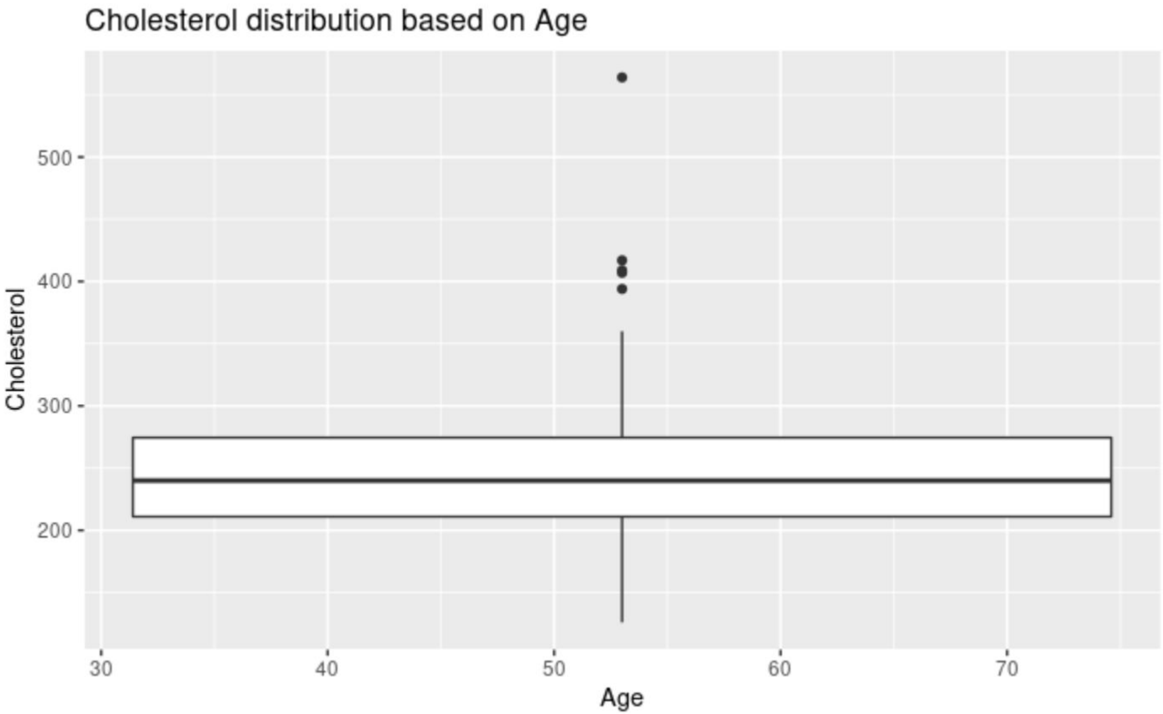


```
age            chol            trtbps          thalachh
Min.   :29.00  Min.   :126.0   Min.   : 94.0   Min.   : 71.0
1st Qu.:47.50  1st Qu.:211.0   1st Qu.:120.0   1st Qu.:133.5
Median :55.00  Median :240.0   Median :130.0   Median :153.0
Mean   :54.37  Mean   :246.3   Mean   :131.6   Mean   :149.6
3rd Qu.:61.00  3rd Qu.:274.5   3rd Qu.:140.0   3rd Qu.:166.0
Max.   :77.00  Max.   :564.0   Max.   :200.0   Max.   :202.0
```

