Sustainability Exploration

Pavana Teja Viswanath Josyabhatla, Satya Sai Koppalu, Colin Glen Fernandes

Question

Is there a linear trend in the unemployment rates for different regions across the globe from the year 2010 to year 2019, that affect the sustainability goal of "Decent work and Economic Growth" of United Nations?

Sustainability Development Goal

For this project we have chosen the "Decent Work and Economic Growth" sustainability development goal form the United Nations website.

The advent of development and progress of societies around the world drastically increased with the use of currency. Till date, currency plays an important role in every individual's life and the worlds future. Although the past few centuries saw great economic growth, there has been a drastic decrease in the last few years due to various reasons- increase in population, inflation, pandemics, etc. With the world population predicted to peak at a 11 billion in the year 2100 (Source: World Population Prospects 2019), these problems of stagnating or decreasing economic growth can be solved by promoting sustainable economic growth by providing decent employment and work to people. Problems of unimaginable scale are still faced by countries around the world which can be solved by a sustained and positive economic growth.

One of the major misconceptions identified in this goal is- the data used to estimate the economic growth around the world can be biased when the analysis is done by grouping all countries together since each country has different factors that may have to be considered. While one country has less population therefore resulting in a higher employment rate, another different country might have a higher population therefore resulting in a lower employment rate even though both the countries have the similar number of people employed. To test this, we decided to compare the total unemployment rates of different countries in different regions for the years 2010-2019 to see if there is a significant difference in the unemployment rates and if we can conclude more from the trends.

Variable

Our topic is concerned with the sustainability goal of promoting economic growth by providing decent work and employment. The variable has been selected as the Total percentage of unemployed workers in the regions of North America, Asia, and Europe for each year from 2010-2019.

As stated by the world bank, unemployed people consist of individuals who do not have jobs, those who are seeking employment in the current years, those who are available to work, and those who lost their jobs. The people who aren't currently working but have secured future opportunities to work are also considered unemployed. The labor force of the population acts as the base for this indicator.

Our group chose this topic because it can be related to the sustainability goal of having Decent Jobs and Economic Growth. The unemployment rates of the population can help us indicate the productivity of the country and the number of jobs available in that country through the last 10

years. This indicator also gives us an insight as to how various countries across the globe are working towards reducing the unemployment of their population in the last 10 years.

Comparison of Countries Within Their Region

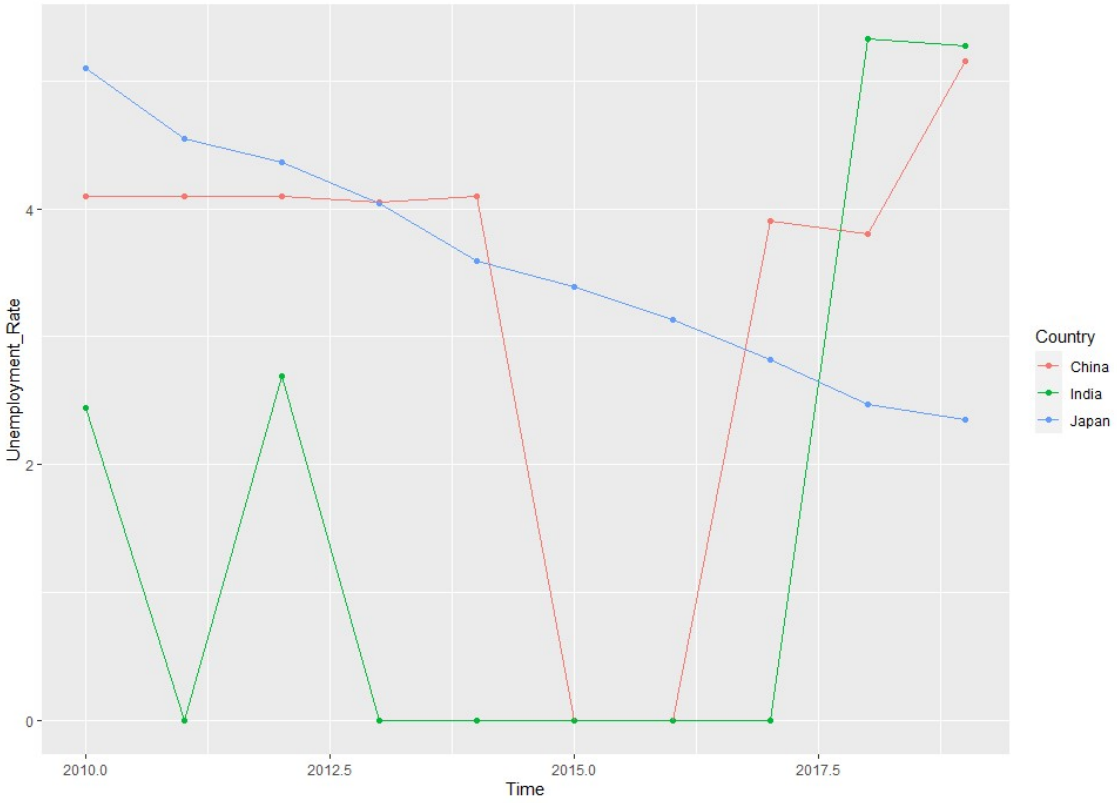
Problem Statement:

- Question Sustainability exploration for a particular world development indicator
- Series: Unemployment, total (% of total labor force) (national estimate)
- Time: 10 years (2010-2019)

Comparing the change in Unemployment rate for China, India and Japan from Asia

by, Pavana Teja Viswanath Josyabhatla

- **PLAN**
  - **a)** Question: Sustainability exploration for particular world development indicators
  - **b)** Region: Asia
    Countries: China, India, Japan
- **Graph**
  Below are the given scatter plots for the countries and the time



  - **a)** Question: Is there a difference in the trends between the three countries?
  - **b)** Comparing the trends between countries in Asia (China, India and Japan) for the rate of unemployment we can say that:

1) China and India's unemployment rates over the 10 years have considerable number of outliers within the plot. The relationship is not strong and positive.
2) China and India each have 3 possible outliers.
3) On the other hand, Japan has a decrease in the total unemployment rates over the years. The plot for Japan is Strong, Negative and linear with no considerable outliers.
4) Japan has a more linear decrease in the unemployment rates, where as China and India have spikes which means the change is not constant over the years.

- **Pearson's Correlation Coefficient**

```
> cor(ThreeCountriesSus$Japan, ThreeCountriesSus$Time)
[1] -0.9950739
```

- The coefficient of -0.9950739 indicates the relationship is negative and strong for Japan

