**STATISTICAL ANALYSIS ON FACTORS INFLUENCING LIFE EXPECTANCY**

1. **INTRODUCTION**

Life expectancy is a critical measure of the overall health and well-being of a population. Many studies have been conducted to identify the factors that impact life expectancy, including demographic variables, income composition, and mortality rates. Previous studies have often used multiple linear regression based on data from a single year for all countries, which may not accurately reflect the complex interplay of factors affecting life expectancy over time.

In this study, we aim to address these gaps in the literature by formulating a regression model based on mixed effects modeling and multiple linear regression. We will also consider important immunization factors, such as Hepatitis B, HIV AIDS as well as other health-related, economic, and social factors. By identifying the key factors contributing to lower life expectancy, this research will contribute to efforts to improve the health and well-being of populations around the world.

1. **DATASET**

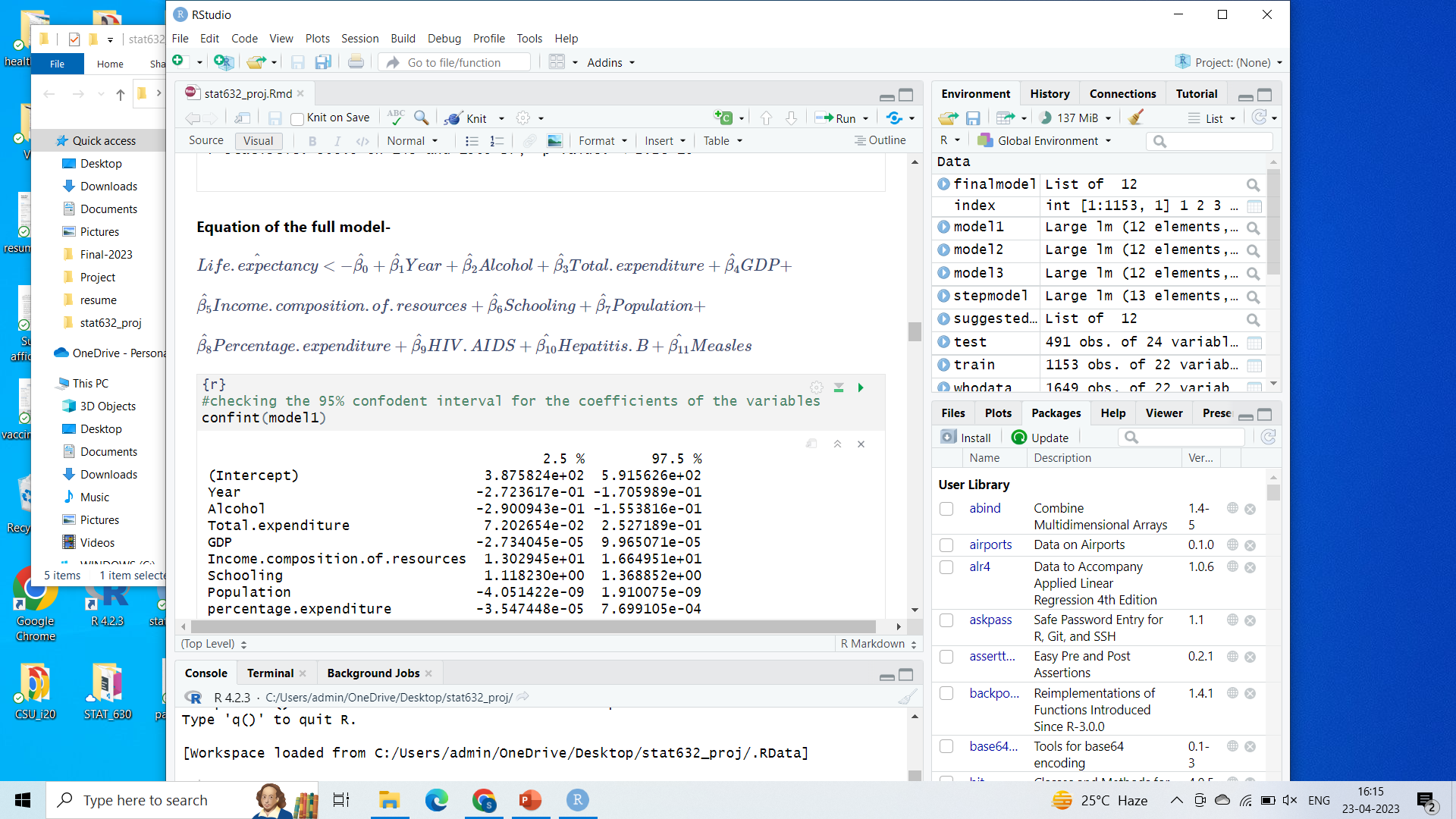
The dataset used in this research project consists of health and economic data from 193 countries, collected from the Global Health Observatory data repository and the United Nations website. It covers the period from 2000 to 2015 and includes 2938 observations and 21 variables (divided into four broad categories: immunization, mortality, economic, and social factors). Out of 21, the following 11 variables are chosen as the predictor variables in determining the life expectancy (response variable)-

* Year-the year data was gathered
* Alcohol- consumption of alcohol, recorded per capita (in liters)
* Total expenditure-government expenditure on health as a percentage of total government expenditure (%)
* GDP-Gross Domestic Product per capita (USD)
* Income composition of resources-Human Development Index (ranging from 0 to 1)
* Schooling-Number of years of Schooling
* Population-Country’s Population
* Percentage Expenditure-Expenditure on health as a percentage of GDP
* HIV AIDS-Deaths per 1 000 live births HIV/AIDS (0-4 years)
* Hepatitis B-immunization coverage among 1-year-olds
* Measles - number of reported cases per 1000 population

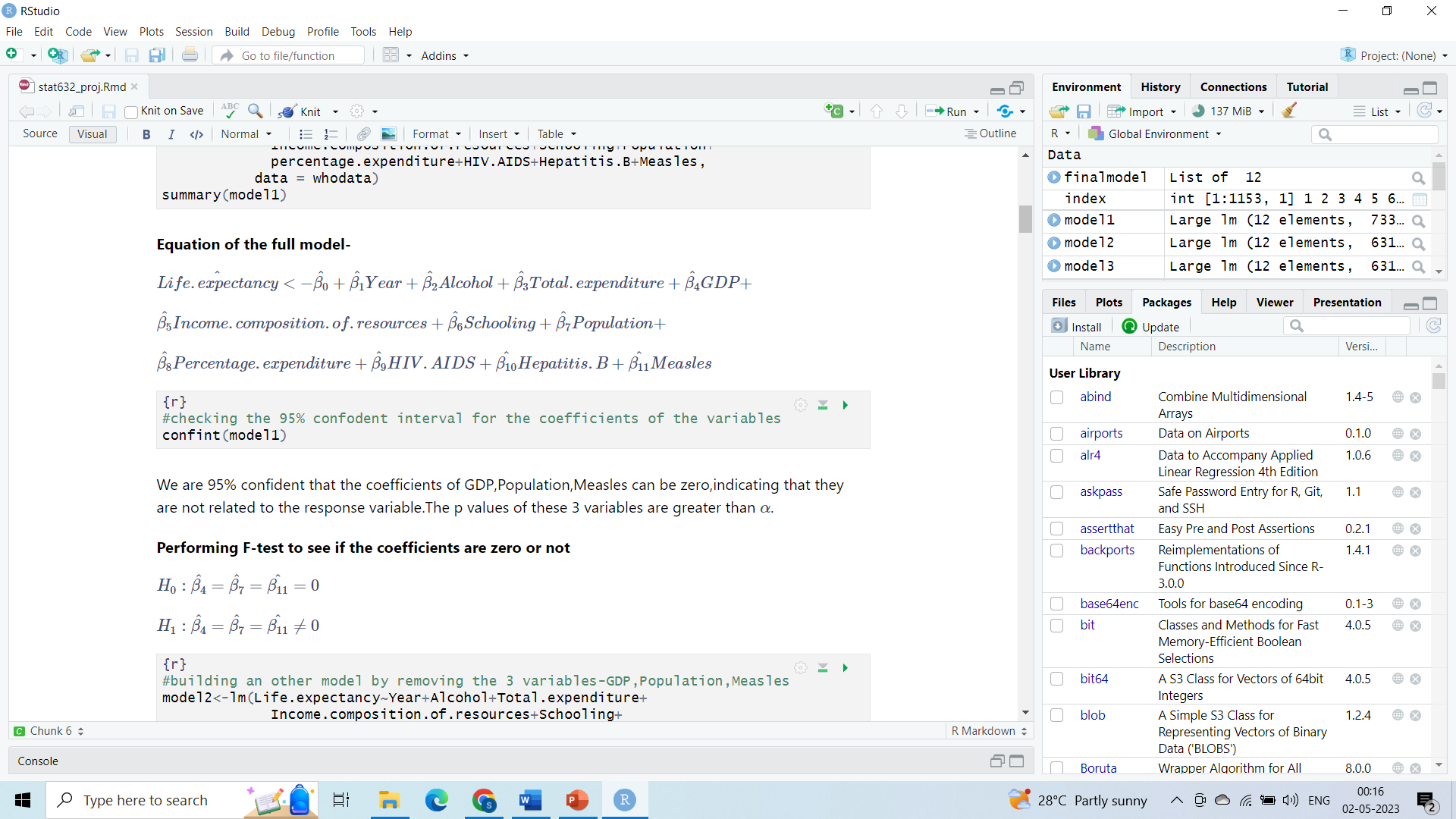
1. **METHODS AND RESULTS**

**3.1 Methods:**

The method we are going to use is Multi Linear Regression (MLR). MLR allows you to estimate how a dependent variable changes as the independent variable changes. MLR model using Life expectancy as a response variable and other variables as predictor variables was built in R language using lm function. The equation of this full model is-



The p-values of GDP, Population and measles are greater than 0.05. The confint () function is used to determine the confidence intervals for the coefficients of these variables. The output of the confint() indicates that there is a 95% confidence level that these three variables' coefficients can be zero.. Partial F-test is also performed to check if the coefficients of these variables are zero or not.