Histogram for Age and Cholesterol



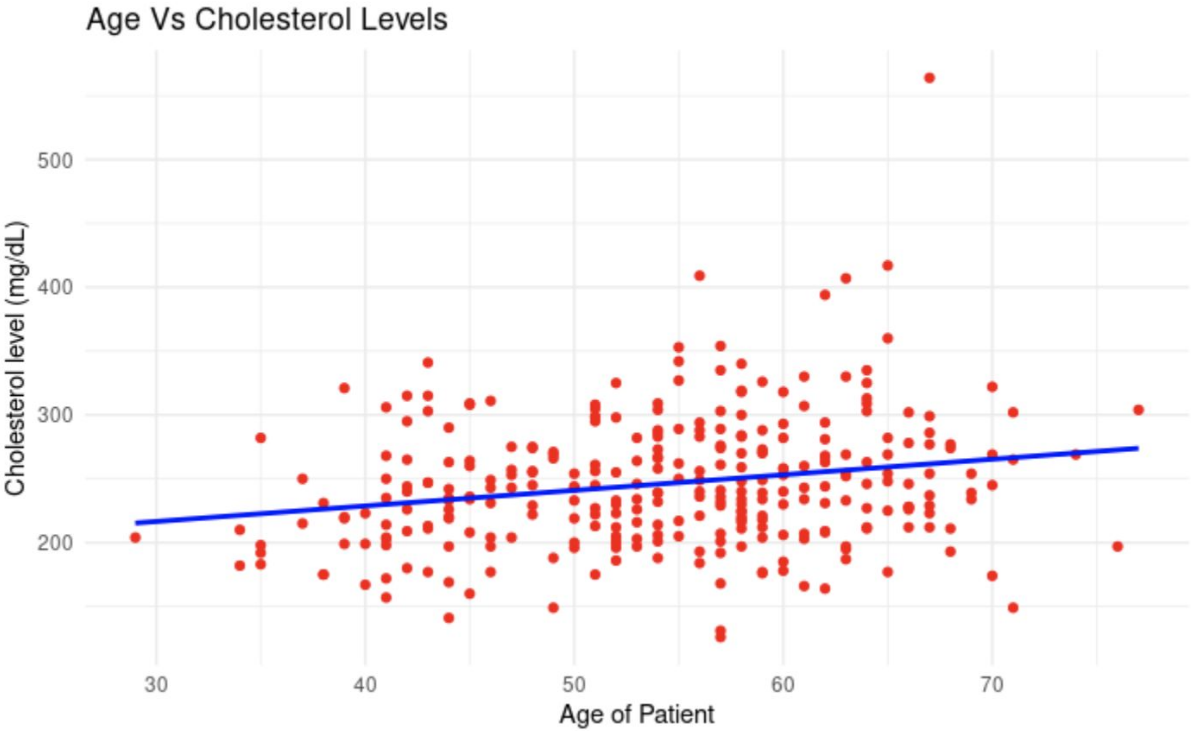
Boxplot for Cholesterol VS Age



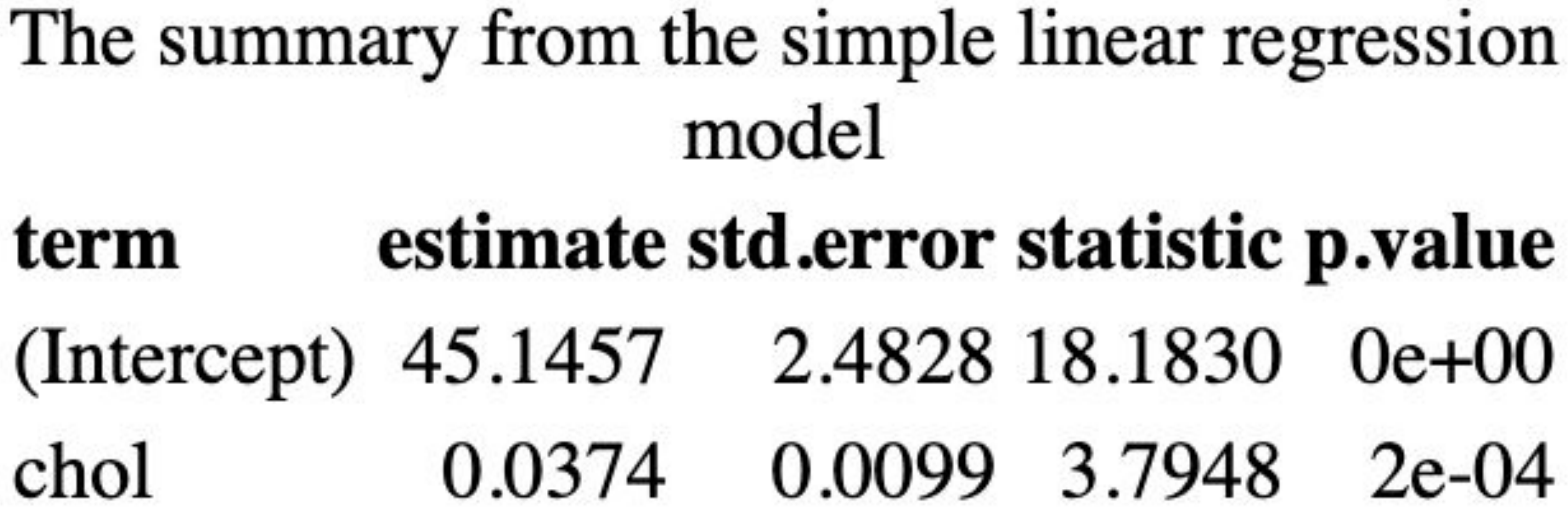
Analysis Of Numerical and Visual Summaries

- We see that the mean (average) age of patients is 54.37 years where the youngest patient is 29 years and eldest patient is 77 years. The middle age for the patients is 55 years.
- The mean (average) of cholesterol levels for patients is 246.3mg/dl with the lowest level recorded as 126 mg/dl and the highest as 564 mg/dl. The middle level of cholesterol is 240mg/dl.
- In both our histograms for age and cholesterol, we see that they are both mostly normally distributed and are both skewed to the right (few extreme values on the right side of the graph)
- The outliers for age lie from 70-80 years and we also see patients having an attack between age 20-30 years on the lower extreme.
- There are a few higher extreme values for cholesterol lying from 400-600 mg/dl and 100-150 mg/dl on the lower side of the graphs acting as outliers.
- For our boxplot, we see that the highest outlier for cholesterol was greater than 550mg/dl.