- **Simple Linear Regression**

The regression model for Japan for the unemployment rate over 10 years

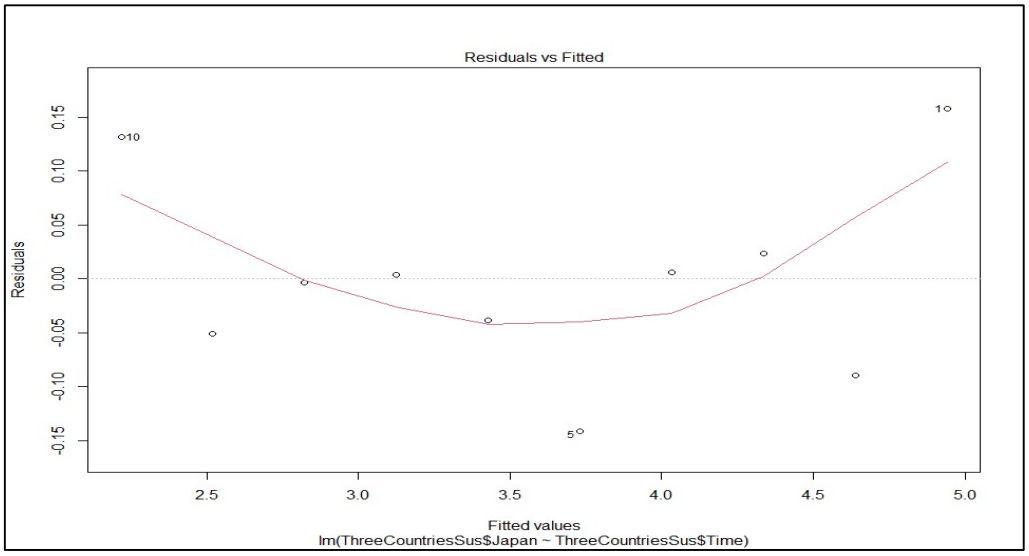
```
> sustainableModel <- lm(formula = ThreeCountriesSus$Japan ~ ThreeCountriesSus$Time)
> summary(sustainableModel)

Call:
lm(formula = ThreeCountriesSus$Japan ~ ThreeCountriesSus$Time)

Residuals:
      Min        1Q    Median        3Q       Max
-0.141333 -0.047667  0.000333  0.019000  0.158000

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             613.30202   21.47656   28.56 2.44e-09 ***
ThreeCountriesSus$Time   -0.30267    0.01066  -28.39 2.56e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09683 on 8 degrees of freedom
Multiple R-squared:  0.9902,    Adjusted R-squared:  0.9889
F-statistic:   806 on 1 and 8 DF,  p-value: 2.561e-09
```

1) Interpretation:

- Interpret the slope: The expected decrease in the unemployment rate is about 0.30267 per year.
- Interpret the y-intercept: Since we do not have a x-value that is 0 on the ggplot, we do not interpret the y-intercept. Value of y-intercept = 613.30202
- T test value for slope: -28.39
- P-Value: Almost 0

2) Assumptions:

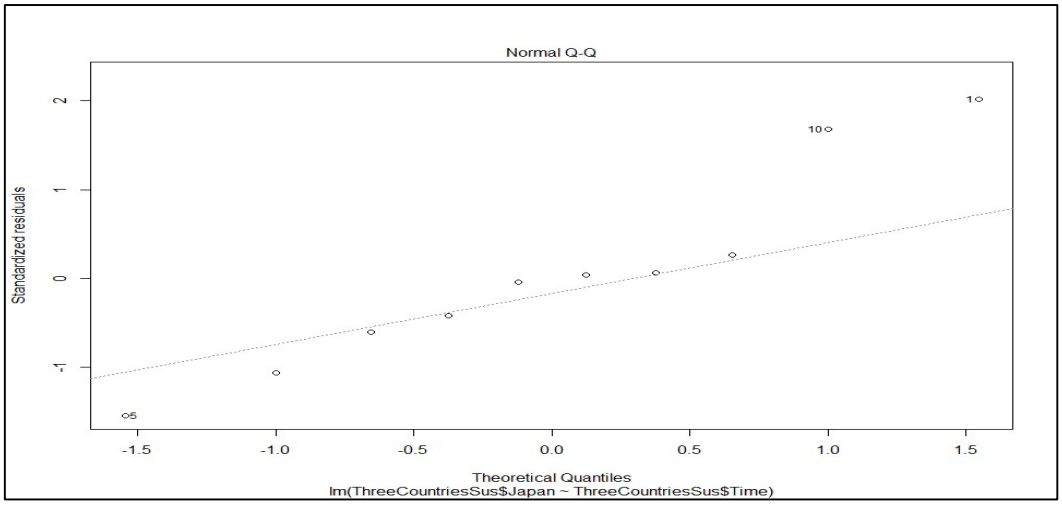
Residuals vs Fitted



- Linearity: No, there is a cone shaped pattern.
- Zero Mean: Yes, the residual mean is almost 0

```
> mean(SustainableJapanResids)
[1] -9.370217e-18
```

- Uniform Spread: We have a cone shape on the residual vs. fitted plot so no uniform spread
- Independence: Random Samples
- Randomness: Yes, the samples are random
- Normality: We can see some outliers otherwise its ok.



