STAT 403 Final Project

This is a final project completed by Xiaoyang Yang (ID: 301364025) for the purpose of STAT 403.

Background

Smoking cigarette is a common human behavior that, according to many research and studies, has a few negative impacts on health. Cigarettes is part of the human culture for its psychological effect, as stereotypes often relates cigarettes with stresses; they do calms people down. The negative effects of cigarettes, however, is a non-neglectable issue, as cigarettes can, commonly known, induce addiction, permanent damage to respiratory system, or, more severely, chronic diseases.

Frequent exercise is beneficial to health. Research shows that people who do cardio exercises regularly has a lower chance of suffering from diseases. Exercises are, in the common view, related to words like “healthy”, “energic”, “uplifting”. Our interest is to find out whether exercise and smoking have a correlation to human systolic blood pressure: Let us say a healthy living person who runs regularly, does the person’s smoking behavior contribute to the high blood pressure he or she may have? Among the health hazards brought by cigarette and the benefits of sporting, we want to find out if smoking cigarette will affect one’s circulatory system, to be exact, one’s blood pressure.

The objective of this project is to research whether there is a correlation between human systolic blood pressure and the behavior of smoking cigarette, and exercise among males between the ages of 40 - 80. The nature of this projects requires extensive and detailed data on human blood pressure and subjects’ habits (whether the sample subjects smoke cigarette). Limited by the current situation and ability to carry out such a survey in real life, however, the author instead turns to a simulated population, the islands.

“The islands” is a simulated world with its own history, events, and residents. Individual resident in the simulation has distinct and random traits on many deterministic characteristics. The characteristics, or “the entry” of an individual ranges from basic physical information, like height, age, weight, to abstract human behaviors, such as marriage status, sexual orientation, or preference of the weekend activity. The user of “the islands” simulation could obtain such information either by examine the virtual resident’s personal profile or by interviewing. Given the factors of large population, well representation of human behavior in this simulation, the author considers it is feasible to use the simulation to gather data that are needed for this question.

Data collection

From the many residents of “The islands” simulation world we selected a simple random sample. We first selected a random sample of houses, which is the residencies of “The island” residents, each house may contain one or more individuals. From each selected house, we randomly selected one eligible household member, in this case, male in the 40 – 80 age group.

Using the previously described method, we recorded our selected houses, around 180 of them, on a spreadsheet. The column of the spreadsheet are identifiers such as location (of the three islands), city, household numbers. We noticed that although some recorded household has multiple residents, none of which falls in the category of test subject: male between 40 and 80 years old. To address this, we delete the household. After handling the null values in the selected data set. We are left with 80 individuals.

Objective formulation

As discussed before, the objective of this project is to find out whether smoking and sports can affect one's blood pressure. But to restate this general question, we want to find out:

1. Does smoking cigarette influence the systolic blood pressure of males between the ages of 40 – 80, that live on the islands?

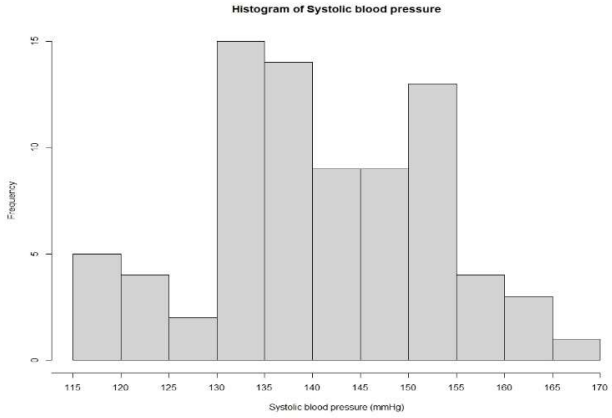
2. Does the amount of moderate exercise (in minutes) completed during the week influence the systolic blood pressure of males between the ages of 40 – 80, that live on the islands?

Hence, we have are quantitative response variable: systolic blood pressure; our categorical explanatory variable: smoking behavior, measured by yes or no; and our quantitative explanatory variable: the amount of moderate exercise per week, measured by minutes.

We are targeting all males who live on the Islands between 40 – 80 years of age and we are drawing the data from South Island, one of the largest islands in the simulation world. South Island is a good representation of the overall target population as its immigration background and largest number of cities/towns, which is an ideal subset of the target population.

Data

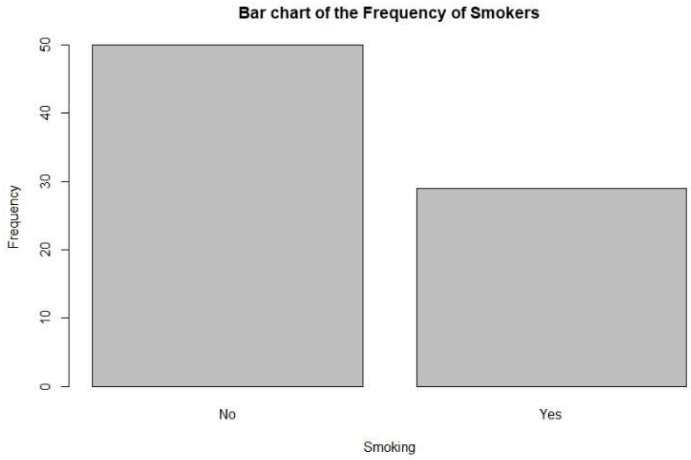
We start by look at the distribution of the response variable: systolic blood pressure.



From the graph we can see the systolic blood pressure (mmHg) looks to be bimodal with the largest peaks reaching a frequency of 15 between 130 – 135 mmHg, 14 between 135 – 140 mmHg, and a finally a frequency of 13 between 150 – 155 mmHg.

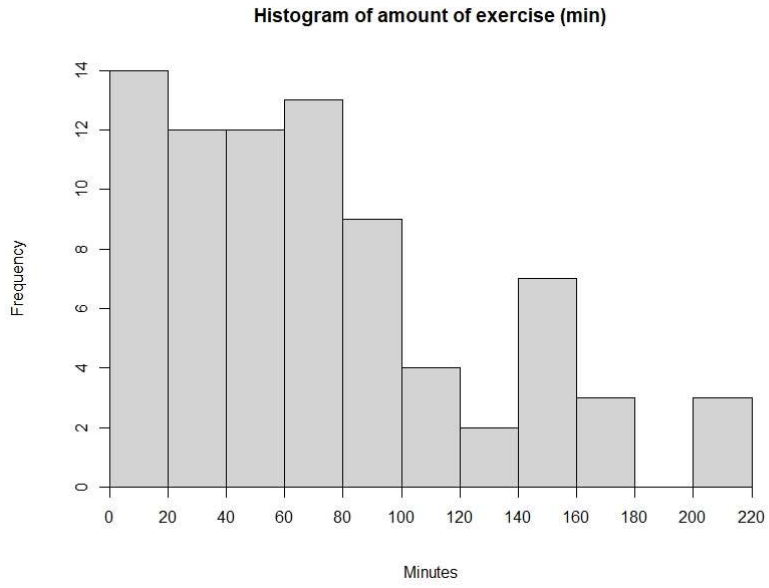
Then we shall look at the distribution of the explanatory variables.

Categorical explanatory variable: smoking



From the plot we can see a majority of the individuals sampled are non-smokers. Out of the total saple size of 80 individuals, 50 were non-smokers, and 30 were smokers.

Quantitative explanatory variable: exerpcise completed in one week, in minutes.



From the plot we can see the distribution is skewed to the right. The lowest amount of exercise has the largest frequency of 14.

Analysis

In order to example whether there is a difference between smoking and non-smoking's impacts on the systolic blood pressure, we shall perform a two-sample t-test. Our assumption is that smoking males between the ages of 40 – 80 years old tends to have a higher blood pressure level that those of the non-smokers.

Hypothesis:

μ_1 = smokers α-level:0.05

μ_2 = non-smokers

H0: $\mu_1 = \mu_2$

Ha: $\mu_1 > \mu_2$

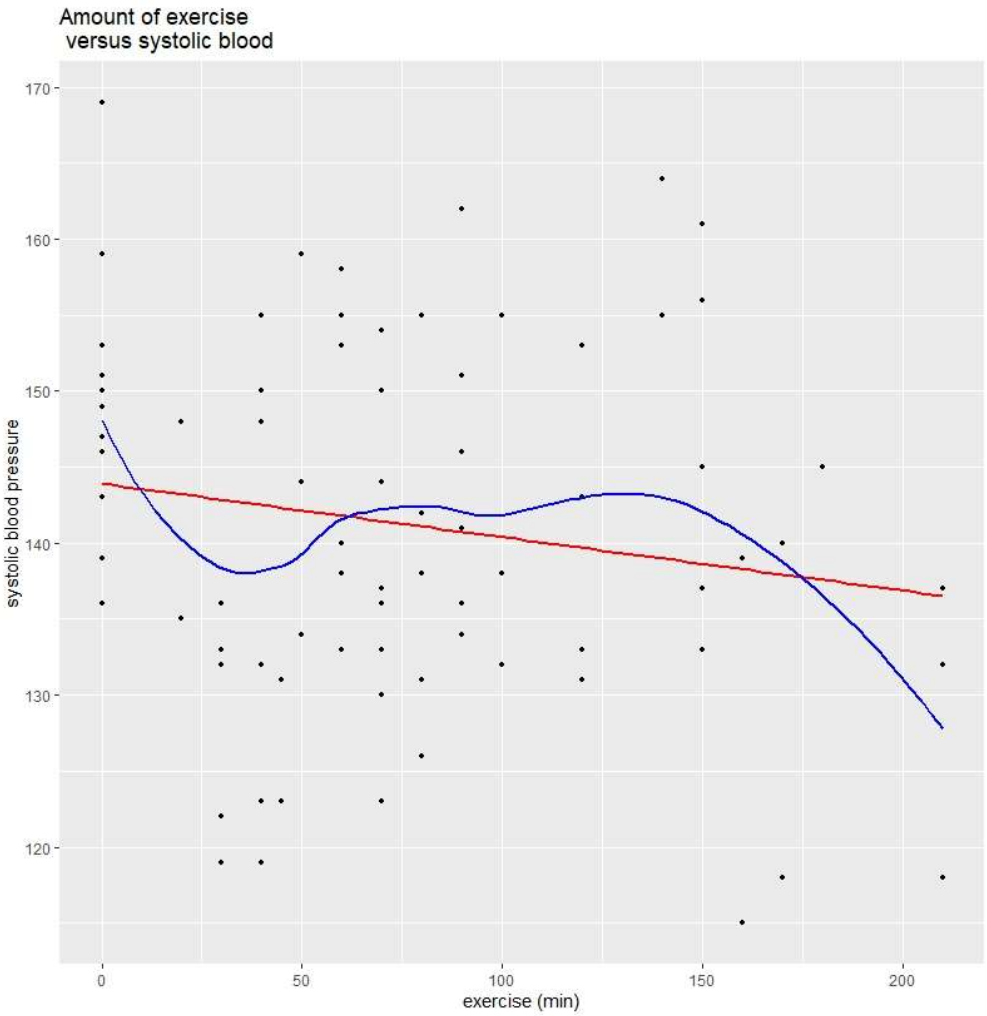
From R, we draw the follow result.

```
Two Sample t-test

data:  No.adg$BP and Yes.adg$BP
t = 0.12239, df = 77, p-value = 0.5485
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 5.126154
sample estimates:
mean of x mean of y
  141.420   141.069
```

The p-value is 0.5485 >= 0.05, hence we failed to reject the null hypothesis. There is insufficient evidence to conclude that there is an association between the blood pressure of male smokers and non-smokers between the ages of 40 – 80.

The analysis does not stop here, the following is to analyze whether exercising impacts the blood pressure. We used simple linear regression analysis for the quantitative variable. Before the analysis, let us look at the distribution of the data to acquire better sense of the objectives.