Linear Regression Model



Numerical Summary of our Linear Regression model



The summary from the simple linear regression model

term	estimate	std.error	statistic	p.value
(Intercept)	45.1457	2.4828	18.1830	0e+00
chol	0.0374	0.0099	3.7948	2e-04

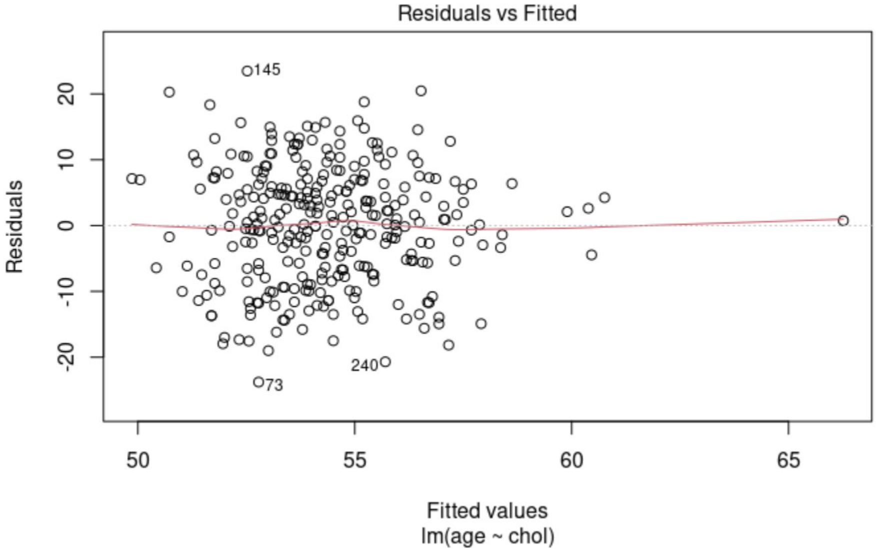
Answering the Research Question

- The linear regression model helps us find the relationship between age and cholesterol levels.
- We see that the regression line in our model is positively sloping and as age increases, the level of cholesterol also increases.
- The slope of the line is 0.0374, which suggests that with every one year increase in age, the cholesterol level of the patient increases by 0.0374mg/dl.
- The correlation between age and cholesterol level comes out to be 0.213678 which suggests that there is a positive correlation between age and cholesterol.
- We can also perform hypothesis testing using the p values given in the summary for our linear regression model

Assumptions for Linear Regression Model

1. Linearity

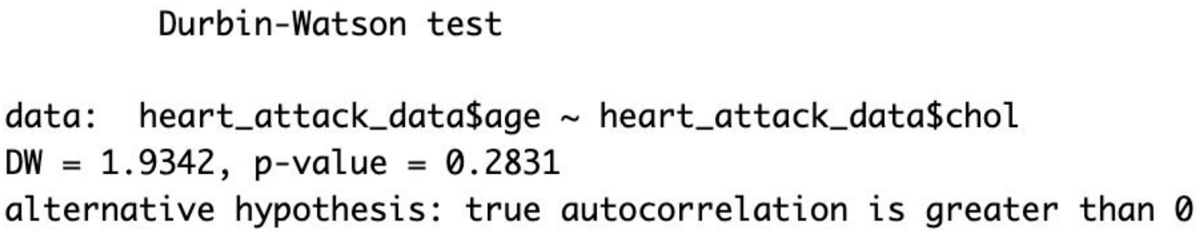
Relation between age and cholesterol levels of patients must be linear.



The Residual vs Fitted plot (residual is the difference between observed value and fitted value on our regression line) for our regression model has an almost straight flat line which **satisfies our assumption for linearity**.

2. Independence

- Independence means that there is no relation between the different observations that we have in our dataset. (random selecting of patients for their age and cholesterol levels)
- We use the Durbin Watson Test to find whether the observations (errors) are auto correlated.
- We have a p-value of 0.2831 which is larger than 0.05, implying that we have enough evidence that our independence assumption is met (fail to reject the null hypothesis).