The partial F-test's p-value is 0.05364 with F-value 2.334. As a result, the coefficients for these three variables are zero and we failed to reject the null hypothesis. A new model has been built without the variables GDP, Population, and Measles because they are not significant. The new model does not meet the MLR assumptions. The necessary transformation has been performed out. All the MLR's assumptions are met after transformation.

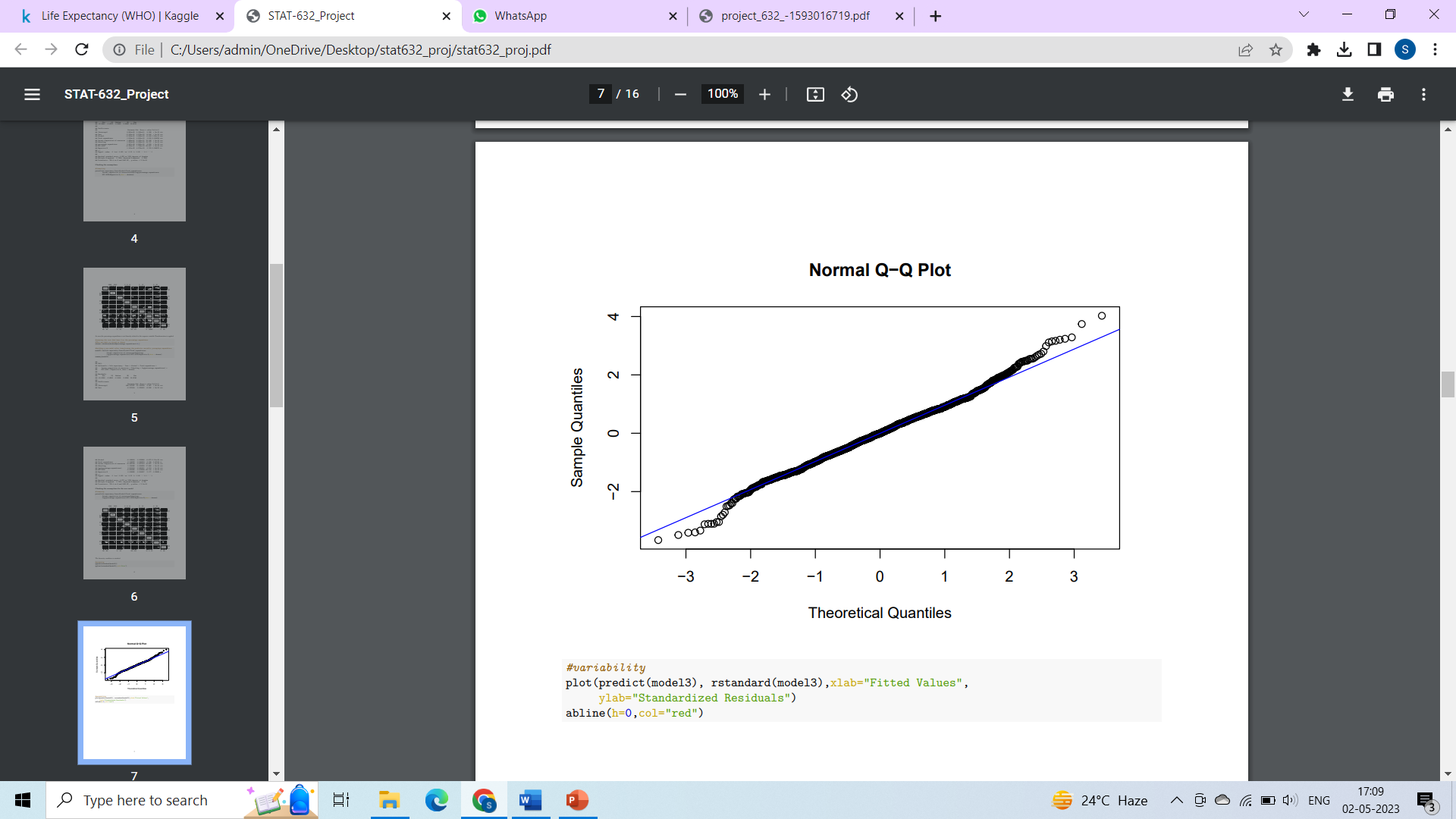
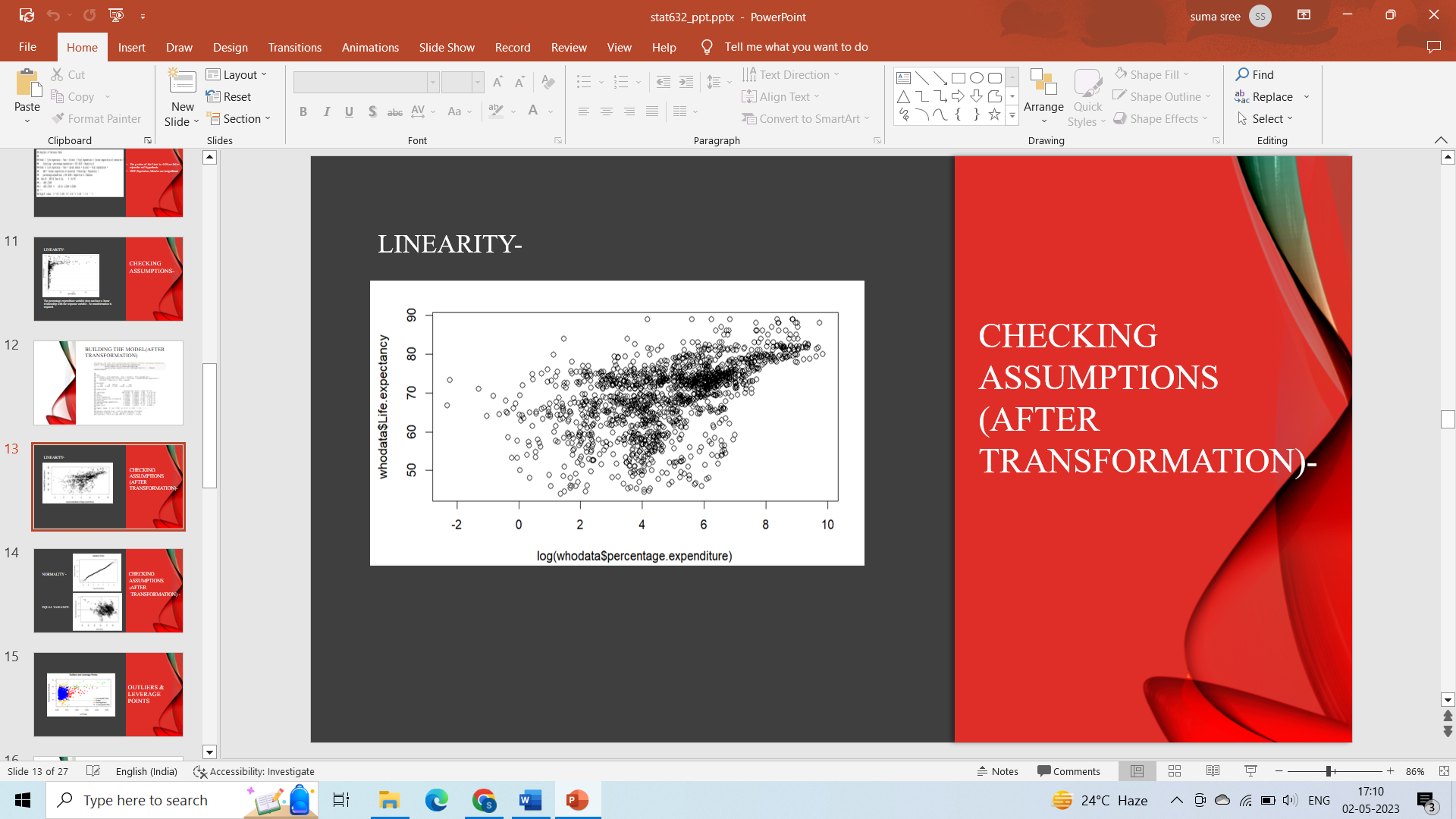
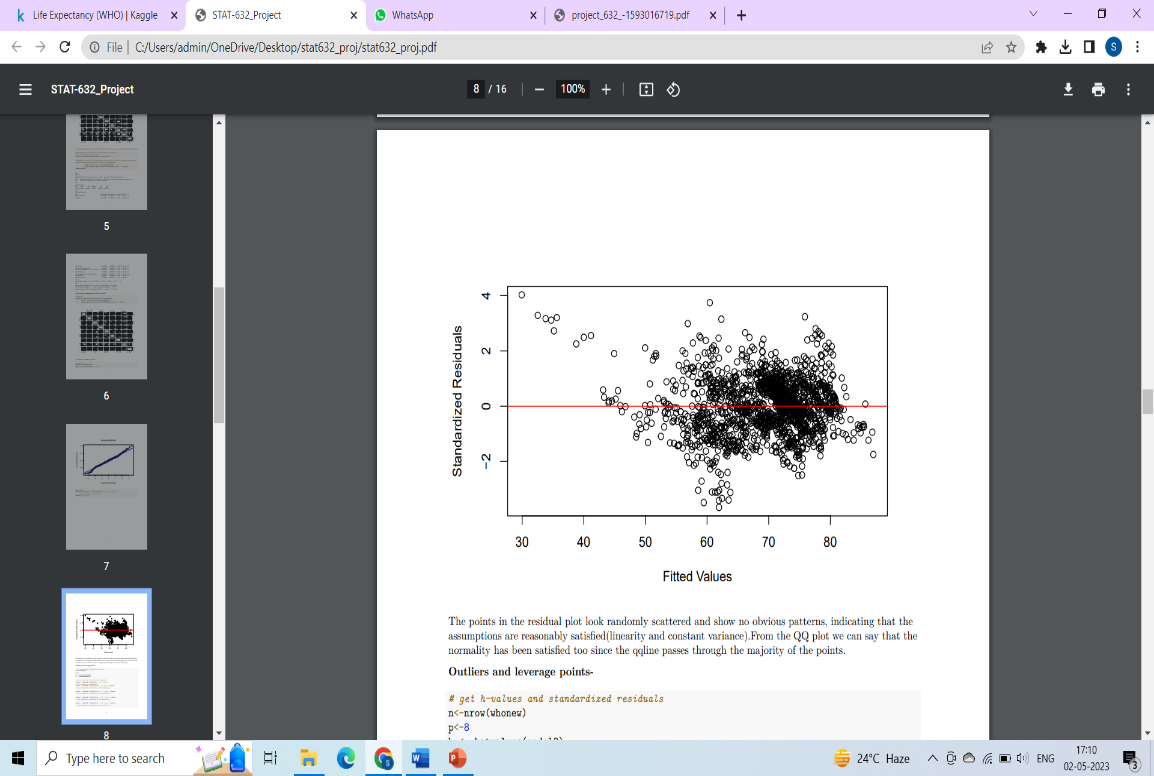