- **Confidence Intervals:**

1) Question: Find the confidence interval and the prediction interval for the next year after the data was collected

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           613.30202   21.47656   28.56 2.44e-09 ***
ThreeCountriesSus$Time  -0.30267    0.01066  -28.39 2.56e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09683 on 8 degrees of freedom
Multiple R-squared:  0.9902,    Adjusted R-squared:  0.9889
F-statistic:   806 on 1 and 8 DF,  p-value: 2.561e-09
```

- Hypothesis:
  Null Hypothesis: H_0:β_1=0
  Alternate Hypothesis: H_a: β_1 not equal to 0
- T value = -28.39
- P-Value is almost 0
- **Conclusion**: With a p value of almost 0 we have very strong evidence that the true value of unemployment rate for each year is different from zero.

2) Confidence interval for slope:

```
> confint(sustainableModel, level = 0.95)
                                2.5 %      97.5 %
(Intercept)             563.7769754 662.8270610
ThreeCountriesSus$Time  -0.3272509  -0.2780824
```

- We are 95% confidence that the change in unemployment rate per year is in between -0.327 and -0.278 percent.

3) Confidence Interval and Prediction Interval for average y

Prediction Interval

```
> predict.lm(susJaplm, latestx, interval = "prediction")
       fit      lwr      upr
1 1.915333 1.644906 2.18576
```

- We are 95% confident that 95% of the unemployment rate for Japan in the year 2020 is between 1.64% and 2.19%

Confidence Internal

```
> predict.lm(susJaplm, latestx, interval = "confidence")
       fit      lwr      upr
1 1.915333 1.762792 2.067875
```

- We are 95% confident that the average unemployment rate for the Japan in the year 2020 is between 1.76% and 2.07%

Comparing the change in Unemployment rate for United States of America, Canada and Mexico from North America

by, Satya Sai Koppalu

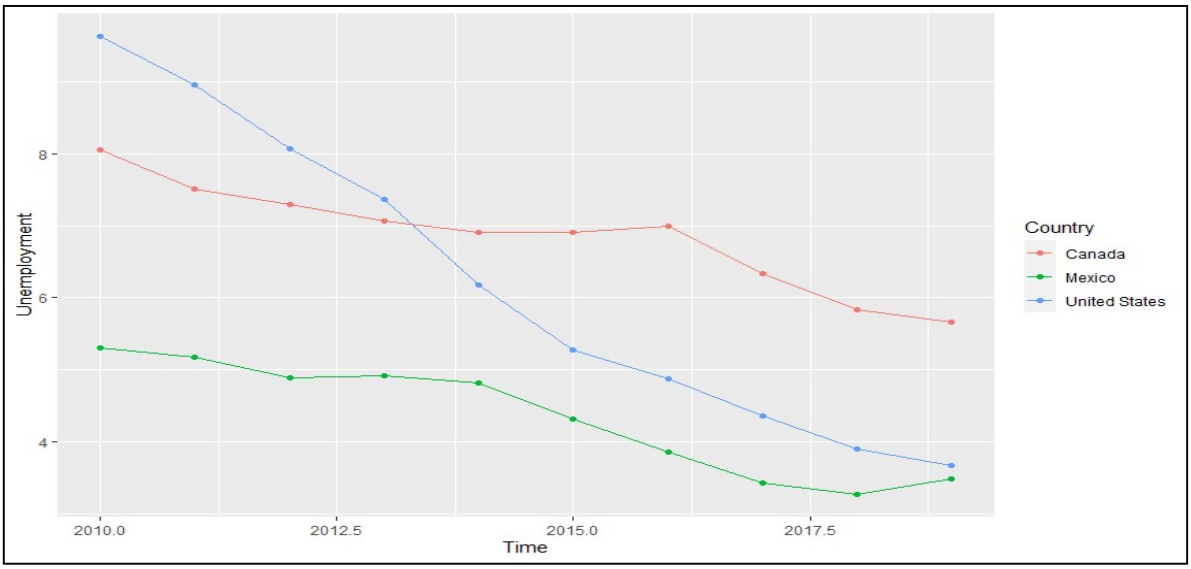
- **Plan:**

a) Region: North America

b) Countries: United States (USA), Canada, Mexico

- **Graph:**

Below are the given scatter plots for the countries and the time



a) Question: Is there a difference in the trends between the three countries?

Interpreting the graph above-

- Canada has have a negative, strong and linear relationship, with a single outlier.
- Mexico has have a negative, strong and linear relationship, with a single outlier.
- United States has have a negative, strong and linear relationship, with no possible outliers.
- Overall, all the three countries saw a decrease in the total unemployment rates over the ten years. Canada and Mexico saw a slight increase in unemployment rate in the middle of the decade although the change was corrected quickly and followed the previous trend.
- Out of the three countries, United States performed the best- it started with the highest unemployment rate and saw the biggest drop over the years.

Country chosen: United States (USA)

- **Pearson's Coeffecient:**

Finding the Pearson's Coeffecient