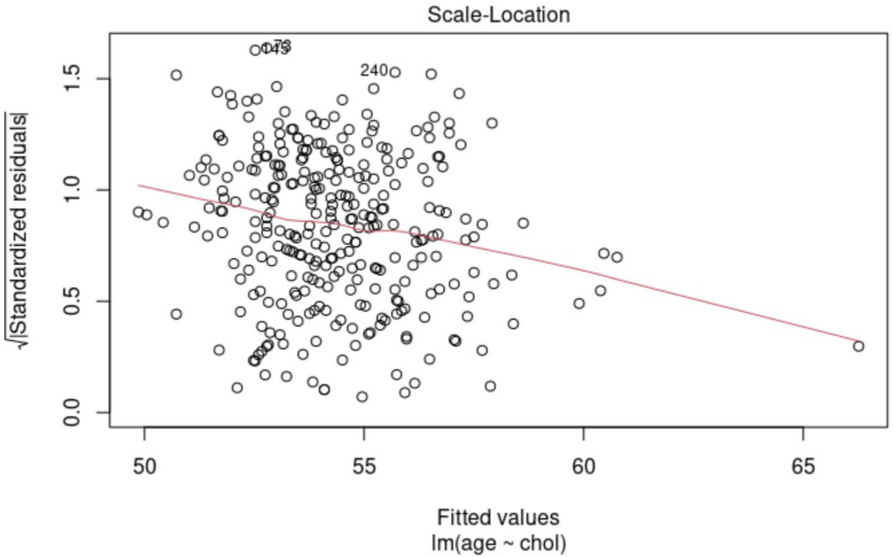
```
	Durbin-Watson test

data:  heart_attack_data$age ~ heart_attack_data$chol
DW = 1.9342, p-value = 0.2831
alternative hypothesis: true autocorrelation is greater than 0
```

3. Homoscedasticity

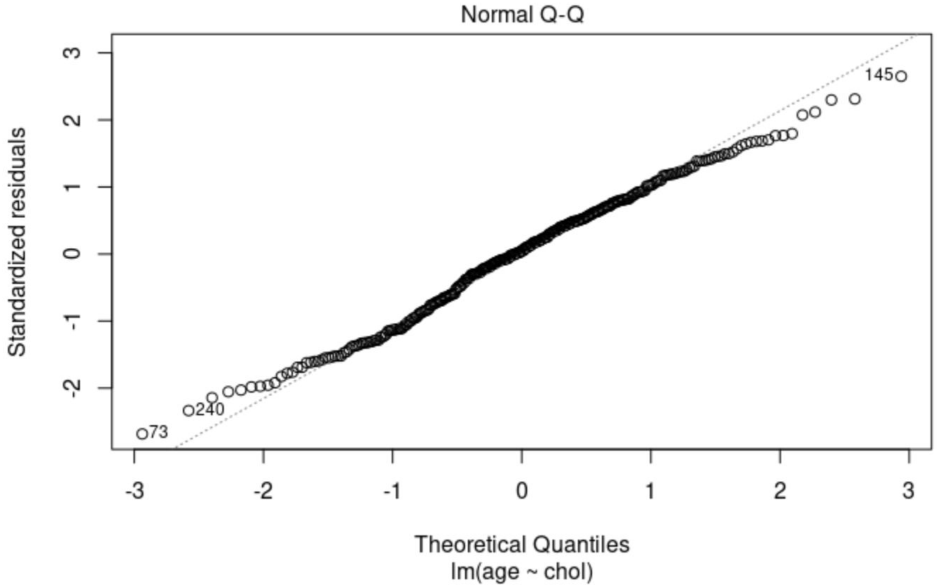
- A condition in which the variance of the residual, or error term, in a regression model is constant.
- For our scale-location plot, we see that with every increase in observation, the variance of residual error value decreases (not constant)
- Therefore, we see that our model does not satisfy the

Homoscedasticity assumption.



4. Normality

- Residual error values should follow a normal distribution that can be shown with a normal QQ plot.
- We see that all our points lie on or very close to the line on the plot suggesting that the deviations satisfy the normal distribution and hence our assumption is satisfied.



References for Research Literature

- BERNS, M. A. M., DE VRIES, J. H. M., & KATAN, M. B. (1988). Determinants of the Increase of Serum Cholesterol with Age: A Longitudinal Study. *International Journal of Epidemiology*, *17*(4), 789–796. https://doi.org/10.1093/ije/17.4.789

- Kronmal, R. A., Cain, K. C., Ye, Z., & Omenn, G. S. (1993). Total serum cholesterol levels and mortality risk as a function of age. A report based on the Framingham data. *Archives of Internal Medicine*, *153*(9), 1065–1073. https://pubmed.ncbi.nlm.nih.gov/8481074/

- Félix-Redondo, F. J., Grau, M., & Fernández-Bergés, D. (2013). Cholesterol and cardiovascular disease in the elderly. Facts and gaps. *Aging and Disease*, *4*(3), 154–169. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3660125/

- Online Link for Dataset: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset