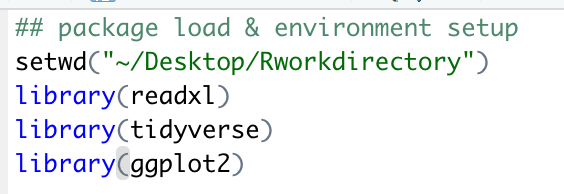
**For our sustainability indicator, we have sustainability immunization DPT statistic as an object of analysis.**

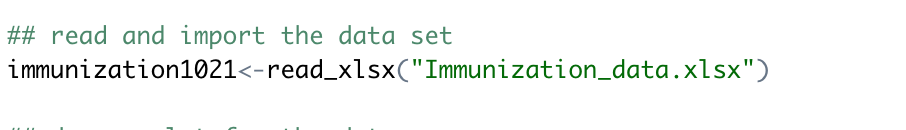
**For Our Research, We have chosen Aisa, North America, Europe, South America, Africa as our study regions.**

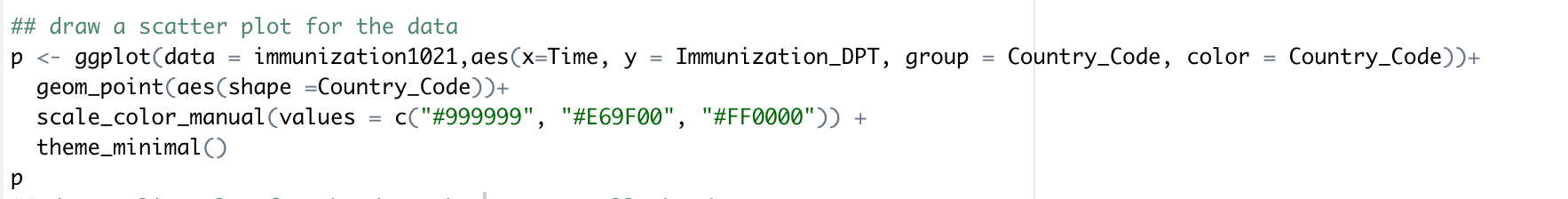
**Xiao Yang**

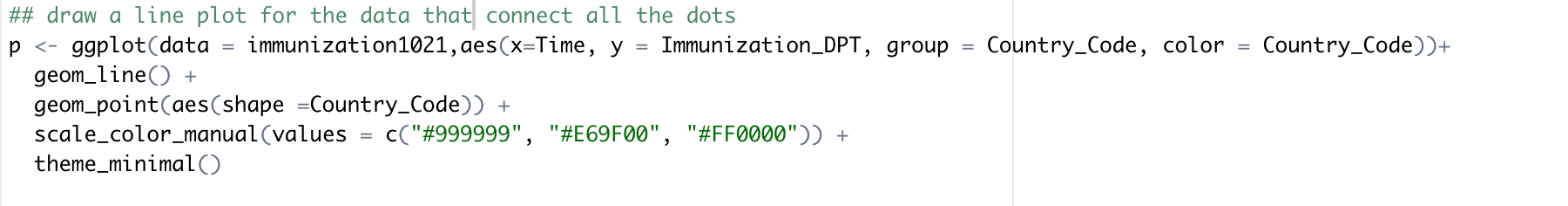
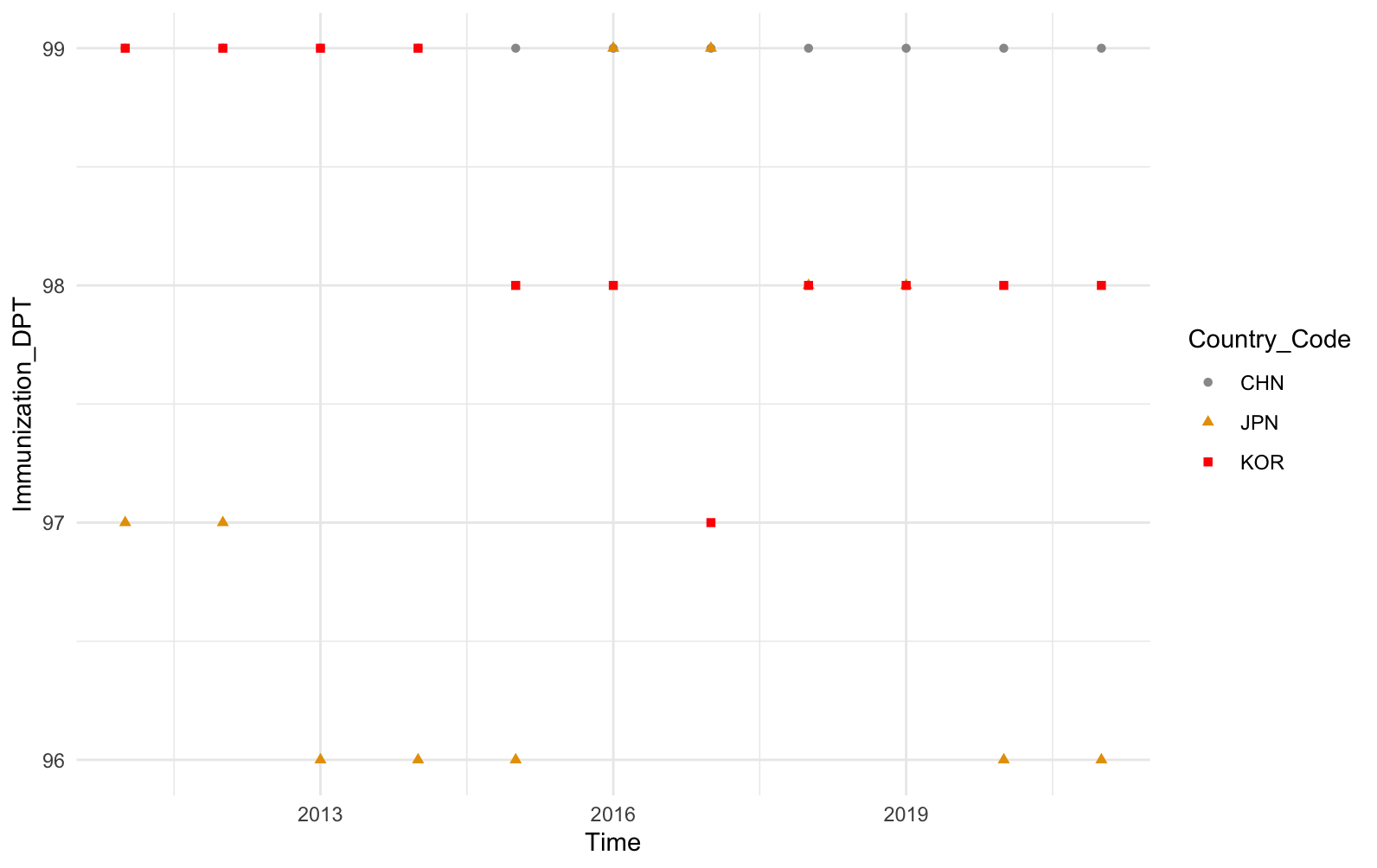
We have chosen the Immunization- DPT as our indicator for sustainability. As immunization is a critical factor of population growth, and human capital growth is a key indicator for sustainability and development.

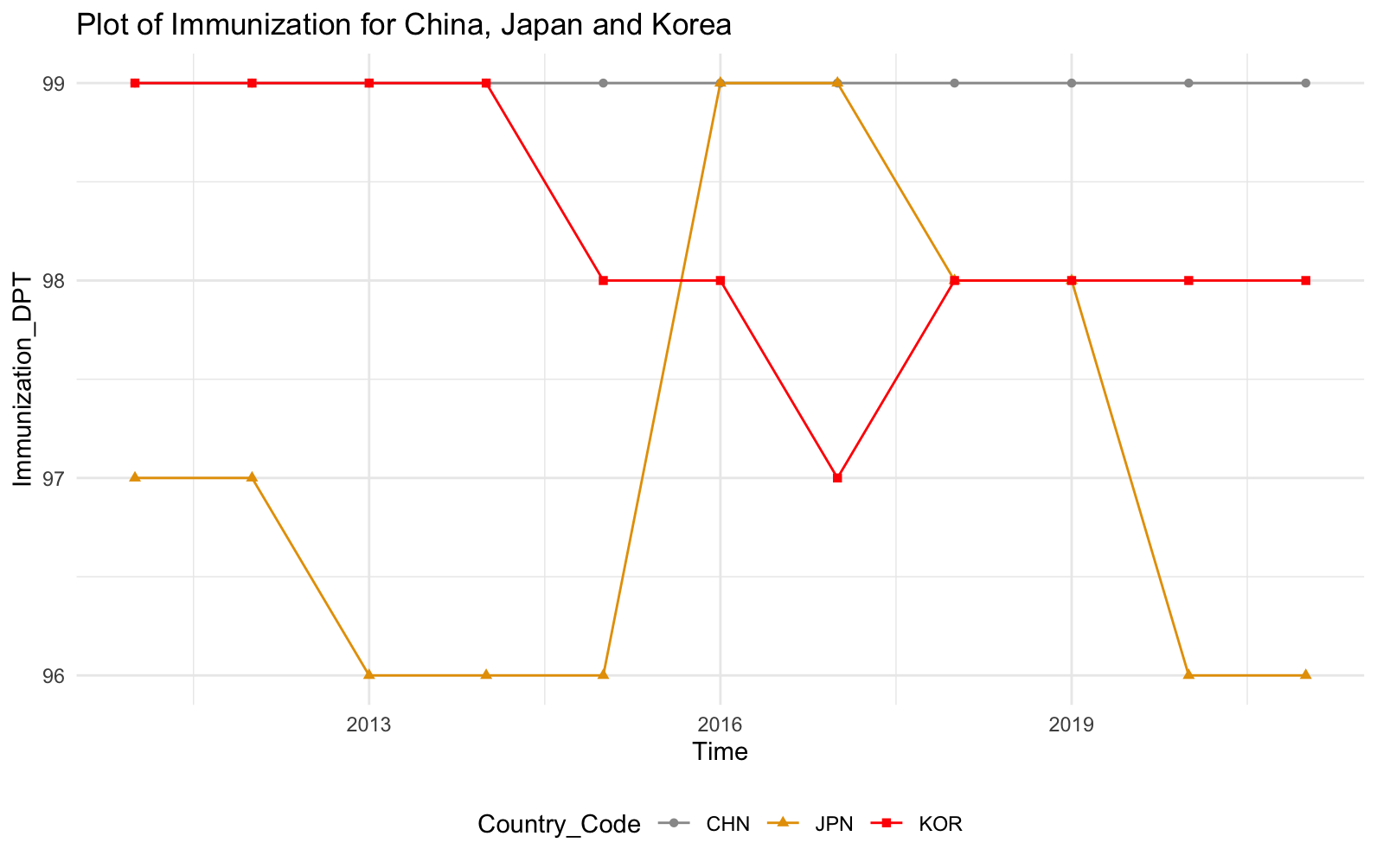


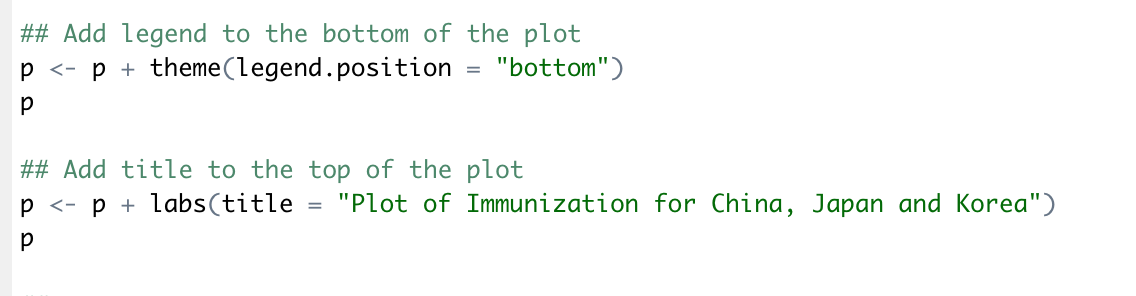
My first step is to setup environment for the project, I have loaded read excel package for data import, tidyverse package for data manipulation and ggplot package for plot



Second step is to do data loading after the data cleaning in excel



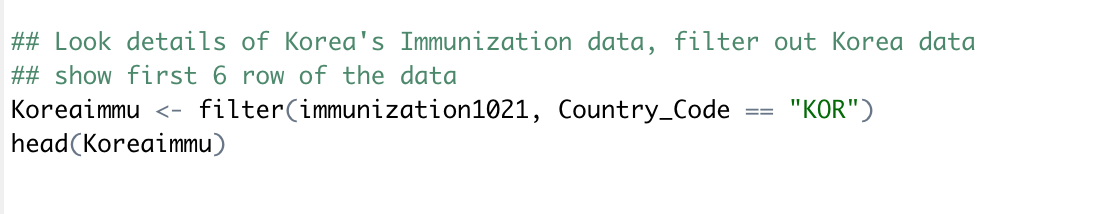




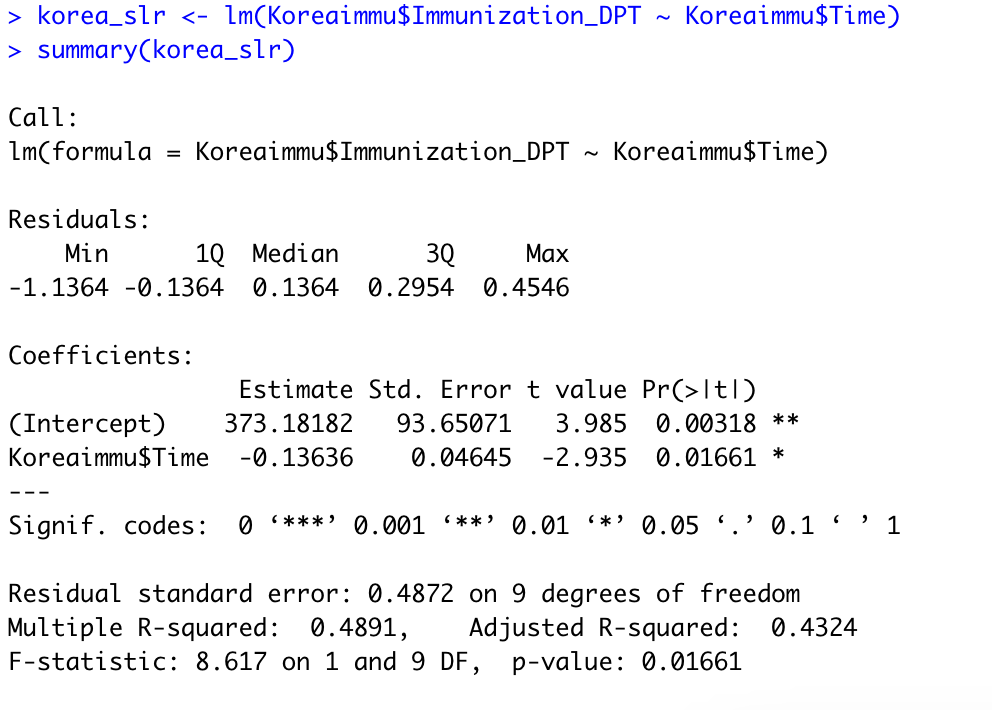
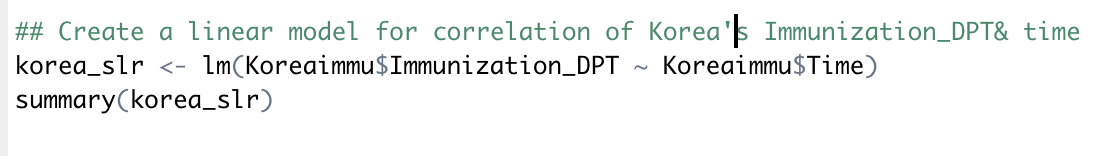
Then we will do a plot with data provided. Add legend and title with code afterward.

Based on the current plot, we can see the Immunization rate data for China is pretty stable across the year through 2010 to 2020, very little correlation exists for Immunization DPT across time for China.

For Korea, it was a slow declining trend from 99 to 98, and a little additional drop to 97 in 2017. It is a weak negative correlation for Korea Immunization data.

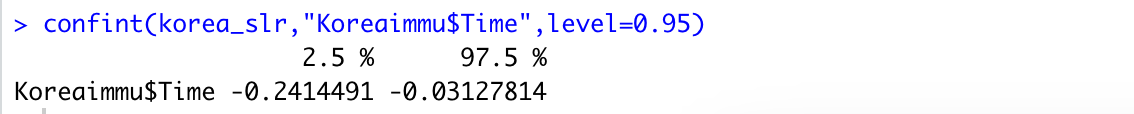
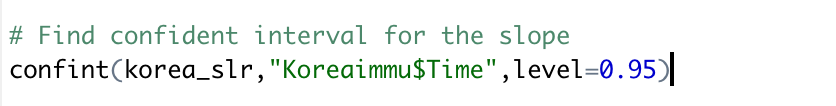
For Japan, the data varies greatly from time to time, so we could not find a linear correlation between time and immunization data. The trend is a somewhat declining trend.

As Korea has a clear linear correlation, I choose Korea as my country to do linear regression model.

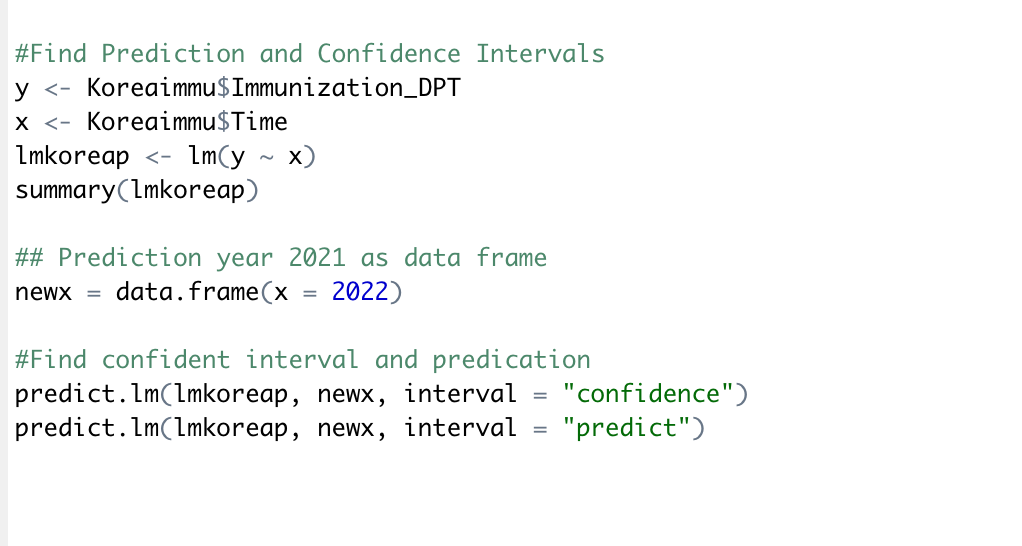


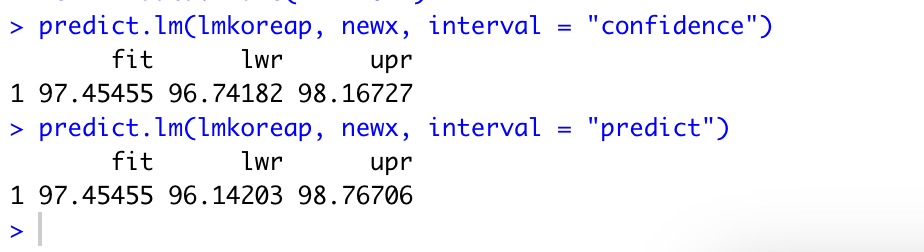
The intercept is 373.18182, and slope is -0.13636, based on adjusted R-square, it is a medium accurate linear regression model, 43.24% of data could be explained by this model.

The Linear regression equation is y=-0.13636x+373.18182

T critical here is -2.935, p-value is 0.01661, it is relatively small which indicate that the correlation is somewhat significant between Korea’s immunization rate and time

Use above function to find the confidence interval for the slope, base on the result, we are 95% confident that the slope is fall between -0.2414 to -0.03128, which mean for each unit of year change, immunizationDPT will decrease 0.2414 unit to 0.03128 unit.





Build a function based on y & x, y as the indicator , Immunization DPT and x as year.

Base on above coding and result, we can conclude that:

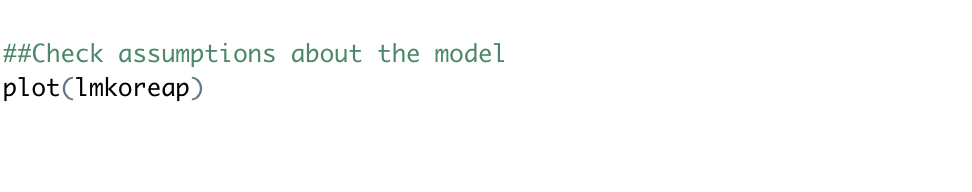
The interval of confident of 95% is between 96.74182 to 98.16727

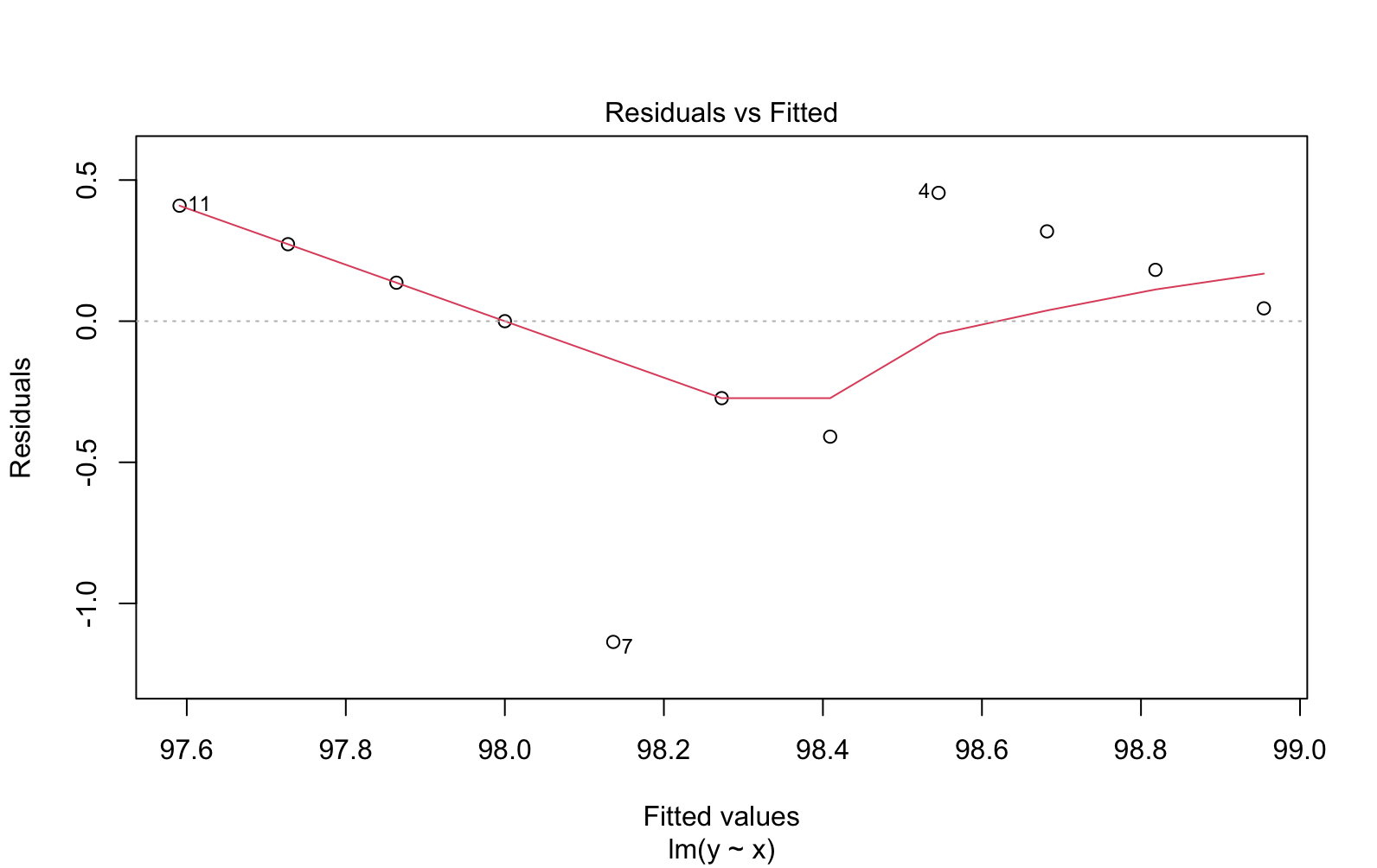
The prediction value of 2022 is between 96.14203 and 98.76706

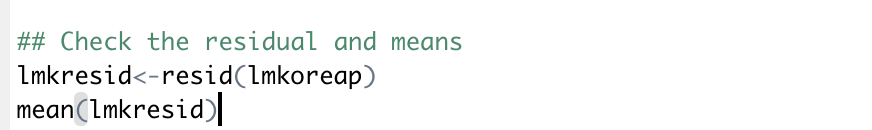
Assumptions

Random selection: We can assume random selection as it is a data set collected by the world bank as stated.

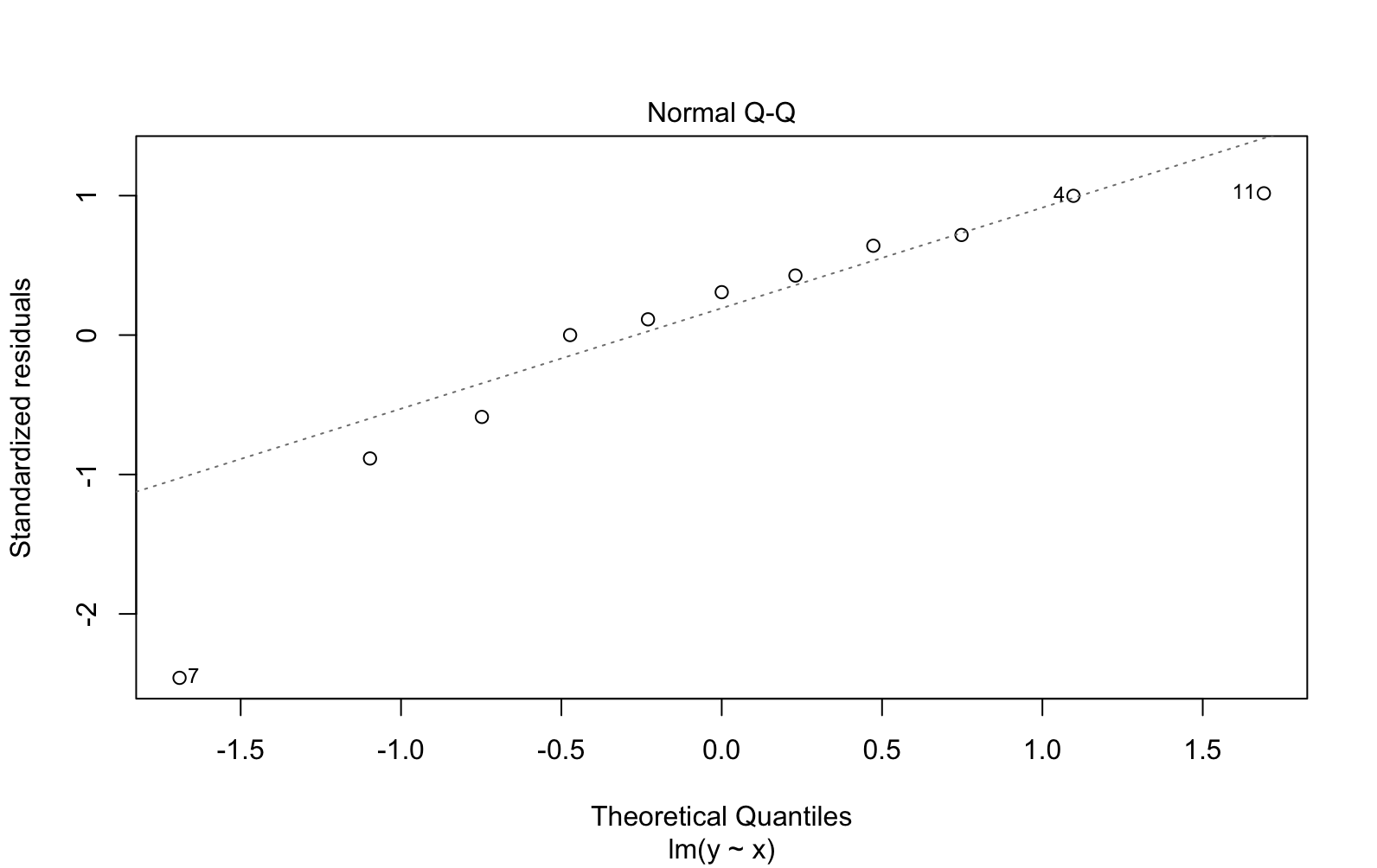
Independent: We are not sure if it is independent as it is not stated.

Linearity: Linear assumption is not met, as it shows a reverse cone shape.



Zero means assumption: Based on the result, we can assume zero means assumption.

Normality



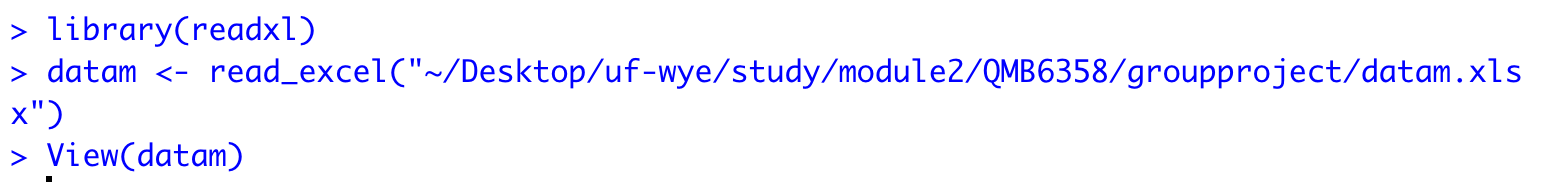
We have 2 outliers 7, and 11.

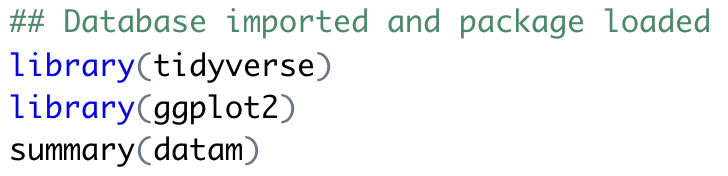
But the line is still in a linear form, so it is still an indication for normality, we can assume normality in this case

Uniformed Spread: Not met, because residual was displayed in a reverse cone shape.

—-----------------------------------------------------------------------------------------------------------------

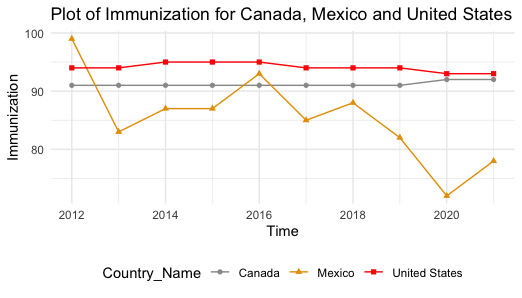
**Yuaner Wang:**

Country: Canada, Mexico and United States



My first step is to setup environment for the project, I have imported database, loaded tidyverse package for data manipulation and ggplot package for plot and summary the data information.



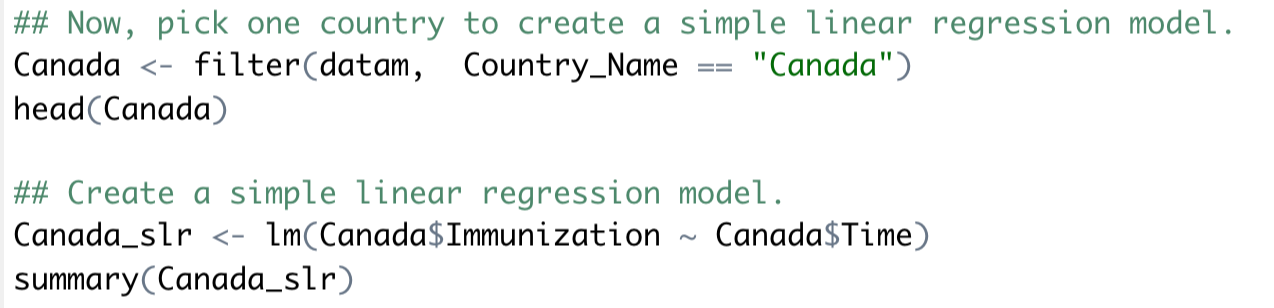


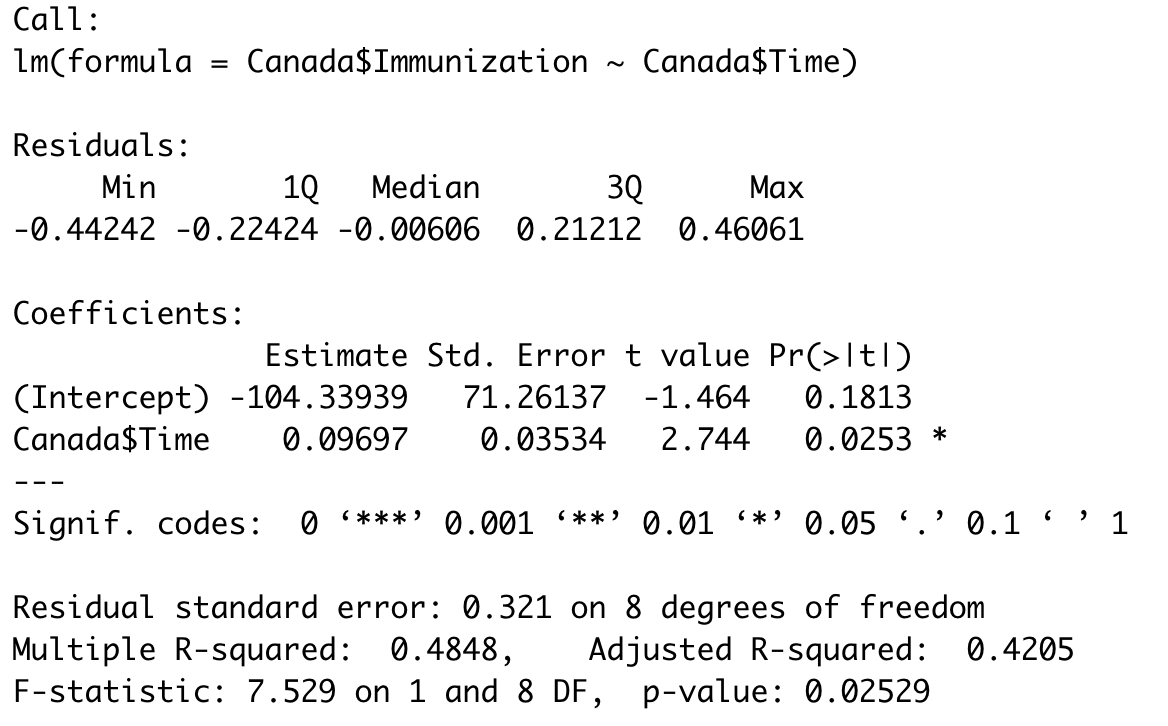
Then we will do a plot with data provided. Add legend and title with code afterward.

Based on the current plot, we can see the Immunization rate data for Canada and United States is pretty stable across the year through 2012 to 2021, very little correlation exists for Immunization DPT across time for Canada and United States .

For Mexico, it’s a wave-shaped plot. so we could not find a linear correlation between time and immunization data.

As Canada has a clear linear correlation, I choose Canada as my country to do linear regression model.



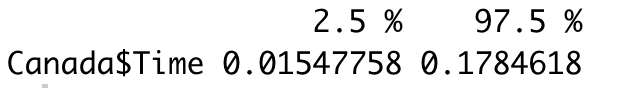


The intercept is -104.33939, and slope is 0.09697, base on adjusted R-square, 42,05% of data could be explained by this model.

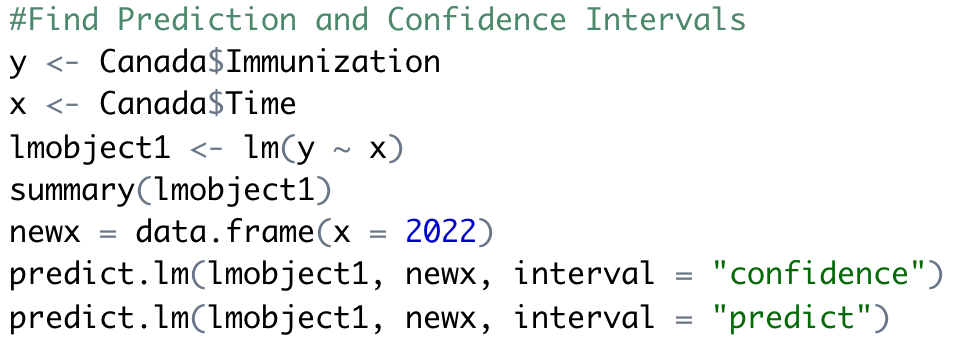
The Linear regression equation is y=0.09697x -104.33939

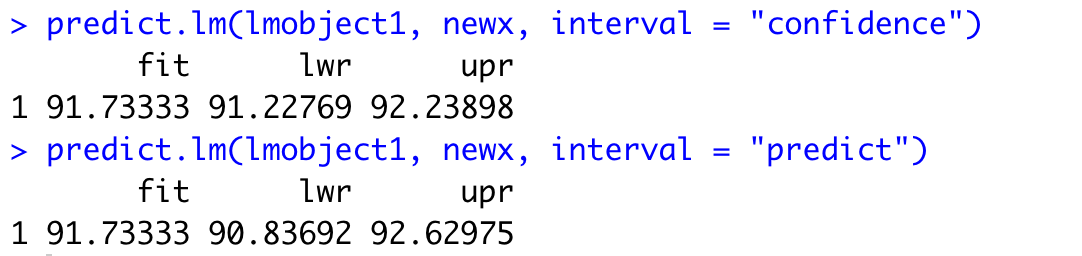
T critical here is 2.744, p-value is 0.0253, it is relatively small which indicate that the correlation is somewhat significant between Canada’s immunization rate and time.





Use above function to find the confidence interval for the slope, base on the result, we are 95% confident that the slope is fall between 0.01547758 to 0.1784618, which mean for each unit of year change, immunizationDPT will increase 0.01547758 unit to 0.1784618 unit.





Build a function based on y & x, y as the indicator , Immunization DPT and x as year.

Base on above coding and result, we can conclude that:

The interval for confidence is (91.22769, 92.23898). We are 95% confident that the average immunization percentage equal to 2022 is between 91.22769 and 92.23898 percents.

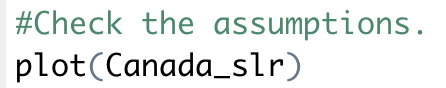
The interval for predict is (90.83692, 92.62975). We are 95% confident that 95% of all immunization with Year 2022 is between 90.83692 and 85.01 92.62975.

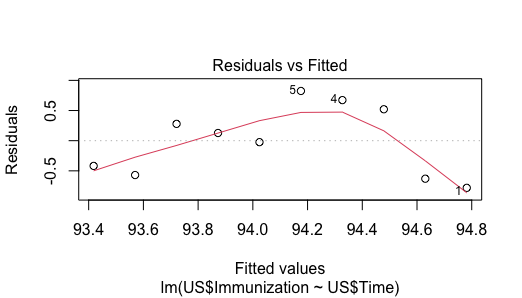
Assumptions

Random selection: We can assume random selection as it is a data set collected by the world bank as stated.

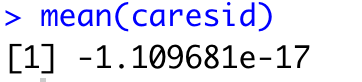
Independent: We are not sure if it is independent as it is not stated.

Linearity: Linearity is not met, because there’s a cone pattern between residual and predicted.

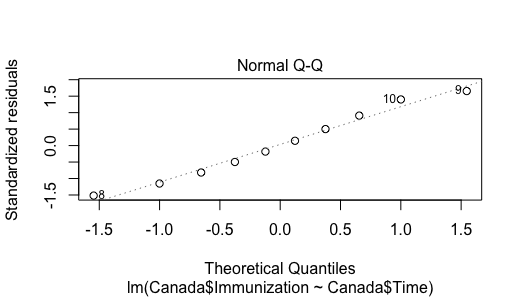




Zero means assumption: Base on the mean almost zero, we can assume zero means assumption.



Normality

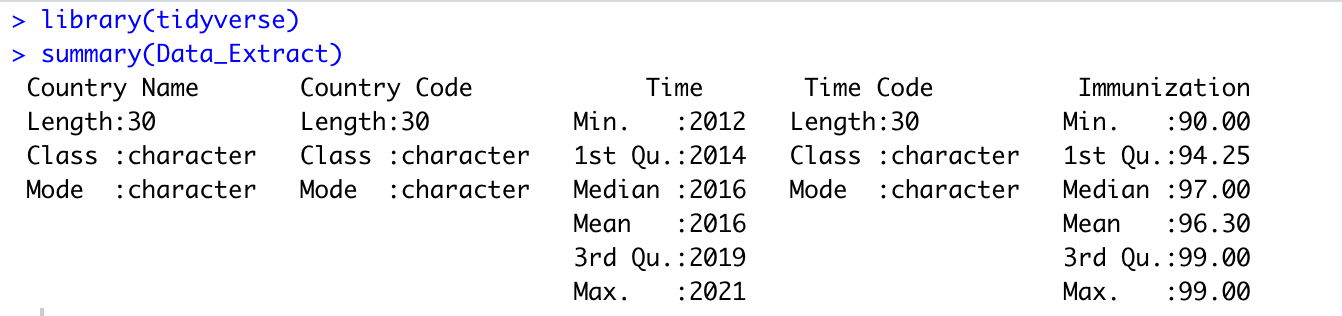


We have three outliers, but it’s okay. So we meet this assumption.

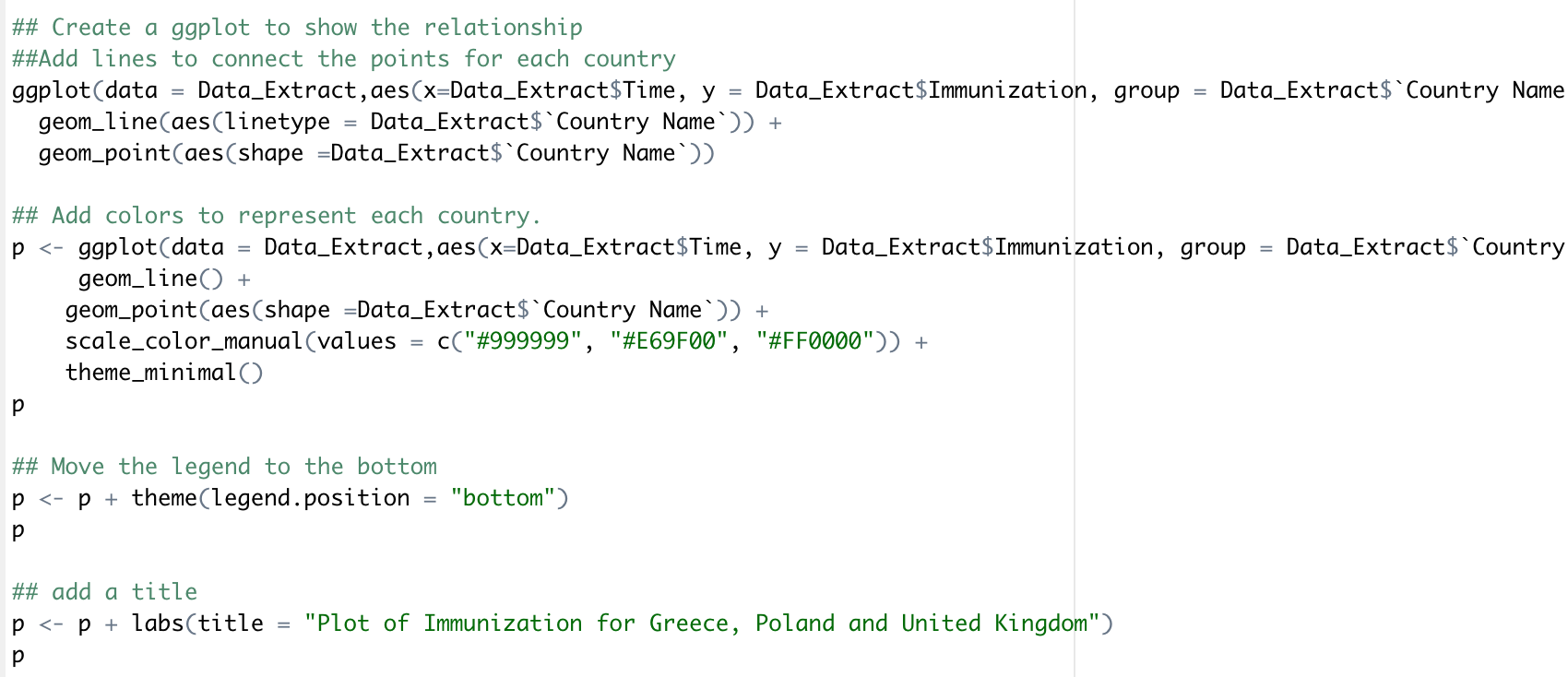
Uniformed Spread: Not met.

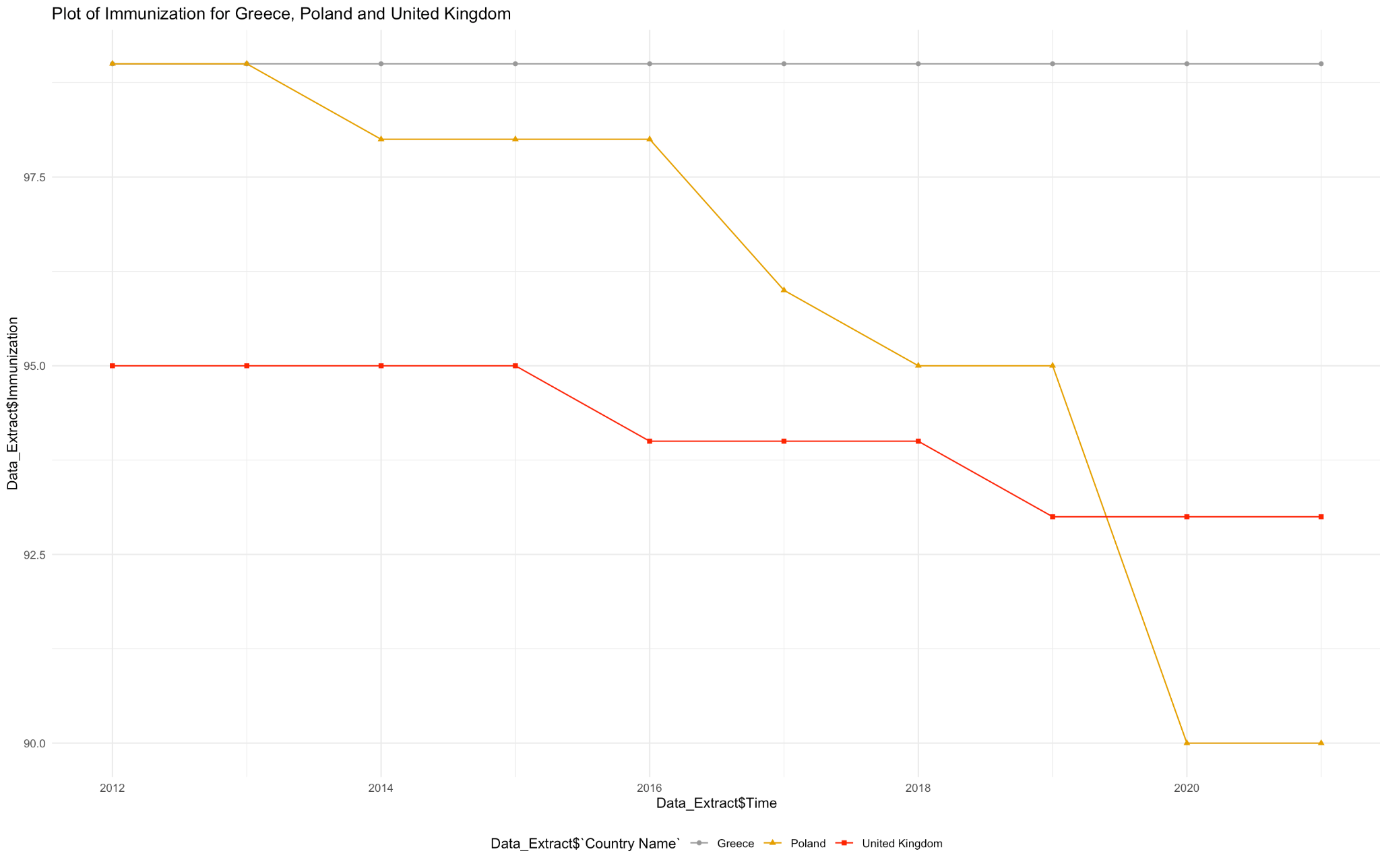
**Wenyu Han**

1. Firstly, setup circumstance, I am going to look at “Immunization, DPT (% of children ages 12 - 23 months)” for Greece, Poland and the United Kingdom.



1. Create a ggplot to show the relationship, adding lines to connect the points for each country



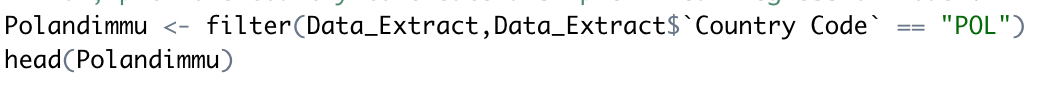


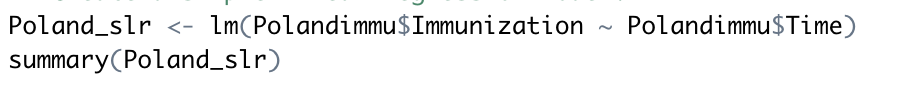
According to the plot picture, the immunization rate for Greece is stable from 2012 to 2021, very little correlation exists for Immunization DPT across time for Greece.

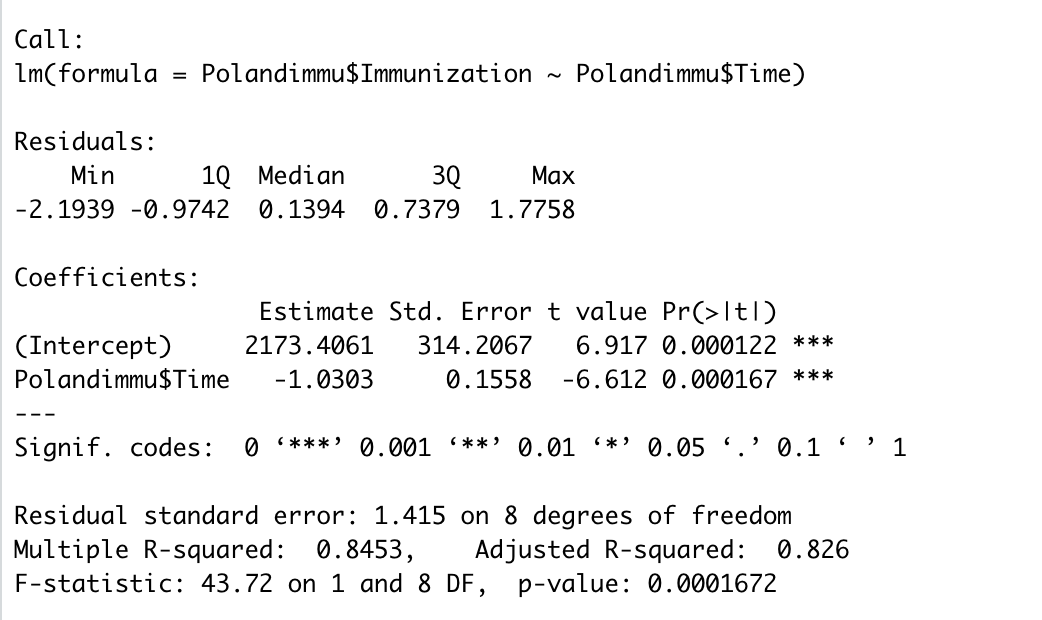
For Poland, the overall downward trend is evidently obvious compared to Greece and the United Kingdom. Especially the data from 2016 to 2018 and 2019 to 2020 declined significantly.

The immunization rate of United Kingdom also has a decreasing trend. However, compared to Poland, the downward trend is relatively moderat. In 2015, the immunization declines from 95 to 94. And in 2019, it drops from 94 to 93.

1. Since Poland has a clear and significant trend, I choose it as my country to do simple linear regression model.



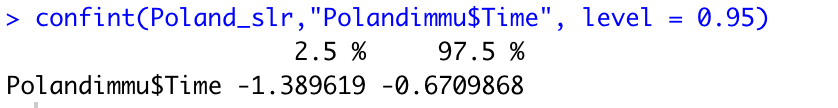




In data summary, the intercept value is 2173.4061, and slope is -1.0303. The p-value for time is lower than 0.01, indicating this variable is statistically significant at the 99% confidence interval. Based on adjusted R-square, it is a highly accurate linear regression model, 82.60% of data could be explained by this model.

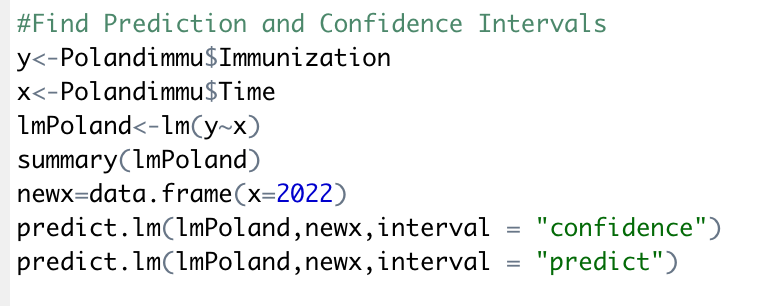
The Linear regression equation is y=-1.0303x + 2173.4061

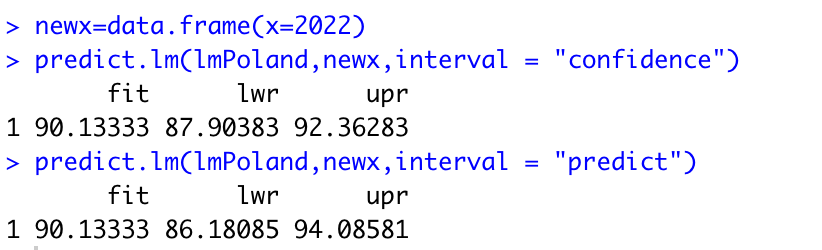
1. Find confident interval for the slope



Use above function to find the confidence interval for the slope, base on the result, we are 95% confident that the slope falls between -1.389619 and -0.6709868, suggesting immunizationDPT will decrease from 1.389619 unit to 0.6709868 unit as time (unit: year) goes by.

1. Find Prediction and Confidence Intervals





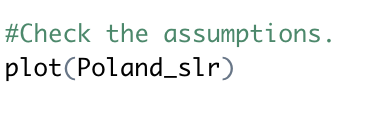
Build a function based on y & x, y as the indicator , Immunization DPT and x as year.