```
> ##finding the Pearson's coefficent
> cor(nacountries$USA, nacountries$Time)
[1] -0.9861495
```

Pearson's coefficient of -0.99 indicates that the relation for USA is negative and strong.

- **Simple Linear Regression**

```
> ##performing SLR
> usaobj <- lm(formula = nacountries$USA ~ nacountries$Time)
> summary(usaobj)

Call:
lm(formula = nacountries$USA ~ nacountries$Time)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59570 -0.25494  0.08779  0.21399  0.60473

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     1421.62686   84.16496   16.89 1.53e-07 ***
nacountries$Time   -0.70261    0.04178  -16.82 1.58e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3795 on 8 degrees of freedom
Multiple R-squared:  0.9725,    Adjusted R-squared:  0.9691
F-statistic: 282.8 on 1 and 8 DF,  p-value: 1.583e-07
```

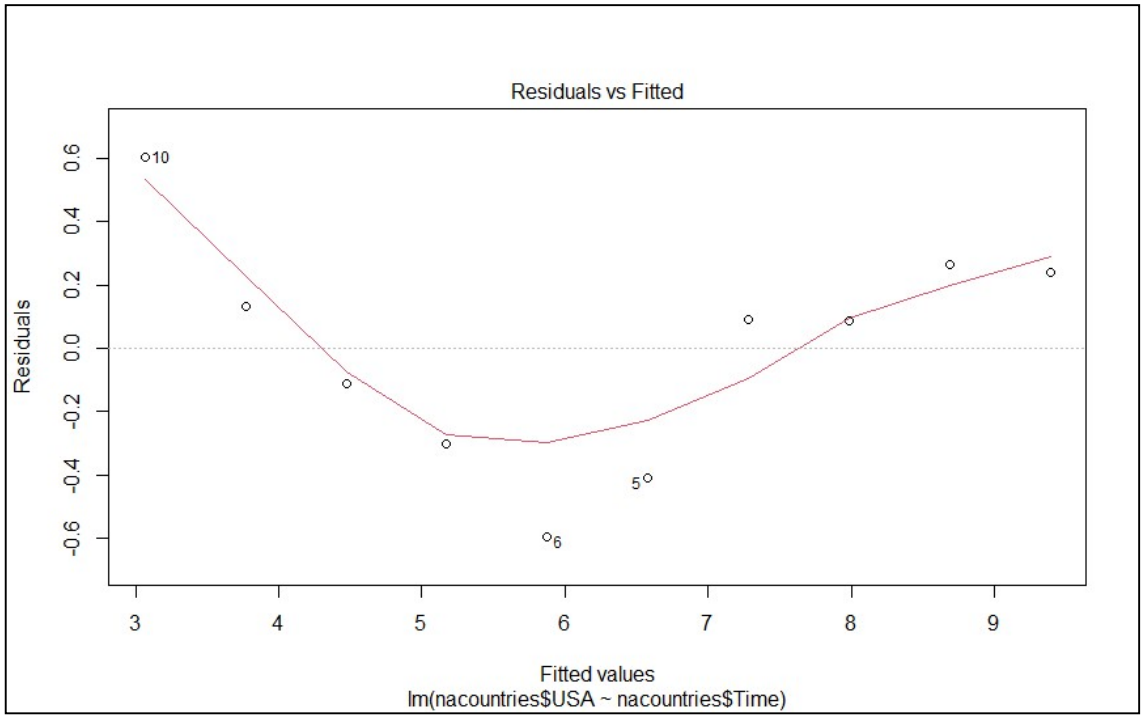
a) Interpretation of the above output:

Slope: The expected unemployment rate on y axis decreases with an average value of -0.7 for every year on x axis.

Y Intercept- Since we do not have a x-value that is equal to 0 we do not interpret the Y intercept with a value of 1421.63

