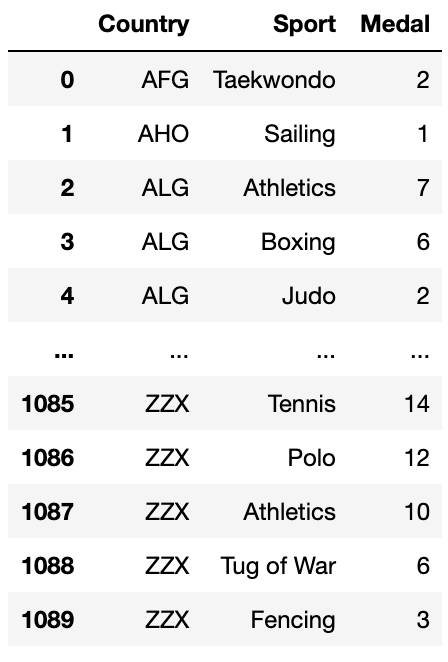
**1.We want to know what sports each country is best at.**

We have processed the number of medals that each country has won in different sports over the years, and obtained the five sports that each country is best at, and the number of medals.



From this result, we can see that for the country numbered ZZX, its teams are good at tennis and polo.

1. **Then we want to know whether GDP and population have an impact on the number of medals**

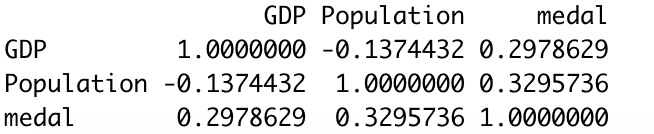
*(1)Data processing and linear correlation analysis between variables*

Because many documents have mentioned that GDP and population will have an impact on the number of medals in this country, so we tried to analyze this。

We used the GDP, population, and the number of medals won by different countries in 2012 as data.

Since the units of each variable are different, model analysis cannot be done directly, so we first standardize the data. That is, using the scale function in R, each number in a group of numbers is subtracted from the average of the group of numbers and then divided by the root mean square of the group of numbers to eliminate the difference caused by different unit variables

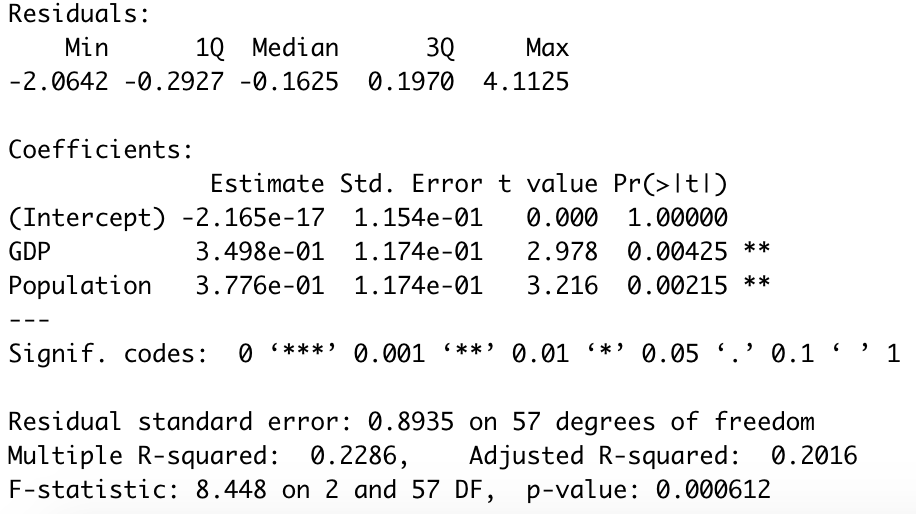
Then we checked the degree of linear correlation between the variables and got the linear correlation coefficient



From here, we can see that there is still a linear correlation between GDP, population and medals, and it is a positive correlation. This shows that with the growth of GDP and Population, the number of medals will also increase

1. *Linear model and variable selection*

We don’t try to do a linear regression model on the medals for each variable, but do a multivariate linear model first, and then check the results for selection, and then consider whether to discard certain variables.Then we did a multiple linear regression analysis of medals on GDP and populations.