Figure 3.3.1: Checking Assumptions

With the data transformed, another MLR model is constructed. There are 21 observations that are both outliers and leverage points out of 2938 observations. Of 2938, 21 seemed negligible. Therefore, these observations are still included in the model.

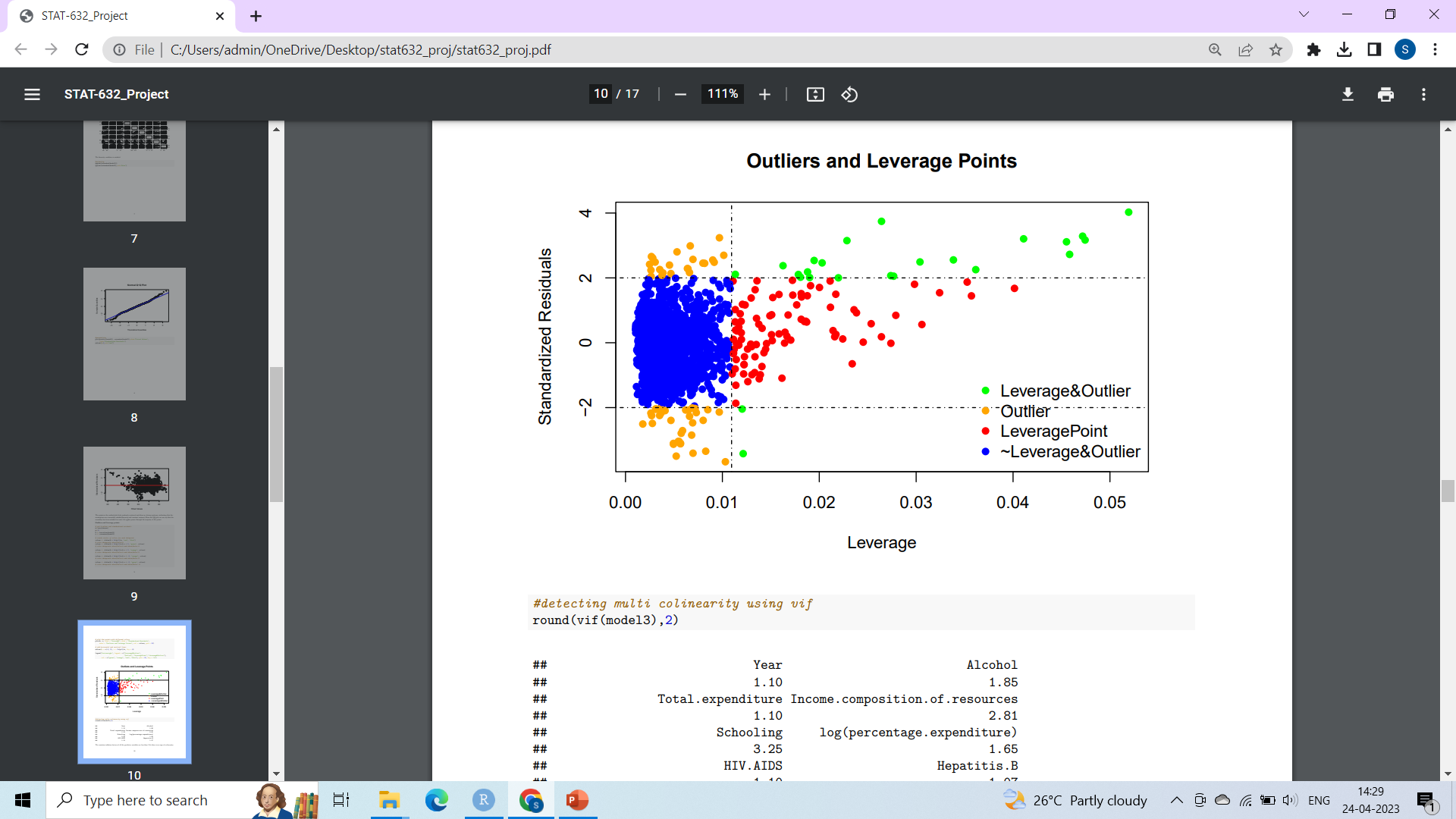
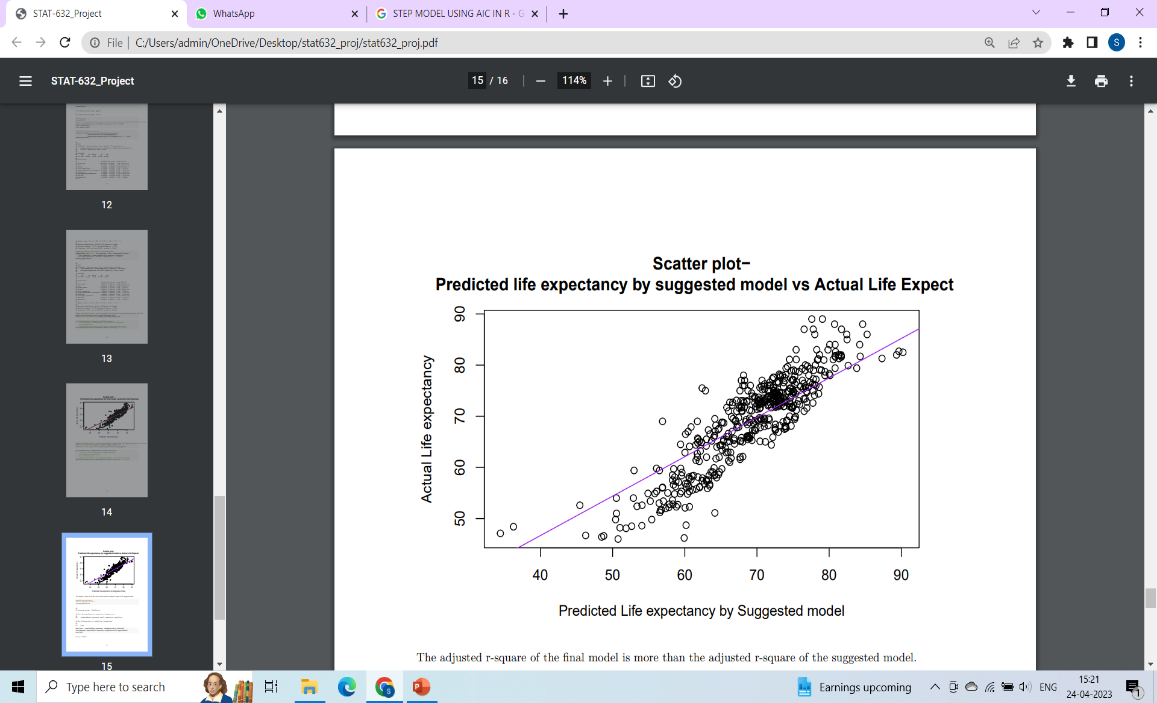


Figure 3.1.2: Outliers and leverage points

Checks are made for multicollinearity in this model. It is crucial to check for any collinearity between the variables since multicollinearity diminishes the statistical power of your regression model by reducing the precision of the calculated coefficients. The vif values of the variables are less than 5 indicating there is no sign of collinearity. To determine which model is the best fit for forecasting life expectancy, AIC is run using the step () method. The model obtained through the step function is identical to the model produced by eliminating insignificant variables. The dataset is divided in a 70:30 ratio as training and testing data respectively. The models—one acquired after transformation (say final model) and one derived via the step () function (say suggested model)—are run on training data. Both models yield predictions when the test data is provided. The predict () method is used to find the predictions.