Based on above coding and result, we can conclude that:

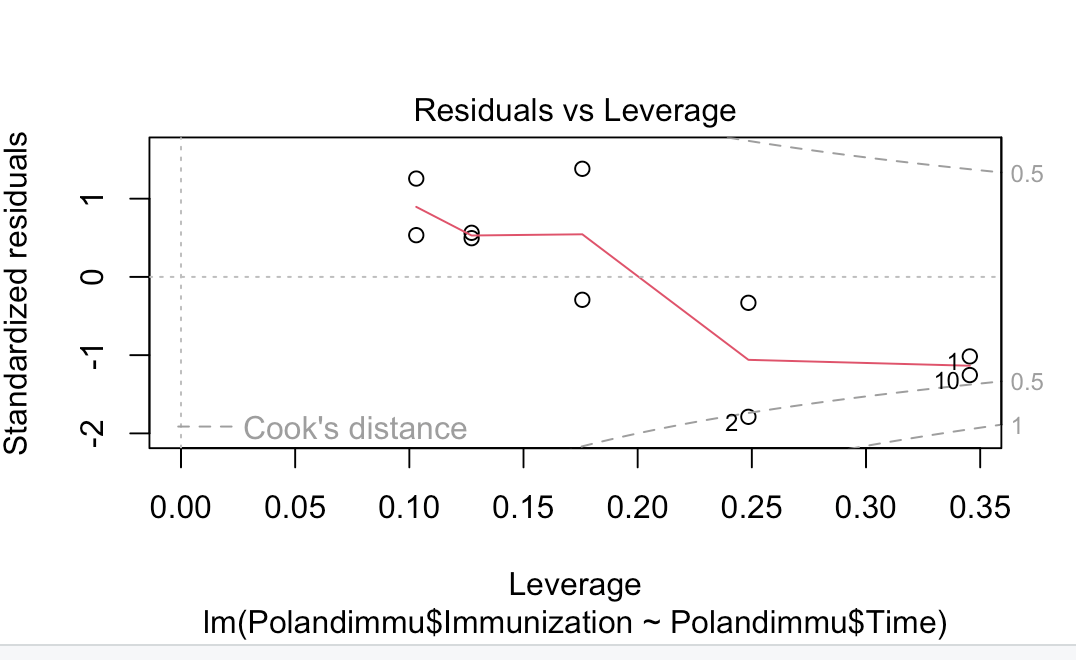
The interval of confident of 95% is between 87.90383 to 92.36283

The prediction value of 2022 is between 86.18085 and 94.08581

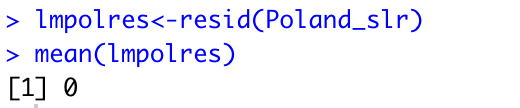
1. Check the assumptions.



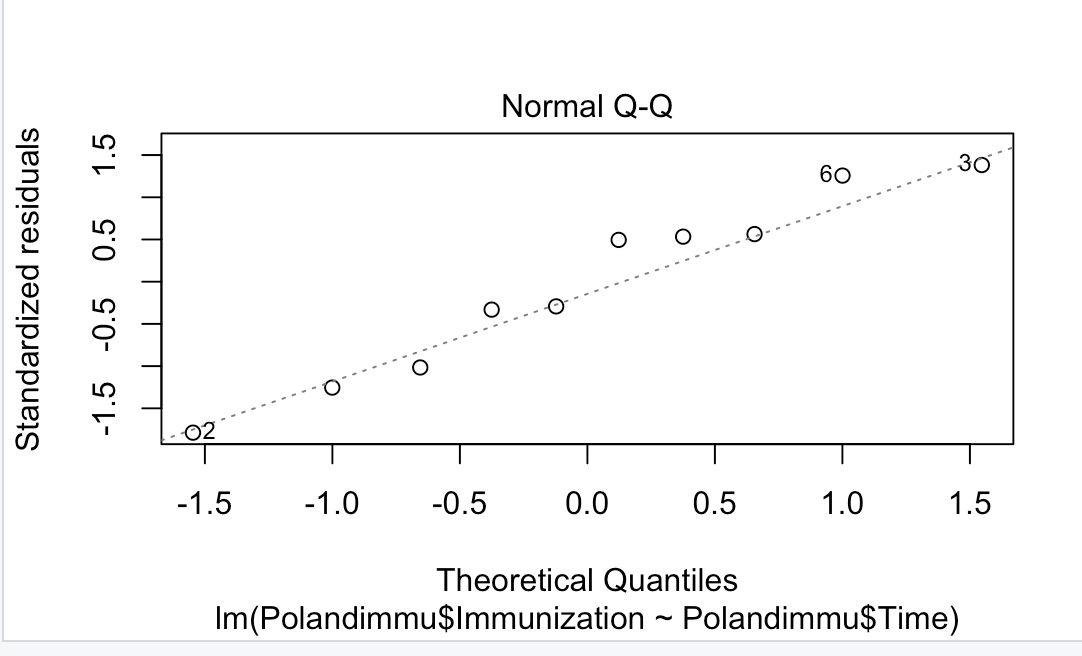
1. Linearity Assumption → Linearity is not met, because there’s a cone pattern between residual and predicted.



1. Zero Mean→ Assumption is met, the mean is zero.



1. Uniform Spread →This assumption is not met. The residuals are also not uniformly spread because the cone pattern in residual vs. fitted plot.
2. Independence → Yes, we assume the data is independent because it’s random sample.
3. Normality: we can check the normality from the standard Q-Q plot. If the data conform to the normal distribution, all data points should lie on the straight diagonal line. However, for my scenario, most of the points deviate from the diagonal line, implying the normality assumption may not hold.



1. Randomness → Yes, we assume the data is random because the data is randomly picked up from different counties,random sample.

**Ruiying Zhu**

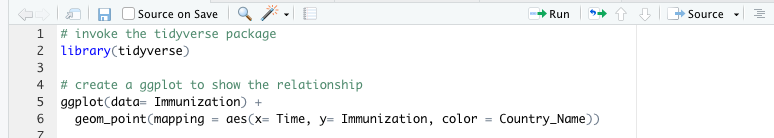
Country: Peru, Brazil, Argentina

Series: Immunization, DPT(% of children ages 12-23 months)

Time: 2012 - 2021

Invoke the tidyverse.

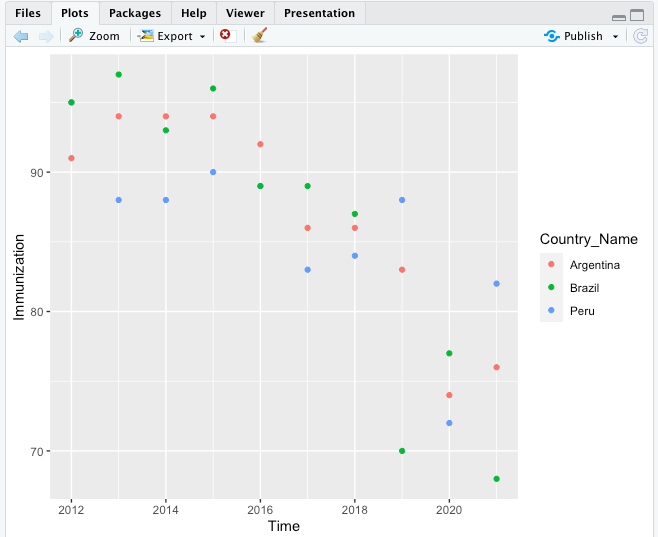
Coding:



First, use the ggplot to make a simple scatterplot of all the datapoints between Time and Immuniztion data and sort them by Country\_Name.

Coding:



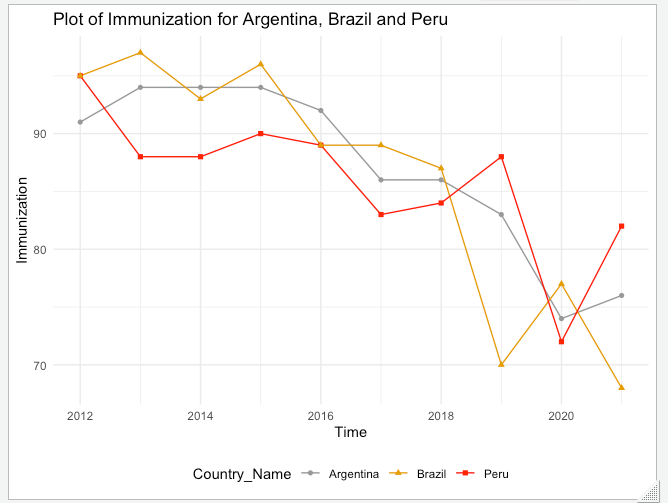


As we could see, there are no extreme outliers (maybe one outliers from Brazil) and it’s a negative and maybe linear relationship between time and immunization data. However, this graph is still not clear enough to demonstrate the relationship of the indicator across time for each country. First, we want to add different lines to connect the points for each country and add colors to each lines. I also add the labels and the title to make the graph more readable.

Coding:



Here’s the final graph.



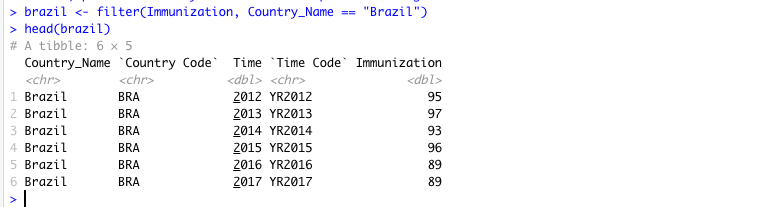
As we could see in the graph, there’s a decrease trend in the beginning and an increase trend in the end for Peru. I assume it has a weak negative and linear relationship between time and immunization because of two increased points in Year 2019 and 2021. No extreme outliers are shown for Peru.

There’s an initially stable but then decrease trend in Argentina and it has a kind of strong negative and linear relationship between time and immunization. No extreme outliers are shown for Argentina, either.

For country Brazil, there’s a strong negative and linear relationship between time and immunization because it has two very low points around 2019 and 2021. It has an obvious decrease trend in the end but no extreme outliers are shown.

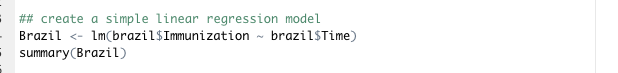
Each person should then pick one country. I chose the country Brazil by using the filter function to filter the Brazil data from the database.

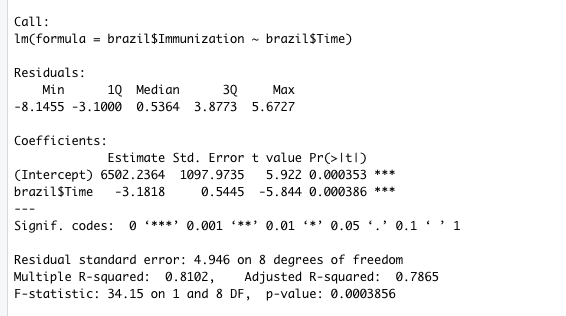
Coding:



Part 1. Now, we can create a linear regression model.

Coding:





We found that the y-intercept is 6502.2364, but this data is meaningless and we do not need to interpret because no data around x = 0. The slope of time vs. immunization is -3.1818. -3.1818 means the average decrease of immunization per increase in one unit of time. The equation should be

Y = -3.1818 X\_time + 6502.2364

Part 2. Conduct the t test for the slope and the confidence interval for the slope.

Parameter: The true change in immunization in one unit increase in time.

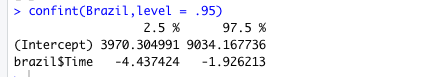
Hypothesis: H0 (null hypothesis): Beta1 = 0

Ha (alternative hypothesis): Beta1 is not equal to zero.

T value = -5.844 and p value is almost zero.

With a p value is almost zero, we have very strong evidence that the true change in immunization in one unit increase in time is not zero.