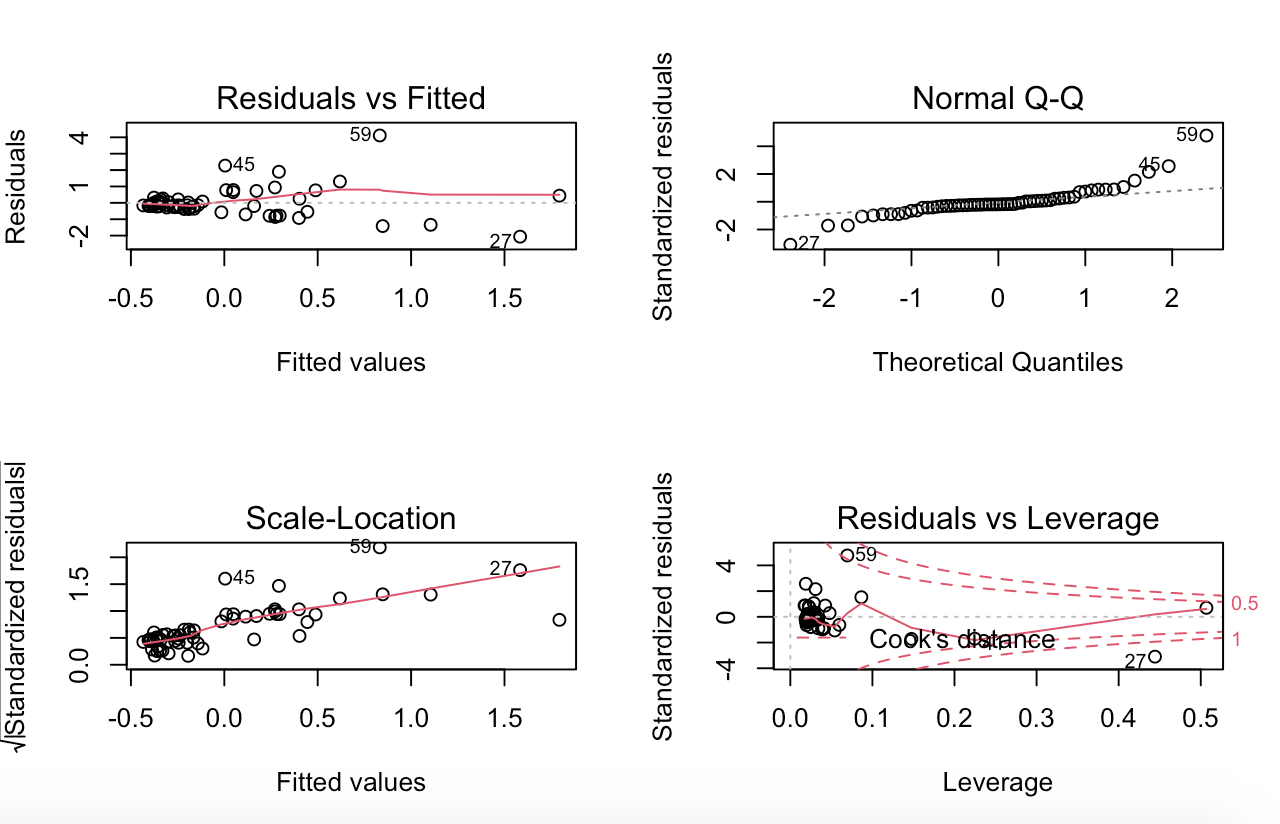


From this result, we can see that firstly, the p-value is less than 0.05, indicating that the model is significant; secondly, looking at the coefficients, we can see signif.codes, we can see that the P value of the constant term is not significant. It is greater than 0.05, but GDP and Population are both significant, so the results are not bad; finally, look at the adjusted R-square, the R-square refers to the degree of interpretation of the data by the model, we can see the R of this model The square is only 0.2, and the closer the R square is to 1, the better, which shows that the model's interpretation of the data is not enough.