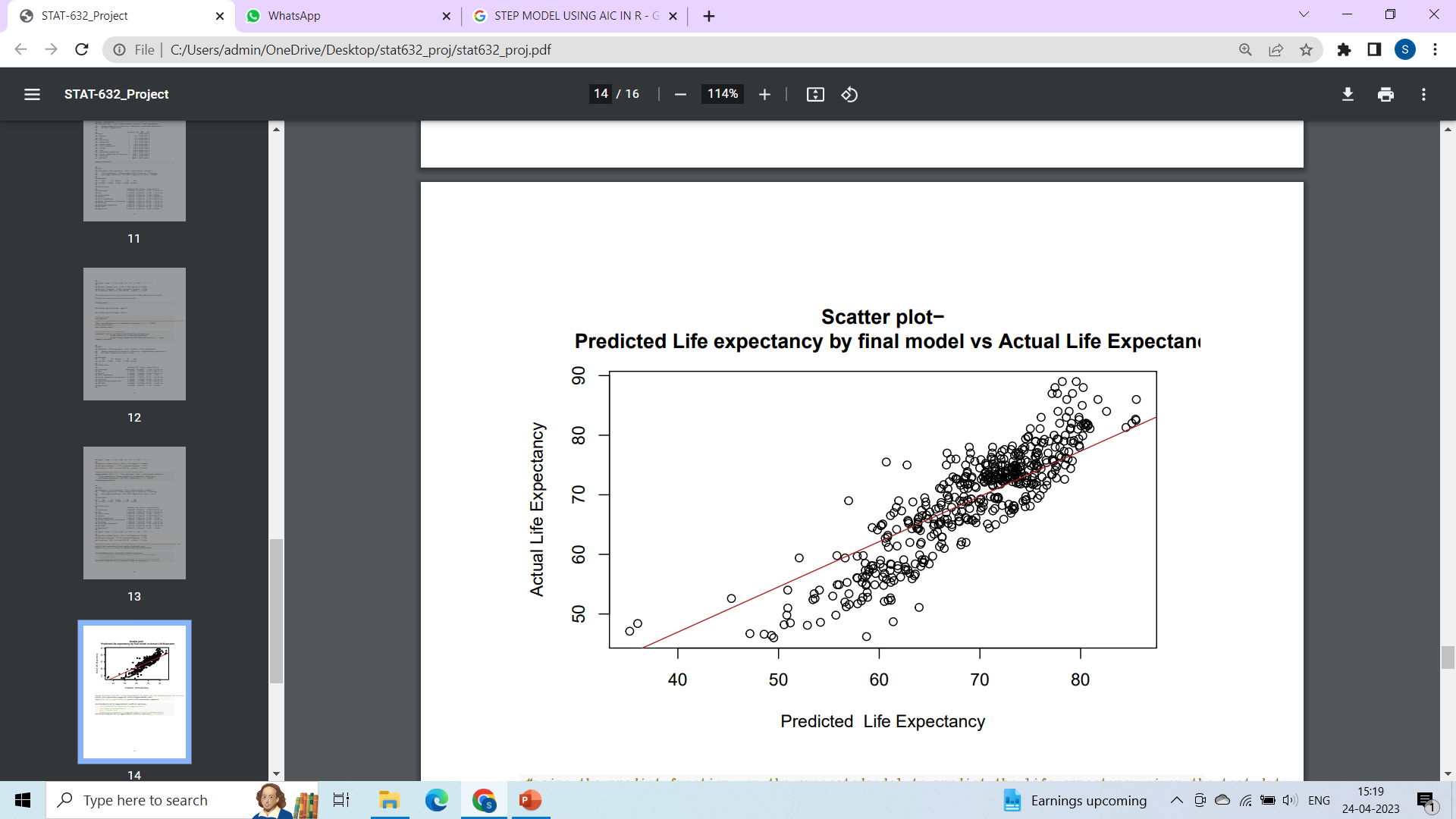


Figure 3.1.3: Predicted life expectancy by final model vs

Figure 3.1.4: Predicted life expectancy by suggested model vs actual life expectancy

actual life expectancy

To determine which model fits the data the best, RMSE is used as an evaluation criterion. The "metrics" package, "ModelMetrics" library, has the "rmse()" function.

**3.2 Results:**

The summary of both these models is given in the table below:

|  |  |  |
| --- | --- | --- |
|  | final model | suggested model |
| p-value | 2.2e-16 | 2.2e-16 |
| Adjusted R-square | 0.7761 | 0.7755 |
| RMSE | 4.649 | 4.0584 |
| Assumptions of MLR | Satisfied | Not satisfied |

Table 3.2.1: Summary of final and suggested models

The final model has a slightly higher adjusted R-squared value (0.7761 compared to 0.7755 for suggested model), indicating that it explains slightly more of the variation in the response variable. Additionally, both models have the same very small p-value, indicating that the models are both statistically significant. The final model has a slightly higher RMSE value (4.0649 compared to 4.058 for suggested model), which means that its predictions are on average slightly less accurate, but the difference is very small.

1. **CONCLUSION AND FUTURE WORK**

**4.1 Conclusion:**

The final model is the best match to reliably forecast life expectancy based on Table 1. This is mostly because all the multi-linear regression assumptions have been met and the model's p-value is below the level of significance. Even though the resulting model's RMSE value is significantly greater than the suggested model's, the difference is extremely minor, meaning that its predictions are generally negotiable and slightly less accurate.

