

STA 202 Mathematical Statistics

**Analyzing the Relationship Between Socio-Economic Factors and GDP Per Capita Across Indian States: A Statistical Analysis (2017-2019)**

Date of Submission: 16-04-24

Student Details (Group 2)

|  |  |
| --- | --- |
| **Roll No.** | **Name of the Student** |
| AU2140213 | Vanaja Agarwal |
| AU2140039 | Nish Parikh |

2023-2024 (Winter Semester)

**Problem Statement**:

*Objective*: Investigate the impact of various socio-economic factors on GDP per capita across 30 states in India from 2017 to 2019.

*Importance*: Understanding these relationships can inform targeted policy interventions to foster economic growth, reduce unemployment, and alleviate poverty.

*Research Questions***:**

1. Relationship between natural population growth rate and GDP per capita.

2. Impact of public expenditure on health on GDP per capita.

3. Correlation between relief expenditure on natural calamities and GDP per capita.

4. Influence of total food grains production on GDP per capita.

**Plan:**

The data for this study has been sourced from the Reserve Bank of India's (RBI) official website. The dataset includes information on GDP per capita and predictor variables for multiple states over a specified time period (2017-2019). Statistical models, such as multiple regression analysis, will quantify the relationships between the predictor variables and GDP per capita. The dataset comprises observations from 30 states in India, excluding Union territories, over three consecutive years (2017, 2018, and 2019). With 90 observations (30 states × 3 years), this study utilizes a sample size of 90 to investigate the relationships between the predictor variables and GDP per capita.

**Data:**

1. **Source:**

- Data collected from the Reserve Bank of India's official website, who collects this data through various sources such as government data sources, surveys and other similar sources.

- Variables include,

*Predictor variables:* natural population growth rate, public expenditure on health, relief expenditure on natural calamities, total food grains production, state-wise invested capital, unemployment rates, *Response variable:* GDP per capita.

1. **Preprocessing:**

- Removed Union territories due to missing data and inconsistencies.

- Imputed missing values using the average imputation method.

- Excluded variables with insufficient data (Average Inflation, Life Expectancy).

- Tested variables for normality using the Shapiro-Wilk Test.

- At first, we had considered including data for 2020-2021 also for training our model, but we decided not to move forward due to missing data.

1. **Organization:**

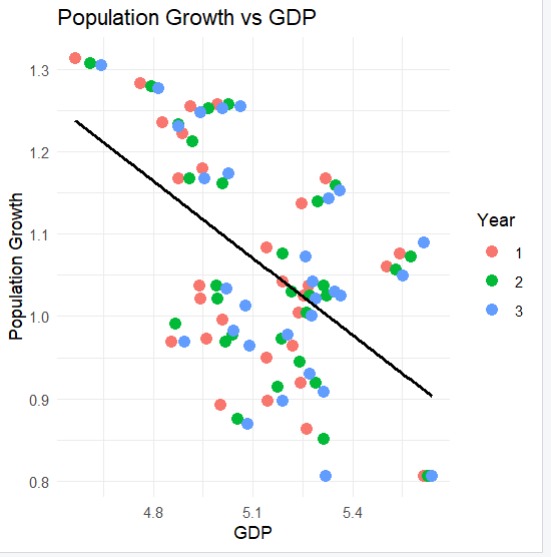
- Consolidated data into a single table with GDP per capita as the response variable and others as predictors (excluding State identifier).