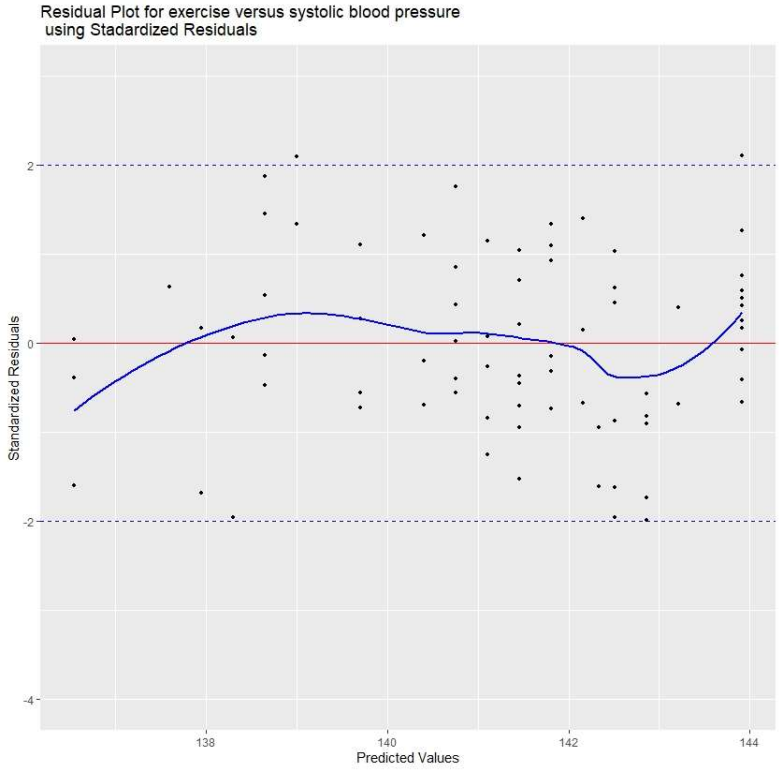


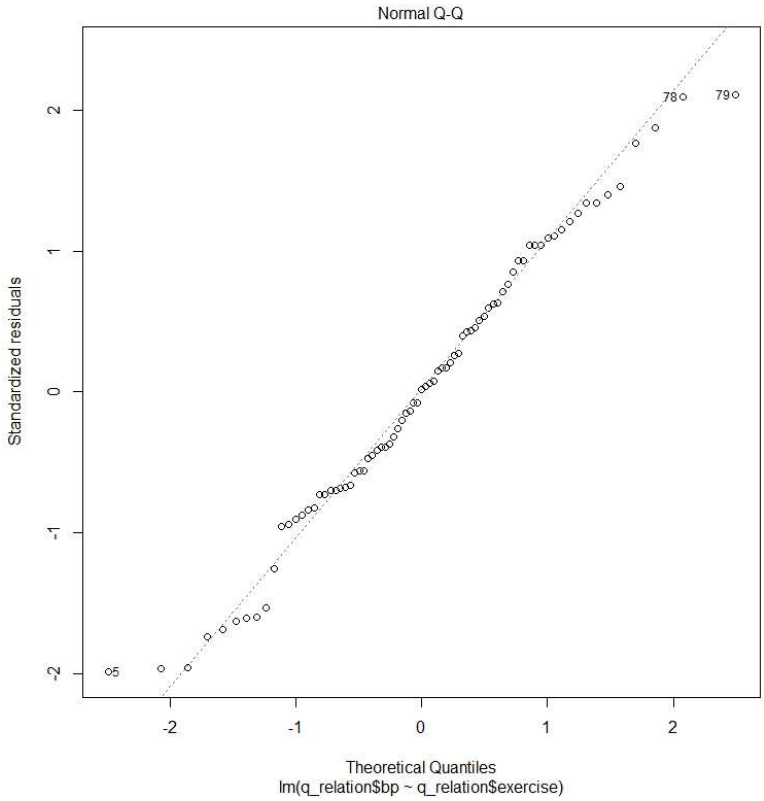
We created a scatter plot with a regression line as well as a lowess curve. Judging by the distribution, linearity is adequately met, constant variance is adequately met. And there exits a certain degree of curvature.

We are fitting the following simple linear regression model for the analysis.

$$y = \beta_0 + \beta_1 X + \varepsilon$$

From this model, we examine further to see if there is any transformation of the model needed.


Residual Plot for exercise versus systolic blood pressure
using Stadardized Residuals
Standardized Residuals
Predicted Values
2
0
-2
-4
138
140
142
144


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(q_relation$bp ~ q_relation$exercise)
78
79
5
2
1
0
-1
-2

From the previous models and plots we can see that the linearity is adequately met as the points are evenly scattered around the residual line. The shape of the overall distribution on the residual plot is constant, hence the constant variance. From the normal QQ plot, most of the points are located on the diagonal line, with one extreme outlier (64). The shape on both ends may indicate there exists longer tails. But we can conclude the normality condition is met.

Conclusion

T t-statistic for the slope is -1.431 and the p-value is 0.1565. The 95% confidence interval is [-0.08, 0.014]. Since the p-value is greater than the alpha-level 0.05, we fail to reject the null hypothesis. Therefore, we cannot conclude that there is an association between the systolic blood pressure and the amount of weekly exercise in minutes complete by males between the ages of 40 – 80. Although we have produced a 95% confidence interval, we cannot be certain that the true value will fall within it.

Reference

Benefits of exercise. (2021, April 22). Retrieved April 26, 2021, from https://medlineplus.gov/benefitsofexercise.html.

Health effects. (2020, April 28). Retrieved April 26, 2021, from https://www.cdc.gov/tobacco/basic_information/health_effects/index.html.

The islands. (n.d.). Retrieved April 26, 2021, from https://islands.smp.uq.edu.au/login.php?req=%2F.