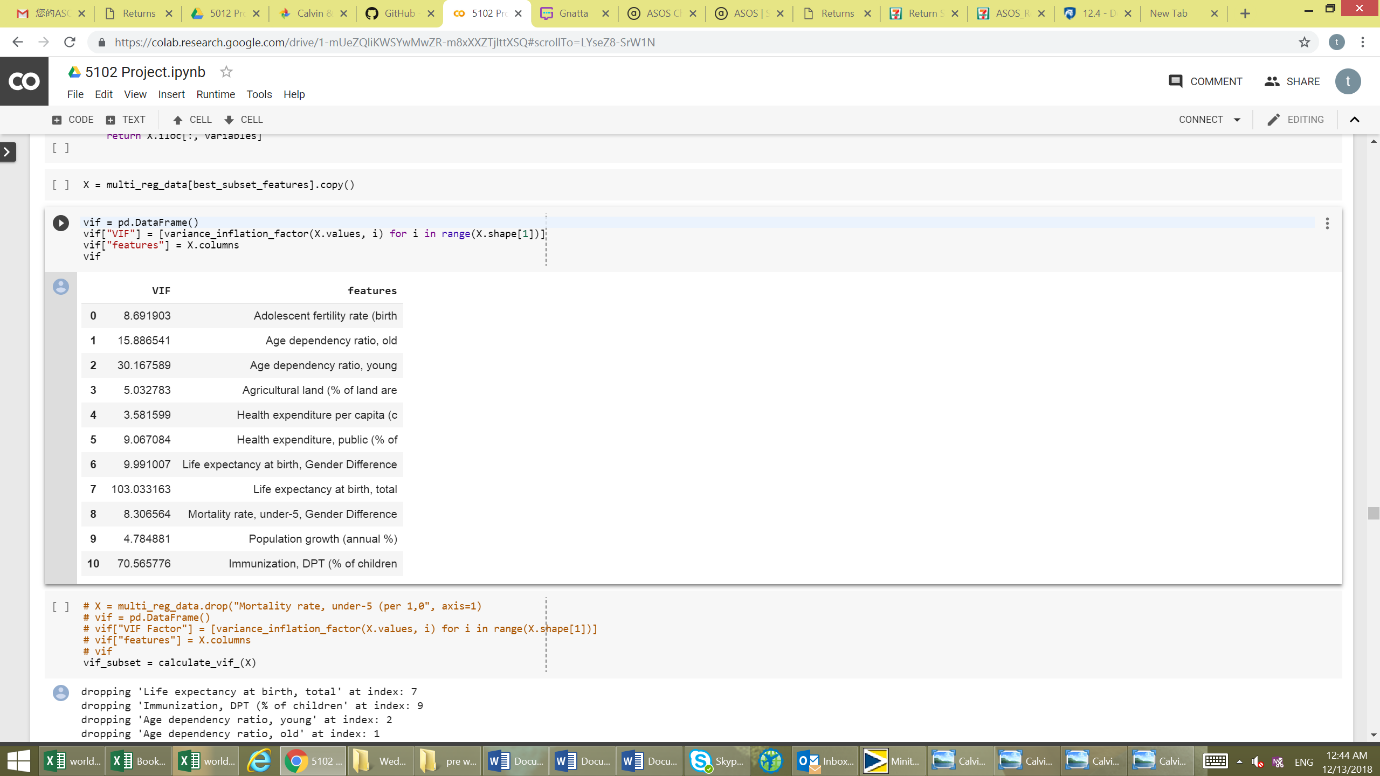
**Multicollinearity and VIF**

After the best selection process, we may check whether Multicollinearity exists among all these selected predictors. Multicollinearity exists when two or more of the predictors are correlated with each other. Multicollinearity can result in several problems, for example, the estimated regression coefficient of one variable depends on which other predictors are included in the model, and this may affect the precision of the model and yield a different conclusion.

A formal method for detecting multicollinearity involves the calculation of Variance Inflation factors (VIF) for individual paramters. One reason why the marginal t-test on are rejected is that the standard error of the estimates are inflated when multicollinearity exists.