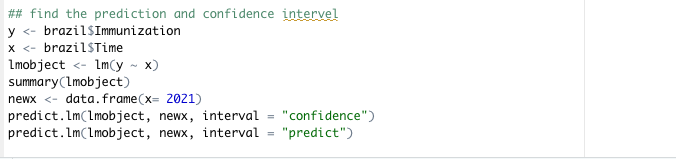
Coding: (confident interval of slope)

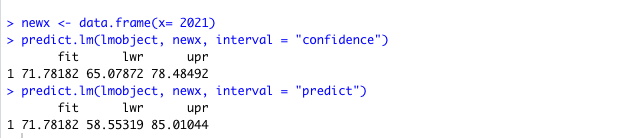


We are 95% confident that the average immunization per time is between -4.43742 and -1.92621 percents.

7 (part3) Find the confidence interval and the prediction interval for the next year after the data was collected.

Coding:





Conclusion:

The interval for confidence is (65.08, 78.48). We are 95% confident that the average immunization percentage equal to 2021 is between 65.08 and 78.48 percents.

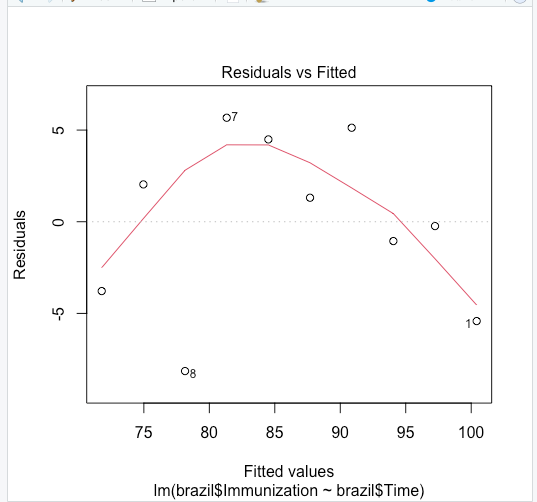
The interval for predict is (58.55, 85.01). We are 95% confident that 95% of all immunization with Year 2021 is between 58.55 and 85.01 percents.

Part 4. Now we want to check the assumptions:

1. Linearity Assumption → Linearity is not met, because there’s a cone pattern between residual and predicted.

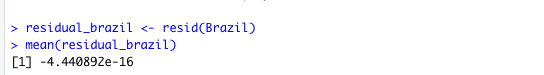
Coding:



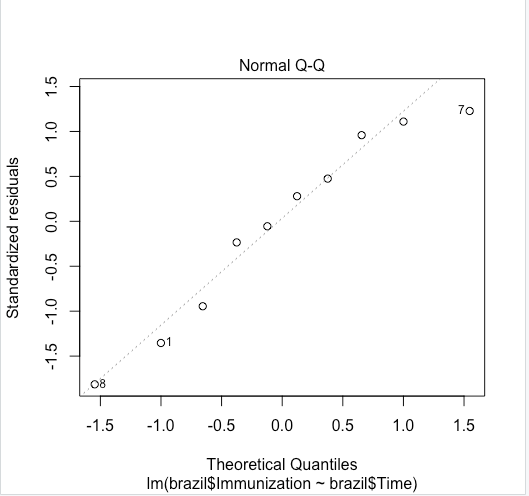


1. Zero Mean→ Assumption is met, the mean is almost zero.

Coding:



1. Uniform Spread →This assumption is not met. The residuals are also not uniformly spread because the cone pattern in residual vs. fitted plot.
2. Independence → Yes, we assume the data is independent because it’s random sample.
3. Normality → We have 2 outliers, but it’s okay. So we meet this assumption.



1. Randomness → Yes, we assume the data is random because the data is randomly picked up from different counties.

**Clare Huang**

1. Country: Egypt, Kenya, Nigeria

Series: Immunization, DPT (% of children ages 12-23 months)

Time: 2012 - 2021

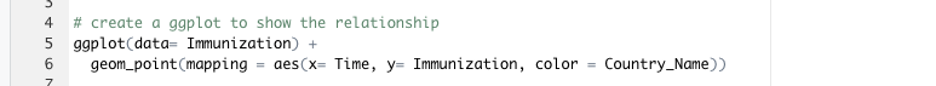
Invoke the tidyverse.

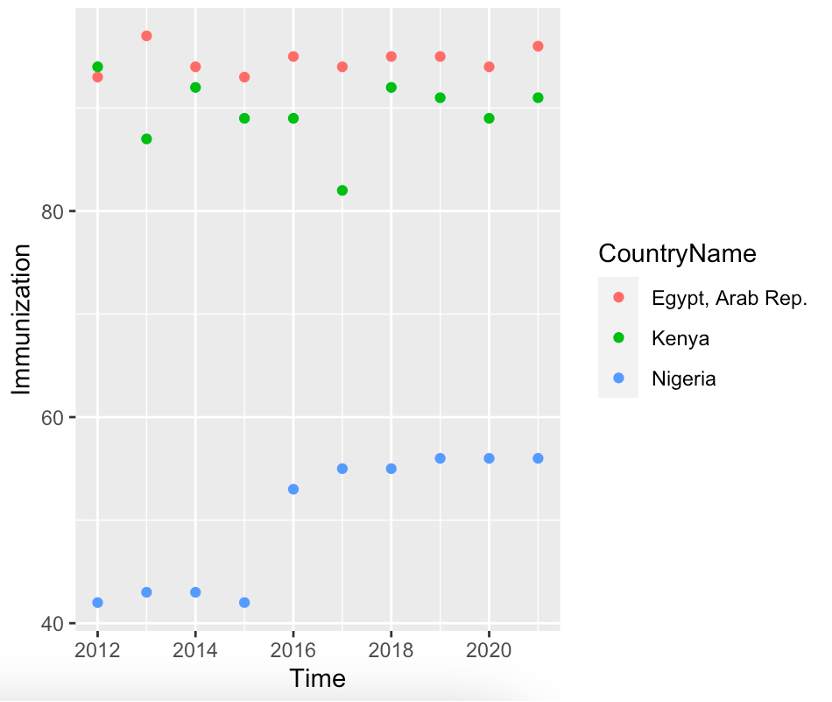
Coding:



First, use the ggplot to make a simple scatterplot of all the datapoints between Time and Immunization data and sort them by CountryName.

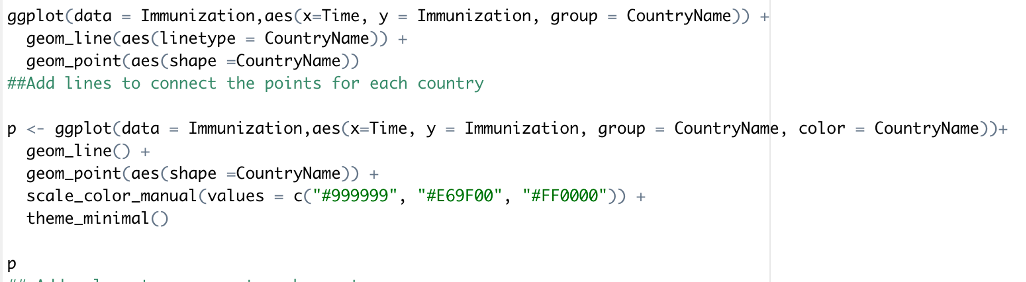
Coding:



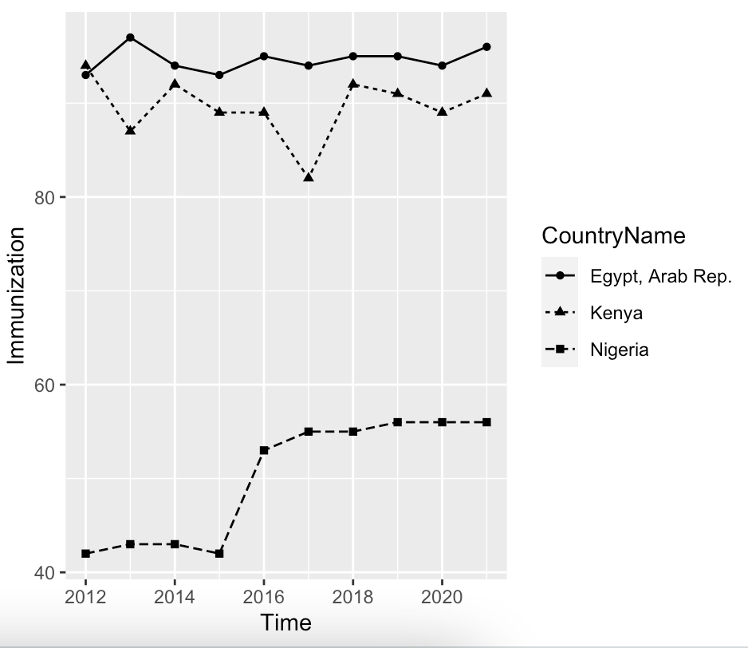


As we could see, there are no extreme outliers (maybe one outliers from Kenya) and there seems to have linear relationship between time and immunization data. In order to make this graph clearer to demonstrate the relationship of the indicator across time for each country, we want to add different lines to connect the points for each country and add colors to each lines.

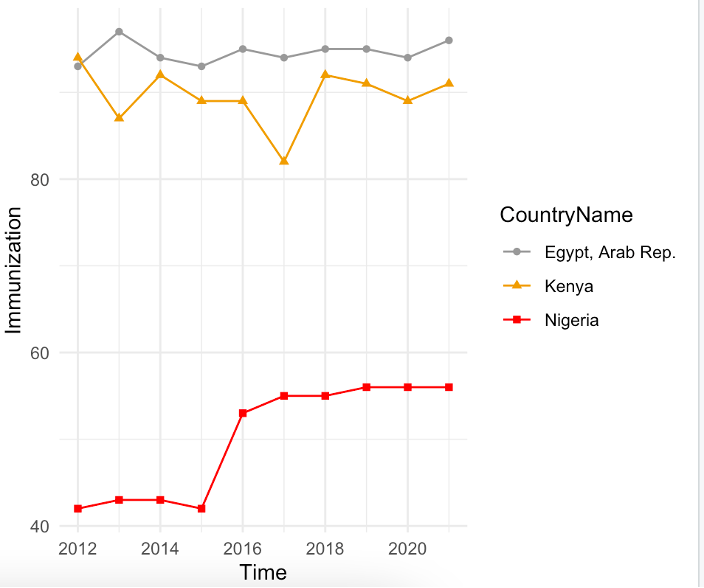
Coding:



Here’s the graph which I added lines to connect the points for each country:



Here’s the final graph:

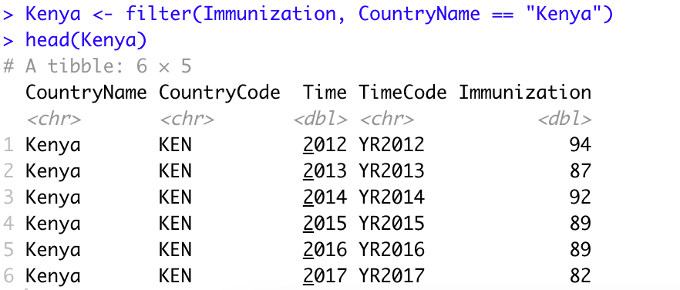


In this graph, we can see that there’s an increase trend in the middle for Nigeria. It is a weak positive correlation for Nigeria Immunization data. The line for Egypt is quite stable. I assume it has a linear relationship between time and immunization because of the suddenly increased point in Year 2016.

No extreme outliers are shown for all three countries. There’s an initially stable but then decrease trend in Kenya and it has a kind of stable linear relationship between time and immunization. It is a weak negative correlation for Kenya Immunization data. No extreme outliers are shown for Kenya, either.