The equation of the model that best fits the data is-

**Life. Expectancy=446.87-0.2006Year-0.202Alcohol+0. 114Total.expenditure+13.46 Income.composition+1.193Schooling+0.626log (percentage. expenditure)-0.638HIVAIDS+0. 010Hepatitis.B**

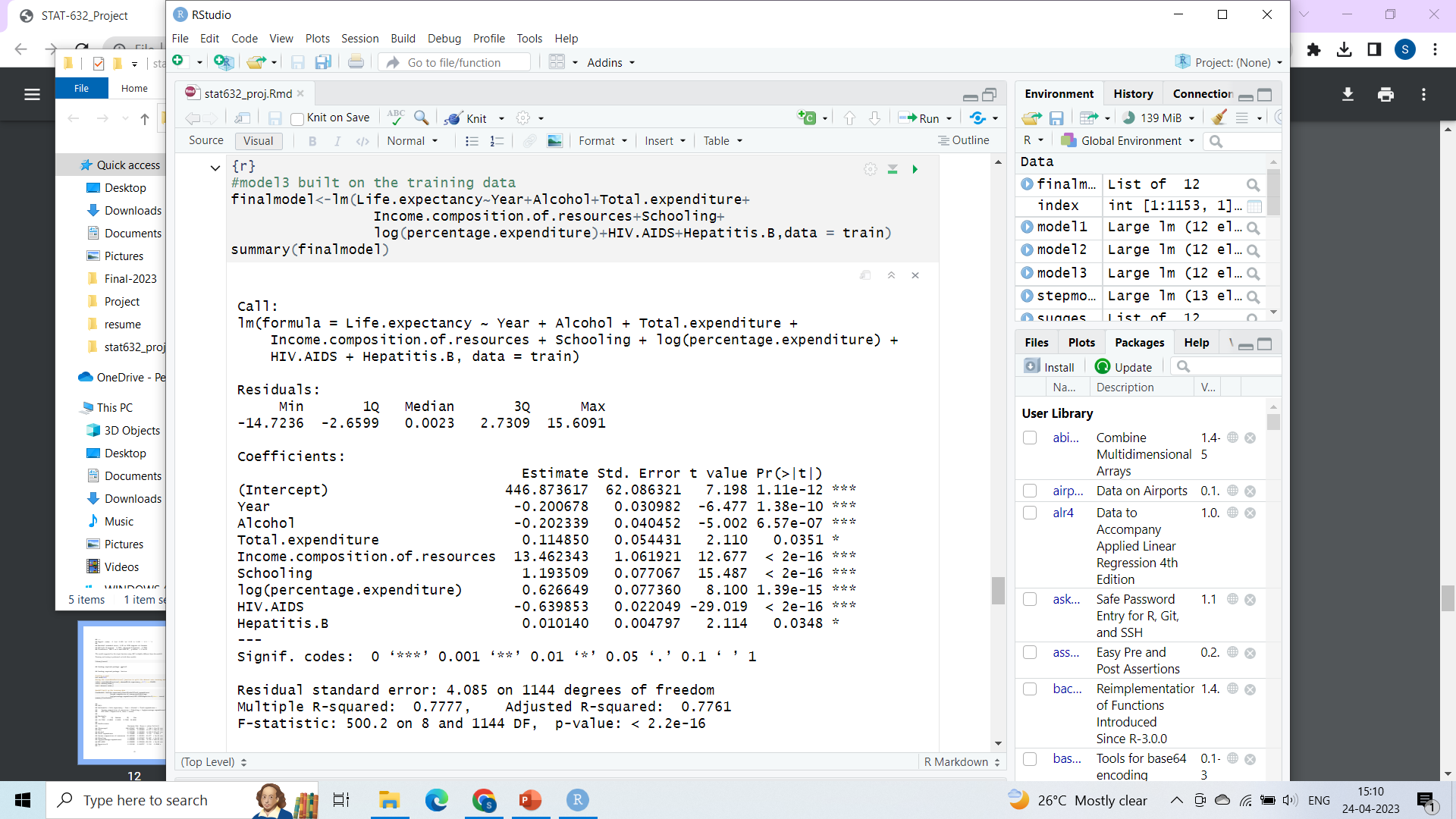


Figure 4.1.1: Summary of final model

**4.2 Limitations and future work:**

* Environmental factors: Environmental factors such as air pollution, water quality, and exposure to toxins can also have a significant impact on health and life expectancy.
* Socioeconomic status: Socioeconomic factors such as income and occupation are known to be strong predictors of health and mortality. However, the WHO dataset does not include measures of these factors.

Future model construction can be made more accurate by taking into account the aforementioned constraints. Predicting the life expectancy of each nation from one time period to the next would be a good idea. By doing so, countries could identify their weaknesses and work to fill them for the good of the nation. It is crucial to research each nation because there are different factors in each one. We were unable to predict the life expectancy country-by-country due to the limited data for each country.

**5. Code appendix:**