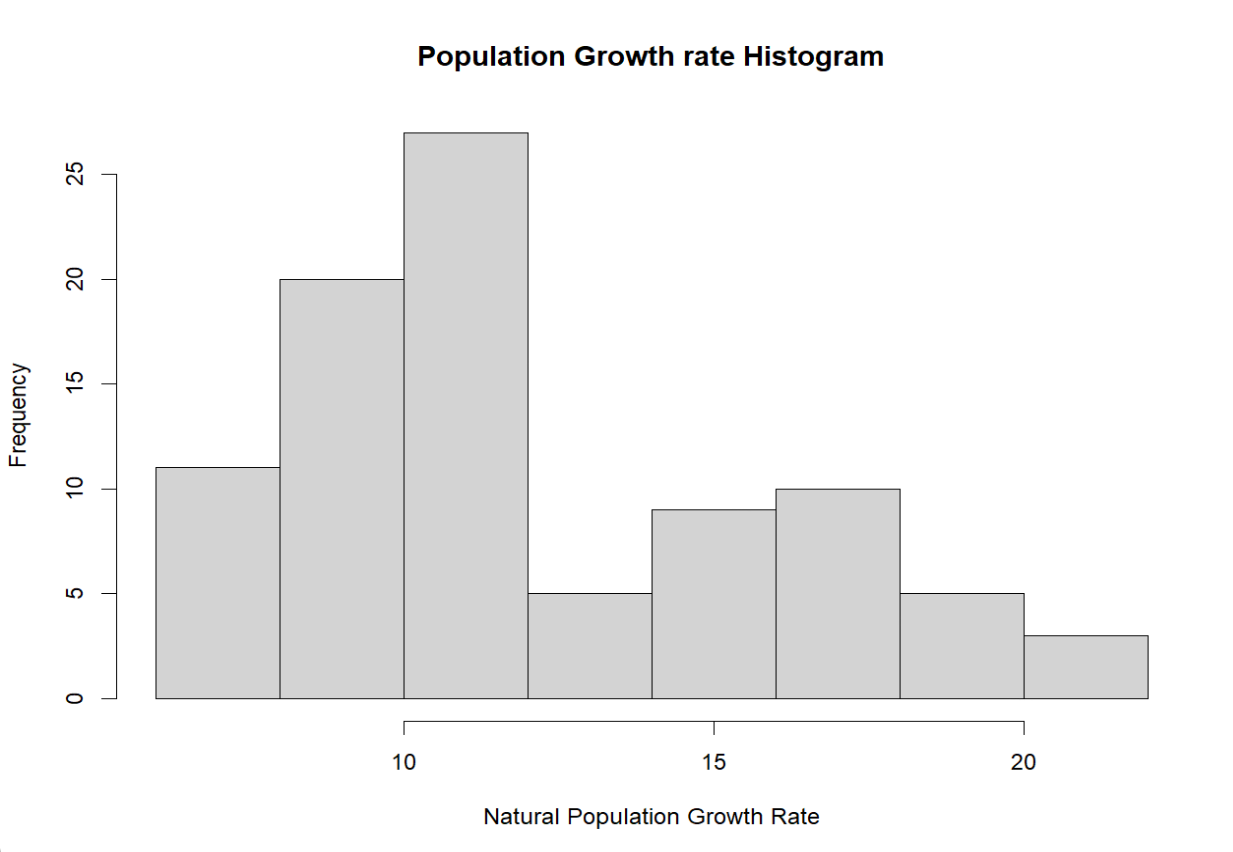
- Log-transformed data for normalization (method of standardization).

- Imported into R-Studio IDE for further analysis.

**Analysis:**

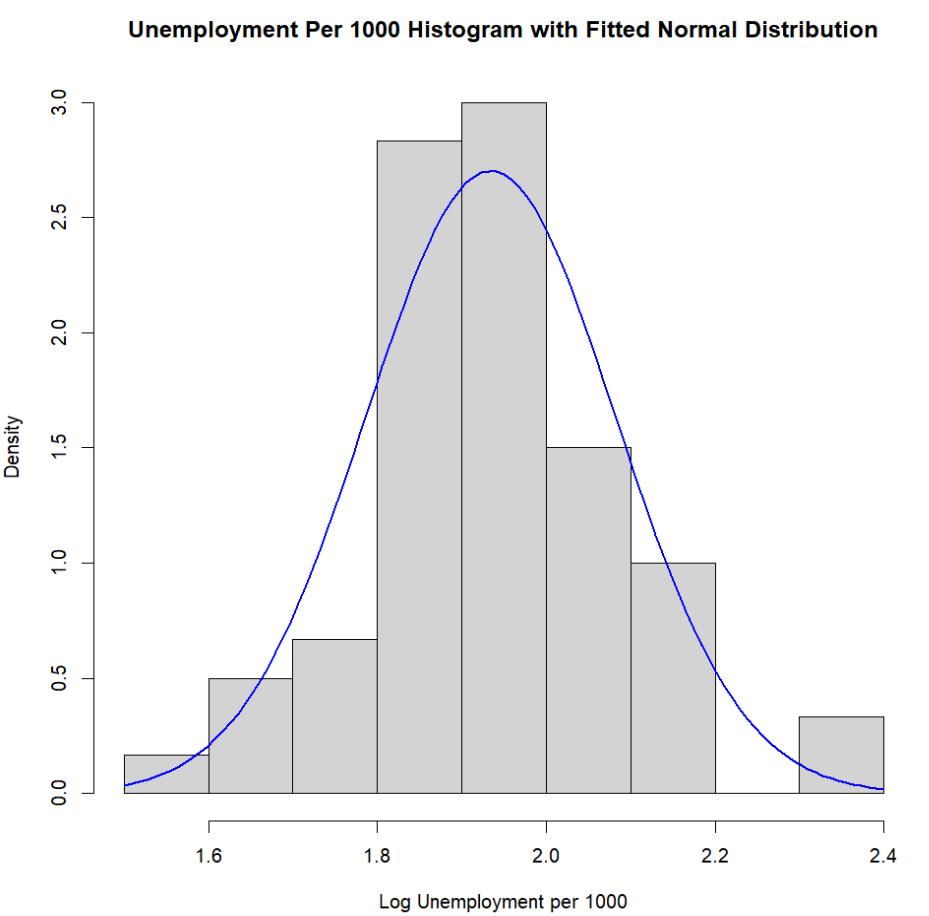
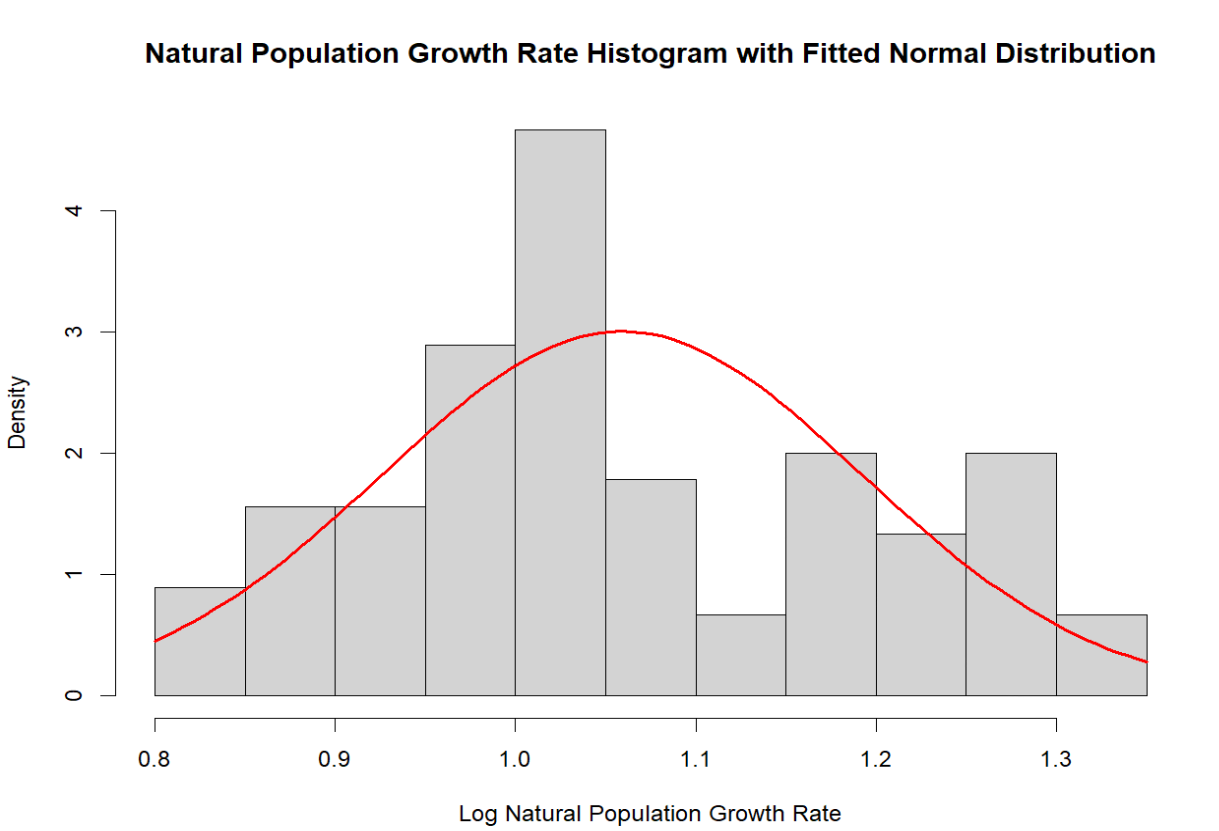
1. **Descriptive statistics**

This is a scatter plot which depicts the relationship between the predictor variable natural population growth rate percent and GDP per capita after log transformation of 90 observations for the years 2017,2018 and 2019. As the majority of values lie near the trendline, we can infer that this variable has a good positive correlation with the response variable (GDP per capita).

