BN2102 ASSIGNMENT 2: DATA ANALYSIS REPORT

PROBLEM 1

PART A

Objective: To determine whether there is any significant difference between Singapore and Malaysia in terms of proportion of satisfied patients.

Since the resulting table containing the nominal data is of size 2x2, Yates correction for continuity will be used in computing the chi-squared value.

Yates chi-squared value	0.8235435490709171
chi-squared critical	3.283020286759539
p-value a	0.3641461483516225

Conclusion: Since chi-squared value < chi-squared critical, data failed to support any significant difference between Singapore and Malaysia in terms of proportion of satisfied patients.

PART B

Objective: To find the number of Thai respondents, or n, who would have to say ”No” in order for this current study to have a p-value that is closest to 0.1.

To do this, we have to iterate through the potential values of n, starting from 14 with an increment of 1. The associated p-value is then computed through each iteration, until the computed p-value is within an acceptable range from 0.1.

n	28.0
p-value b	0.10858543122564468

Conclusion: We are 93% confident that the number of Thai respondents who would have to say “No” for the p-value to be close to 0.1 is 28. Specifically, the p-value associated with this n is 0.1086.

PART C

Objective: Determine whether there is a statistically significant trend of QLS with respect to age among patients, where:
NULL hypothesis (H_0): $\beta_1 = 0$ (there is no correlation between QLS and age)
alternative hypothesis (H_1): $\beta_1 \neq 0$ (there is a correlation between QLS and age)

t-statistic	-2.1921164827646313
t-critical	1.8203851211186555
p-value c	0.029374091444891448

Since the |t statistic| is greater than the t critical value, we can reject the NULL hypothesis and accept the alternative hypothesis. Since the NULL hypothesis was rejected, the associated p-value will be lower than 0.07. Specifically, according to our data, p-value = 0.0294.

Conclusion: We are 93% confident that there is a statistically significant trend of QLS with respect to age among the patients ($\beta_1 \neq 0$).

PART D

Objective: Find the interval estimate of the predicted QLS score for a 25-years old patient.

The original data only contains information on ages 30 to 70.5. Thus, predicted QLS score for a 25-year-old patient must be calculated by extrapolation from the linear regression model.

Upper bound predicted QLS score	120.76695669142345
Lower bound predicted QLS score	69.22771529287957
S_{ypred}	14.15613674288664

One thing to consider in calculating the predicted QLS score is that there is a high uncertainty associated with extrapolated data as compared to existing data. This can be observed when the formula for variance of y_h and y_{pred} is compared.

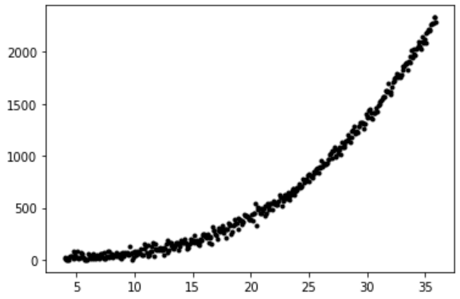
Conclusion: We are 93% confident that the true QLS score for a 25-years old patient is in between 120.77 to 69.23. However, this predicted QLS score is associated with a higher uncertainty, since it is extrapolated from existing data.

PROBLEM 2

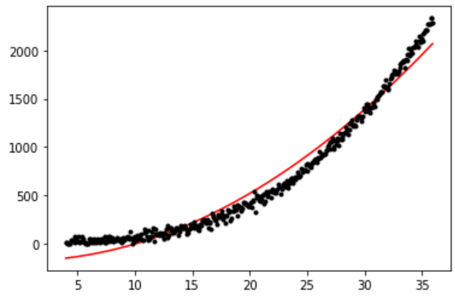
PART A

Objective: To find the values of β_0 and β_1 that best approximate the observations, where $y(x) = \beta_0 + \beta_1 * f(x)$.

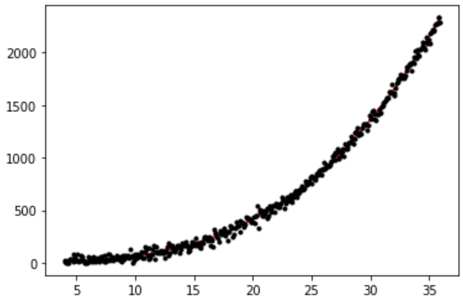
To determine the most suitable f(x) for the regression model, we need to plot the data, with the potential regression models, and determine the best f(x) to use in the regression model.



Experimental data



$f(x) = x^2$



$f(x) = x^3$

It is evident that the data points are closest to the regression model when $f(x) = x^3$.

β_0	13.829722333029022
β_1	0.04992217397562325
SSE	417843.81681322335

SSE is a collective sum. The computed SSE value may seem large, but considering the large amount of data, and comparing the SSE value associated with other regression models, this SSE value may be considered minimal. For example, if $f(x) = x^2$ is used in the regression model, the SSE will be around 4774963.

Conclusion: Using $f(x) = x^3$ in the regression model, the value of β_0 and β_1 that best approximate the observations is 13.8297 and 0.0499.

PART B

Objective: To compute the confidence intervals of β_0 and β_1 using a 98% confidence level.

Upper bound β_0	19.13631984309981
Lower bound β_0	8.523124822958234
Upper bound β_1	0.05020699575201406
Lower bound β_1	0.04963735219923244

Conclusion: We are 98% confident that $8.5231 < \beta_0 < 19.1363$ and $0.0496 < \beta_1 < 0.0502$.