From the above results, it can be seen that both the GDP and population variables have a great impact on medals, and arbitrarily discarding one will only make the model's results worse. In order to better prove this point, we used stepwise regression to select variables. The results also show that when GDP and Population exist at the same time, the AIC value of the model is the smallest, that is, the model at this time is the best. (The Akaike information criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data)

To sum up, GDP and population affect the medals at the same time, we can’t abandon either

1. *Model checking*



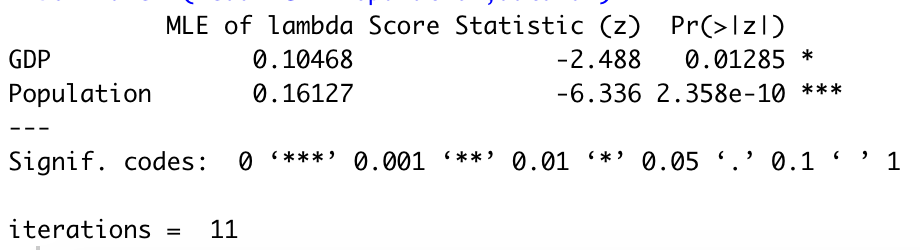
The scattered points in the upper left graph do not show any regularity, which means that there is a linear relationship. The scatter points in the upper right graph are roughly concentrated on the straight line in the QQ graph, indicating that the residuals are normal. The scatter points in the lower left graph may have rules, and the variance increases with the mean. The more the scatter points to the right, the larger the upper and lower spacing, and the more obvious the difference in variance. Check the abnormal points in the lower right picture, there may be abnormal points。

In order to make a more accurate diagnosis, we performed various tests with R. Such as: multicollinearity test, heteroscedasticity test, independence test, etc.

In these tests, We found that the model does not satisfy the assumption of homoscedasticity. In order to eliminate heteroscedasticity, we will change the model

1. *change model*

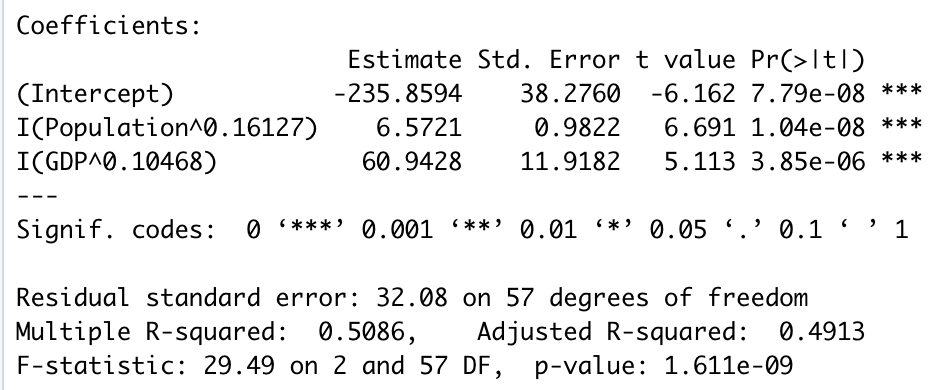
In this regard, we perform a scene transformation on the variables in the model to eliminate heteroscedasticity, and use the car package to determine the power of GDP and Population



The MLE of lambda is the best power of each variable

Get new model as:





From this result, we can see that the constant term has also become significant, and the R square has also been improved.

1. *Conclusion*

Both GDP and Population affect the number of medals, and there is a certain linear relationship. But this linear relationship is not very obvious, and the nonlinear model is more suitable for this data. After the non-linear transformation of GDP and Population, the model has been significantly improved. But the R-square of the model is still not very good. This may be because our data only has two influencing factors, GDP and Population. If we increase variables, such as latitude, weather, etc., the model may be more appropriate. But the choice of variables remains to be considered.