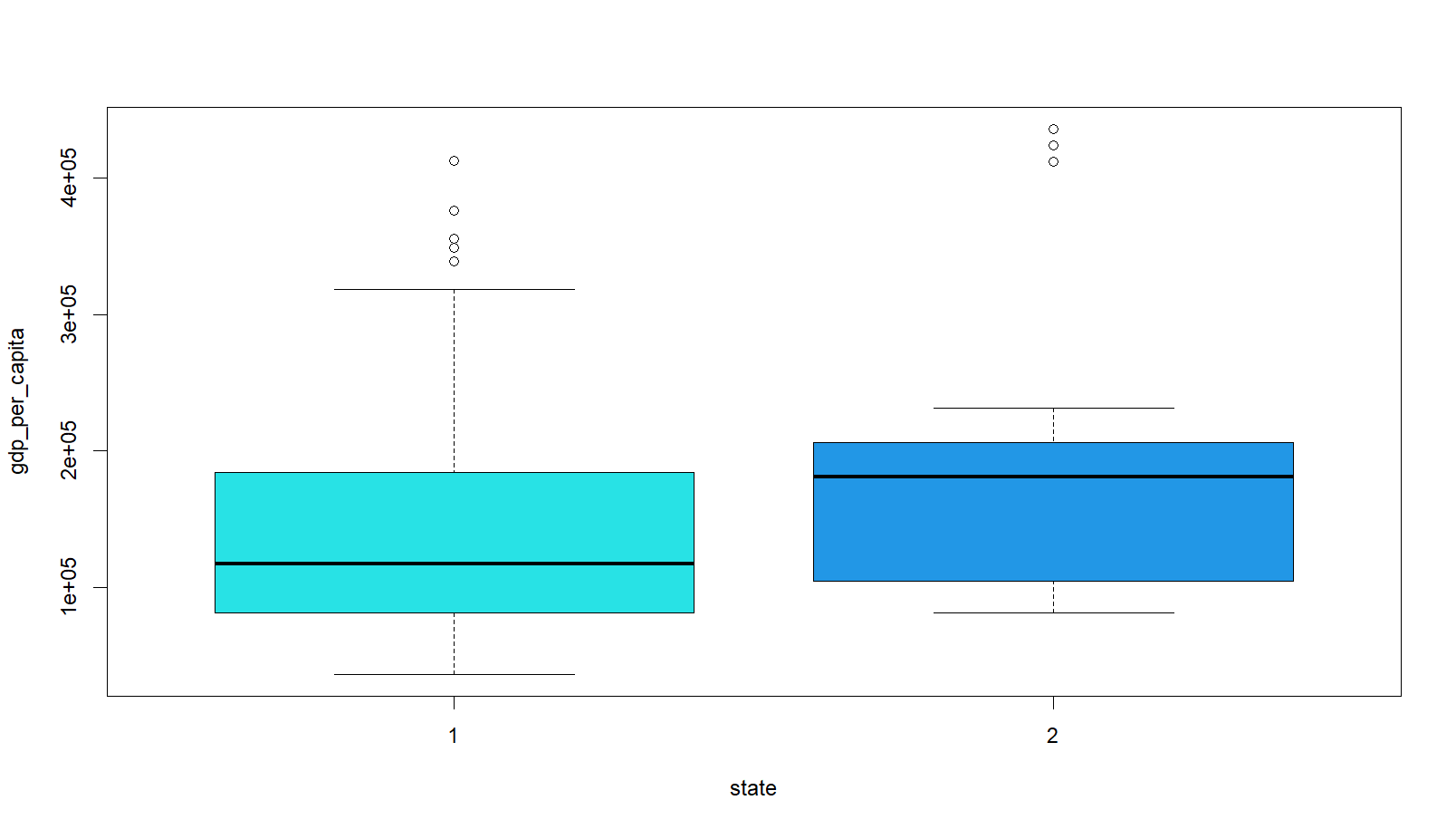
c

This is a histogram of the predictor variable natural population growth rate before the log transformation. From the graph, we can observe that it doesn’t follow a normal distribution.