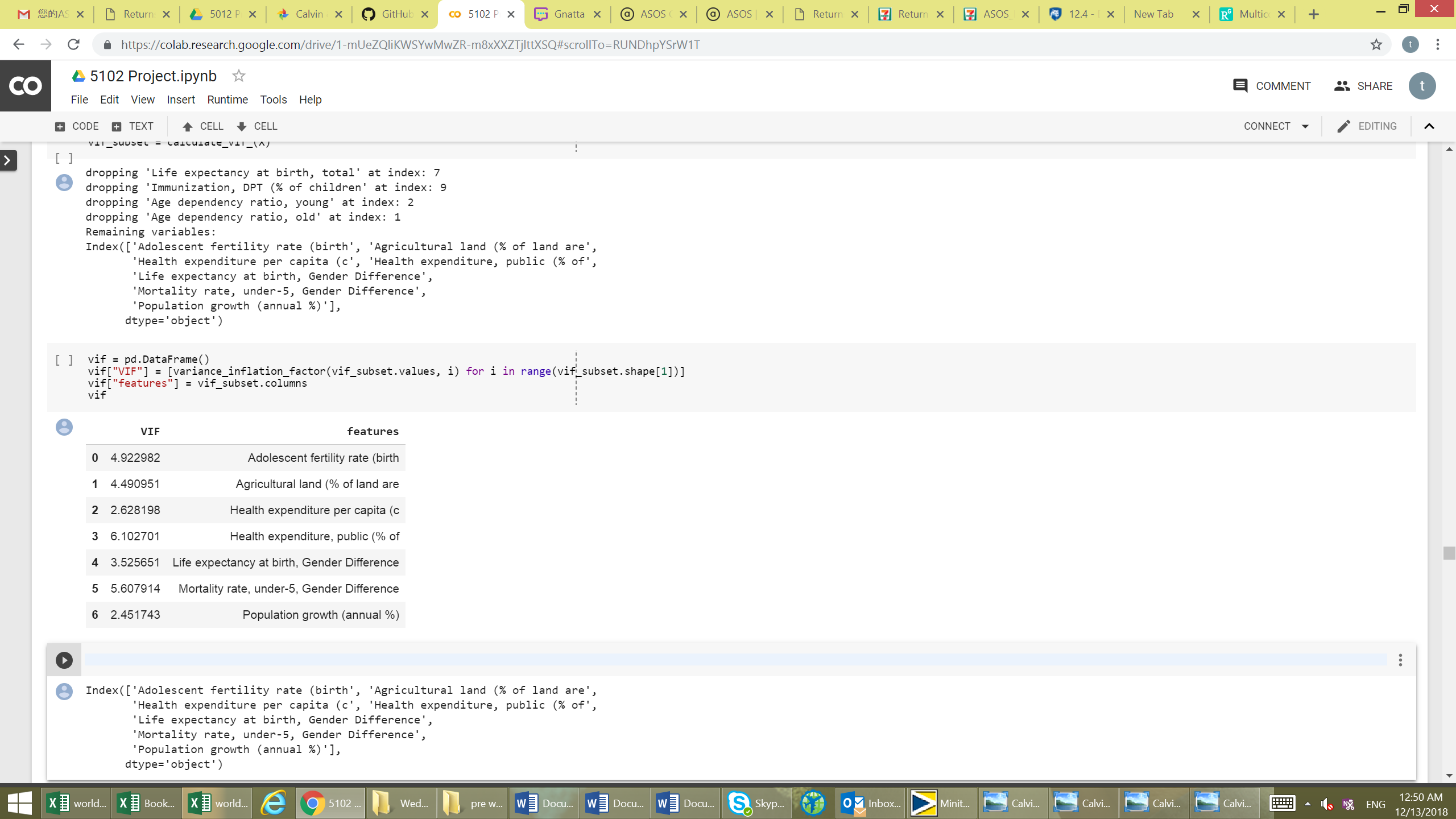
We run the following result

"10" as the maximum level of VIF (Hair et al., 1995)

Therefore, we can conclude there is problems of multicollinearity, and drop the following predictors -

* Age Dependency ratio, old,
* Age dependency ratio, young,
* Life expectancy at Birth,
* total and Immunization, DPT (% of children ages 12-23 months)

And we run the VIF again for the remaining predictors:



From the result, the VIF indicates that the problem of multicollinearity disappears as the index is roughly +/- 5.

**Residual vs Leverage**

Other than VIF, we also check whether there is outliners affect the model. Through Residual vs Leverage Plot:



There are few observations -

* There are serious outliners in Population growth (annual %)
* There is a data point with value greater than 100% in Immunization, DPT (% of children)
* There is polynomial relationship between our response and Life expectancy at birth, Gender Difference