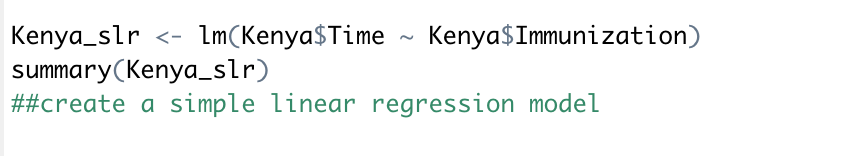
2. Each person should then pick one country. I chose the country Kenya by using the filter function to filter the Kenya data from the database.

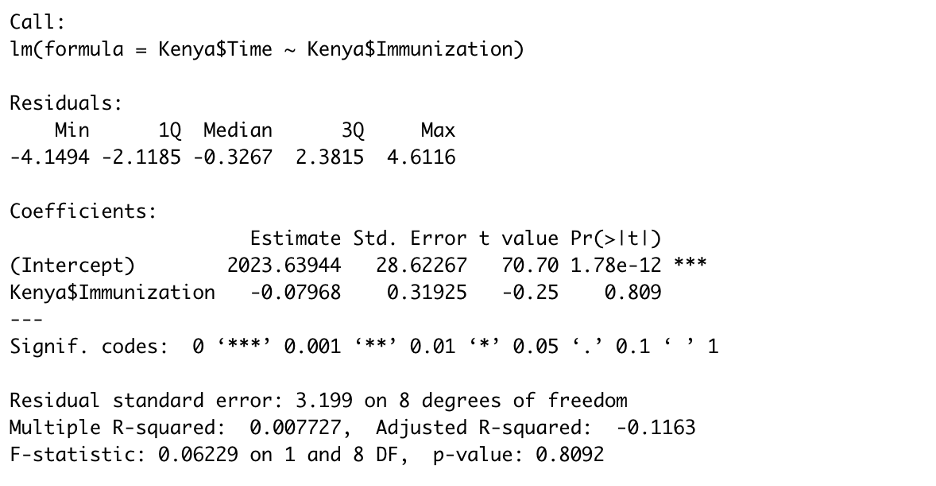
Coding:



**Part 1. Now, we can create a linear regression model.**

Coding:





We found that the y-intercept is 2023.63944, but this data is meaningless and we do not need to interpret because no data around x = 0. The slope of time vs. immunization is -0.07968. -0.07968 means the average decrease of immunization per increase in one unit of time. The equation should be

Y = -0.07968X\_time + 2023.63944

**Part 2. Conduct the t test for the slope and the confidence interval for the slope.**

Parameter: The true change in immunization in one unit increase in time.

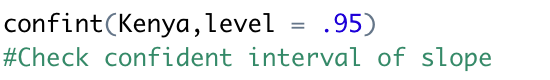
Hypothesis: H0 (null hypothesis): Beta1 = 0

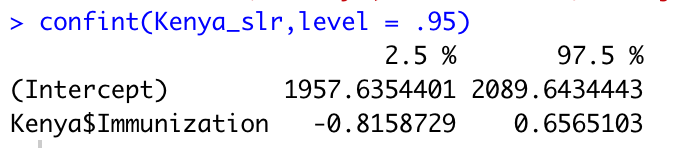
Ha (alternative hypothesis): Beta1 is not equal to zero.

T value = -0.25 and p value = 0.809.

With a p value of 0.809 which is bigger than alpha, we don’t have strong evidence that the true change in immunization in one unit increase in time is not zero.

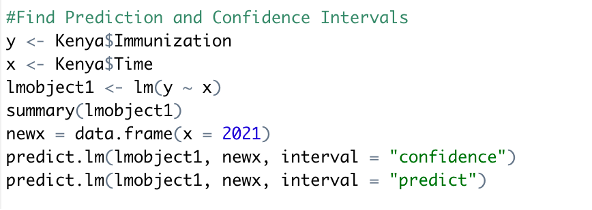
Coding: check confident interval of slope.

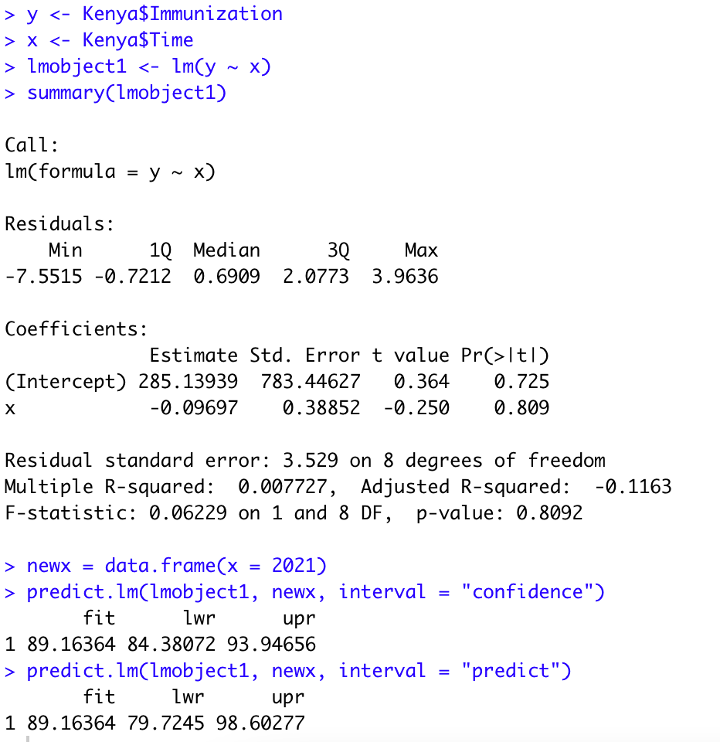




**Part 3. Find the confidence interval and the prediction interval for the next year after the data was collected.**

Coding:





Conclusion:

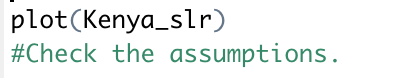
The interval for confidence is (83.38, 93.94). We are 95% confident that the average immunization percentage equal to 2021 is between 83.38 and 93.94 percent.

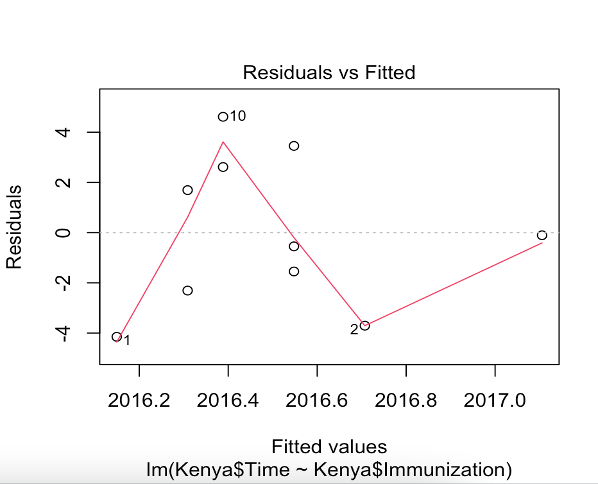
The interval for predict is (79.72, 98.60). We are 95% confident that 95% of all immunization with Year 2021 is between 79.72 and 98.60 percent.

**Part 4. Now we want to check the assumptions:**

1. Linearity Assumption → Linearity is not met, because there’s a cone pattern between residual and predicted.

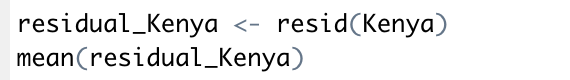
Coding:

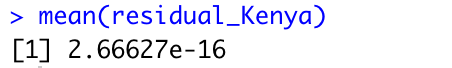




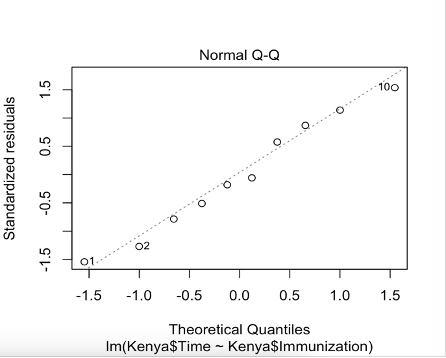
1. Zero Mean→ Assumption is met. Per the result, the mean is almost zero.

Coding:





1. Uniform Spread →This assumption is not met. The residuals are also not uniformly spread because the cone pattern in residual vs. fitted plot.
2. Independence → Yes, we assume the data is independent because it’s random sample.
3. Normality → There are 3 outliers (see numbers 1, 2, and 10 points in the graph), but otherwise, it's normal because it's a linear and straight line.



1. Randomness → Yes, we assume the data is random because of random sample.

Group Question:

**Q1. Poses a question about the world development indicator. Describe the variable that you chose. Please explain why this world development indicator is important and why you chose it. Who collected the data? To which area of sustainability does it relate?**

A: We have chosen the Immunization- DPT as our indicator for sustainability. As immunization in DPT(Diphtheria, pertussis, and tetanus) is a critical factor of population growth. Population indicates human capital growth and workforce for the human aspect of sustainability and social aspect development.

**Q2. Compares the countries within each region. Within each region is there a consistent trend and direction? Does one country stand out?**

Aisa-X.Y.: For the countries in Asia, the trend for Immunization rate change isn’t consistent across the three countries I chose. China’s immunization rate seems not to be affected by time, possibly because of the large population. Japan’s data was varying across the metric, almost no trend existed. Only Korea has a consistent trend of declining.

North America-Y.W.: For the Asian countries, the trends in immunization rates were not consistent for the three countries I chose. Immunization rates in Canada and the United States do not appear to be affected by time. Mexico had a consistent declining trend.

Europe- W.H.: For the European countries, even though there is no consistent trend among the three countries I selected, the general trend is broadly consistent. Immunization rates in Greece don’t be affected by time, maybe since the data is not intact. However, the rates in Poland and the United Kingdom have a consistent declining trend, especially in Poland.

South America- R.Z.: For the South America countries, the trend for Immunization rate change in three countries is kind of not consistent and all of them are in a decrease direction. Even though Peru’s immunization rate increase a little bit in the end, it’s still decrease in the whole picture. There’s no countries stand out in among the three countries.

Africa- C.H.: For the African countries, I was unable to find a consistent trend among the three countries I selected. However, all of the three countries show a consistent trend in immunization rates. Immunization rates in Egypt and Kenya don't seem to be affected by time but the rates in Nigeria showed a tendency of increasing by time.

**Q3. Compare the regions with each other. Across regions, was there a consistent trend and direction? Was there a reason that stood out?**

A: Based on the data analysis we have conducted, we can see that there is a trend of slowly declining immunization DPT rate. The Trend was weak across most of the regions, but the declining was pretty consistent.

We suspect this might be caused by covid spread, which could cause the sample size to become smaller, as researchers are less likely to go out there and collect data. This might cause a certain type of bias. Another possible downside cause by covid is people have less trust in vaccination, and less likely to engage in immunization after covid.

**Q4. Summarizes the narrative.**

A: Upon checking random samples chose from different continents, we can come up with a conclusion that in total, the DPT immunization rate has a tendency to decrease. This phenomenon also means that population that are in need of DPT vaccination is decreasing. Since DPT immunization is required for all children in every country, we can also learn the circumstances of total population growth in different countries from this phenomenon. Additionally, as population growth is a key point for us to observe sustainability, we can indirectly establish a relationship between the DPT immunization rate and sustainability development for different countries.