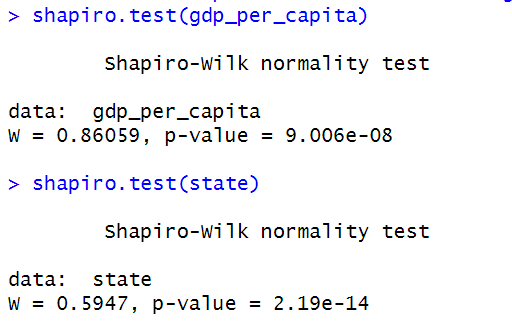
1. **Statistical Inference**
2. **Fitting of the Normal Distribution**

** **

1. **Hypothesis Testing**



boxplot(gdp\_per\_capita~state) depicts that both the variables (state and the gdp\_per\_capita) don’t follow a normal distribution, also inferred by the p-values of the shapiro-test in R as they are less than 0.05.



Data: A sample of 30 states for 3 years (90 observations), GDP\_per\_capita and state (Northern or Southern state)

1. ***Parametric two sample independent t test***

Null Hypothesis: There is no difference between the average gdp per capita of the northern and southern states.

Alternative Hypothesis: There is a difference between the average gdp per capita of the northern and southern states.

**t = -1.9313, df = 88, p-value = 0.05666**

Interpretation: With a p-value of 0.05666, there is not strong evidence to reject the null hypothesis.

Inference: There may not be a statistically significant difference in GDP per capita between the two groups of states.

1. **Non-Parametric Test**
2. **Wilcoxon rank sum test with continuity correction**

Null Hypothesis: The median difference in GDP per capita between the northern and southern states is equal to 0.

Alternative Hypothesis: The median difference in GDP per capita between the northern and southern states is not equal to 0.

**W = 613, p-value = 0.0142**

Interpretation: With a p-value of 0.0142, there is strong evidence to reject the null hypothesis.

Inference: There is a statistically significant difference in median GDP per capita between the two groups of states.

1. **Pearson's Chi-squared test with Yates' continuity correction**

Null Hypothesis: There is no association between GDP per capita and the type of state(northern or southern).

Alternative Hypothesis: There is an association between GDP per capita and type of state(northern or southern).

**X-squared = 5.4056, df = 1, p-value = 0.02007**

Interpretation: With a p-value of 0.02007, there is strong evidence to reject the null hypothesis.

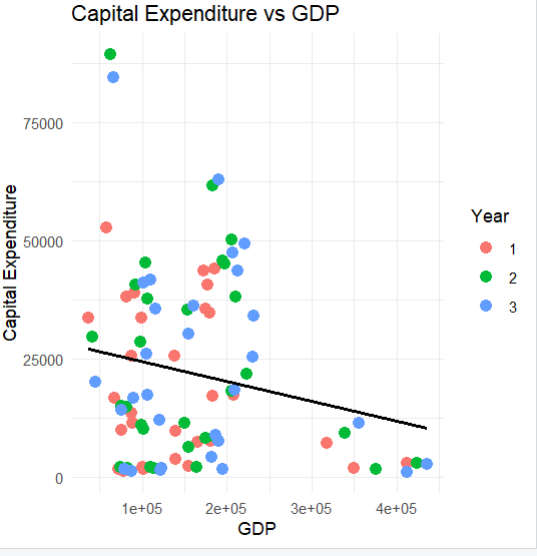
Inference: There is a statistically significant association between GDP per capita and states.

**From the above three tests, we can conclude that our data was not following a normal distribution, the parametric tests gave us inaccurate results drifting towards the null hypothesis whereas the non-parametric tests drifted towards accepting the alternative hypothesis which is more applicable in our case.**

1. **Prediction:**

Correlation Matrix:

The correlation between capital expenditure and health expenditure was found to be high. All others were comparatively low correlation.



Scatter Plot:

This is the scatter plot for one of the predictor variables plotted with the response variable. This scatter plot was created before the log transformation and as we can clearly see a simple linear regression line does not fit the model very well.