We can consider to

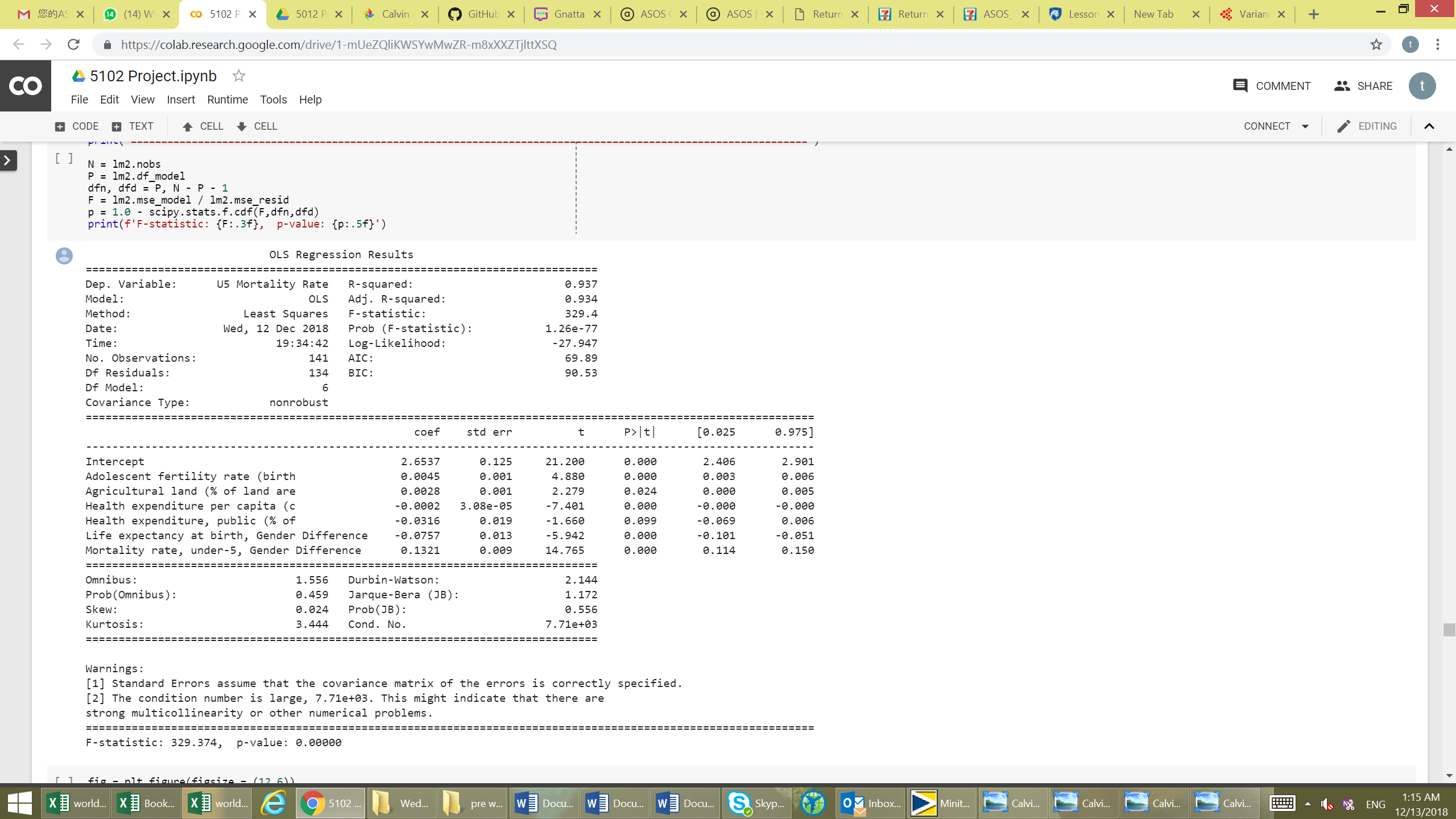
* transform our predictors a little to give a better fit
* Remove outliners
* Remove obviously erroneous data point

**Third Trial – Multiple Linear regression**

After this, we can run a multiple regression with the remaining variables again.

First order model for Mortality rate, under-5 (per 1,000 live births) (Y) as a function of:

* Adolescent fertility rate (births per 1,000 women ages 15-19),
* Agricultural land (% of land area),
* Health expenditure per capita (current US$),
* Health expenditure, public (%),
* Life expectancy at birth, Gender Difference
* Mortality Rate, under-5 Gender Difference



Fitted model:

And we use hypothesis to test whether the model is significant

H0: No linear relationship between Mortality rate and all variables

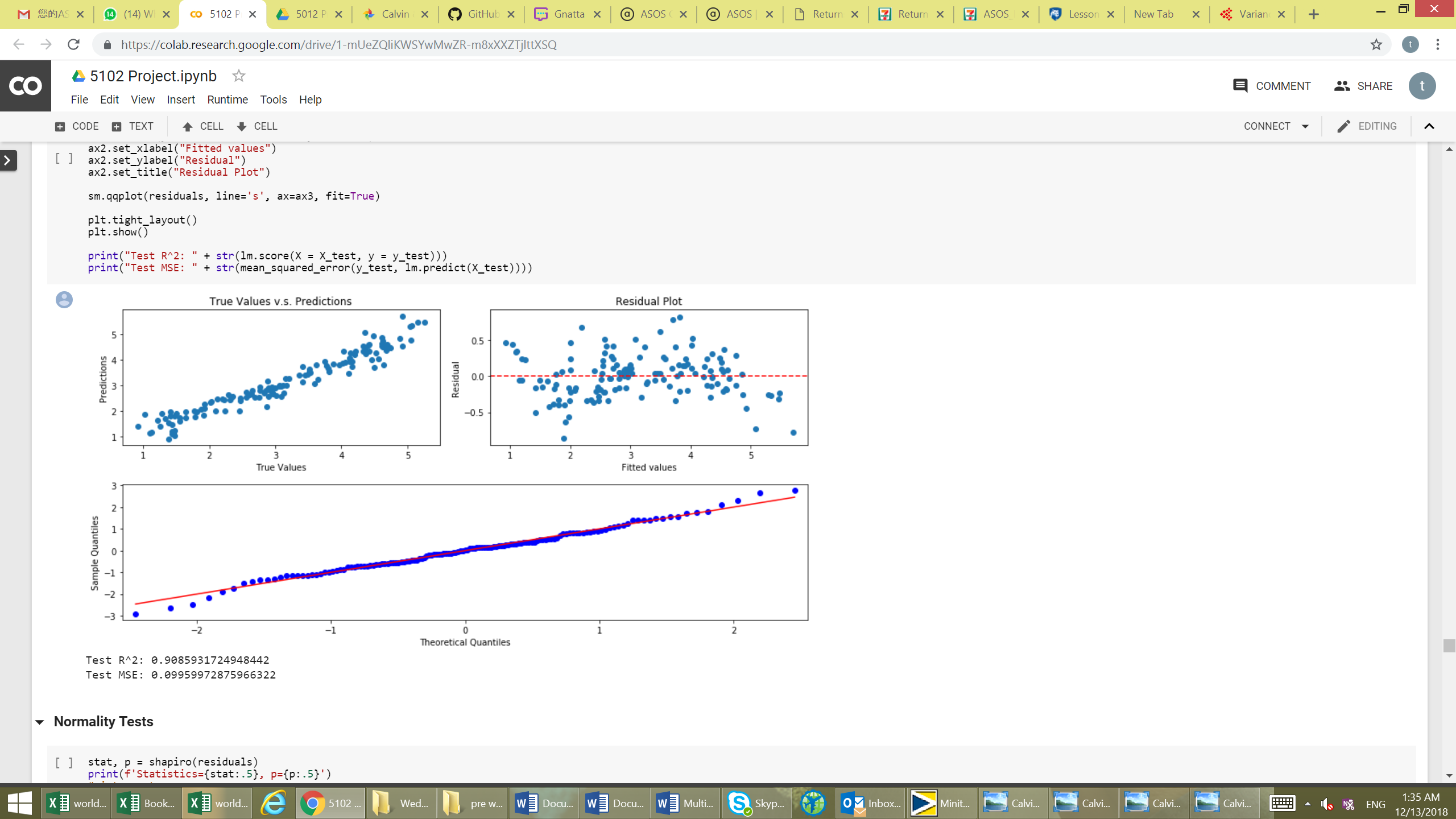
H1: At least one independent variables affects Y

Since the p value is less than 0.0001, we reject H0 and conclude that these independent variables affects Y.

And adjusted R square for this model is 0.934. It indicates that the model explains all the variability of the response data around its mean.

Before making conclusion, we do the model assumption check again.

Here are the graph of Actual vs predict group, residual plot and Q-Q plot below:



From the eye ball test, we can see a good result from the Actual vs predict graph, we have a quite ideal result as the line shows the predicted value is reasonably close to actual value. And then we see the residual plot, we see an improved random pattern in this model, this indicate this is a good fit linear model. From the Q-Q plot, we can observe the residual is normally distributed.

To ensure the residual is normally distributed and has constant variance. We can further go through the normality test and heteroscedasticity test.

We have used few normality test to check for the normality of residual, and we set a hypothesis test H0: Residual follow a normal distribution and H1: residual do not follow a normal distribution

Kolmogorov-Smirnov (K-S) test: Statistics=0.99182, p=0.59095, so we do not reject H0, which we have enough evidence

Shapiro-Wilk test: Statistics=0.28804, p=7.4993e-11, so we reject H0

Anderson-Darling test: Statistic: 0.3324676044750276, p=0.766 ( = 0.05), so we do not reject H0.

Among these test results, we could conclude the residual is relatively normally distributed.

Then we could test for the constancy of error variance through Heteroscedasticity test.

Breusch-Pagan Test: F-Statistics = 1.3572, p = 0.2365, so we do not reject reject H0, so the variance looks constant.

Through the above testing, we can prove the key assumptions of multiple regression are valid in this model.

**Conclusion**