**Linear Regression:**

In our dataset, we binary-coded the State into Northern and Southern regions, and we created three different models (All States, Northern States, and Southern States) to observe and test the significance of each predictor variable in predicting the response variable and the changes in the overall error metrics of the model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Log**  **Transform** | **All States** | **Southern States** | **Northern States** |
| No | Adj. R2 = 0.2373  {IC, FG, NPG} | Adj. R2 = 0.1795  {FG} | Adj. R2 = 0.1485  {UR} |
| Yes | Adj. R2 = 0.6222  {IC, FG, NPG} | Adj. R2 = 0.7915  {HE,CE,UR,IC,NPG} | Adj. R2 = 0.6015  {NPG,FG,IC} |

We noticed a class imbalance (30:60) between southern and northern states. This explains the well fit model on the southern states as it has fewer data points.

**Legend:**

IC: Invested Capital HE: Health Expenditure

FG: Production of Food Grains CE: Capital Expenditure

NPG: Natural Population Growth UR: Unemployment Rate

{}: denotes significant variables

**Prediction:**

We created another dataset for predicting the response variable which included data from all the predictor variables for the year 2017 on which our model had been trained and also for the future year 2020. When collecting the data of predictor variables for the year 2020, some of the data was unavailable for a couple of variables, so we applied a simple linear regression formula on the existing dataset (2017,2018,2019) to predict the missing values upon which we further tested our models.

**Uncertainty Quantification:**

This component helps in assessing the reliability of our predictions. We utilized different multiple regression models to understand the range within which the actual GDP per capita values for 2020 are likely to fall.

*Model 1 - All States:*



*Model 2- South States*



*Model 3- North States*

****

**Discussion:**We decided to use the log transformation to standardize the data as the data was in very different units with tremendous values for different variables as justified in the statistical inference part. Then, we studied the correlation between the predictors and with response variable. This helped us understand why certain variables were better at predicting the response variable (for a particular type of state) more accurately than other variables for different kinds of models namely *All States*, *North States* and *South States*.

**Conclusion:**

Overall, our analysis revealed valuable insights into the relationships between socio-economic factors and GDP per capita in Indian states from 2017 to 2019. While parametric tests didn’t give clear results due to data not following a normal distribution, non-parametric tests provided significant findings. Our predictive models demonstrated promising performance, especially when considering regional variations between northern and southern states. In the future, we can expand the dataset to include more recent years. Adding more socioeconomic variables could also make the analysis more robust and useful. Trying out advanced statistical and machine learning techniques may provide even deeper insights into how socioeconomic factors relate to GDP per capita.

Understanding the nuanced impact of socio-economic factors on GDP per capita can inform targeted policy interventions aimed at fostering economic growth, reducing unemployment, and mitigating poverty at the state level in India.

**References:**

*Reserve Bank of India - Handbook of Statistics on Indian States*. (n.d.). <https://rbi.org.in/Scripts/AnnualPublications.aspx?head=Handbook+of+Statistics+on+Indian+States>

Mike Saint-Antoine. (2023, September 4). *Fitting probability distributions to data in r* [Video]. YouTube. <https://www.youtube.com/watch?v=x-ApQenI5ko>

Research Methodology Advanced Tools. (2021c, November 26). *Wilcoxon Signed Rank Test Using R Studio(nonparametric test)(Wilcoxon test)(rstudio)(signed test)* [Video]. YouTube. <https://www.youtube.com/watch?v=evZZT7SIQvM>

Research Methodology Advanced Tools. (2021b, October 10). *Independent t test using r studio(rstudio)(independent two samplet test)* [Video]. YouTube. <https://www.youtube.com/watch?v=Fvsw2G1Aki8>

Research Methodology Advanced Tools. (2021a, October 5). *Multiple Regression Using R Studio(rstudio)(multiple regression)(r sqaure and adjusted r square)* [Video]. YouTube. <https://www.youtube.com/watch?v=UomnHPBfvBc>

Eugene O’Loughlin. (2021, June 28). *How To. . . Make a Prediction using a Multiple Linear Regression Model in R #102* [Video]. YouTube. <https://www.youtube.com/watch?v=AR6sLpcVcSU>

Equitable Equations. (2020, December 4). *Chi-Squared testing for independence in R* [Video]. YouTube. <https://www.youtube.com/watch?v=XfsDtYuwVzs>