- **Assumptions**

- Linearity- No linearity since there is a presence of a cone shaped pattern.

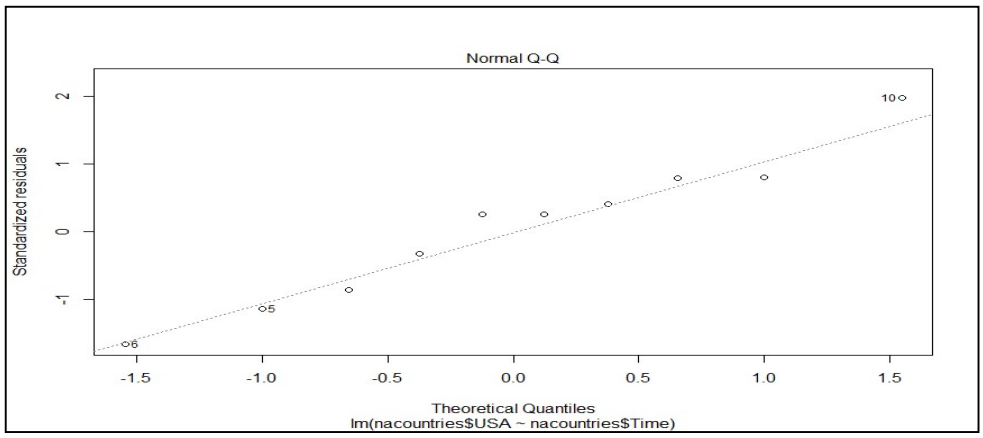


Residuals vs Fitted graph

- Zero Mean- Yes

```
> mean(usaresid)
[1] 1.943974e-17
```

- Uniform Spread- No uniform spread. Since the Residuals vs Fitted graph has a presence of a cone shape pattern and the data is scattered.
- Independence- Random Samples.
- Normality- Yes normality is present with a possibility of a single outlier.



- **T test for slope**

```
> ##performing SLR
> usaobj <- lm(formula = nacountries$USA ~ nacountries$Time)
> summary(usaobj)

Call:
lm(formula = nacountries$USA ~ nacountries$Time)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59570 -0.25494  0.08779  0.21399  0.60473

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     1421.62686   84.16496   16.89 1.53e-07 ***
nacountries$Time   -0.70261    0.04178  -16.82 1.58e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3795 on 8 degrees of freedom
Multiple R-squared:  0.9725,    Adjusted R-squared:  0.9691
F-statistic: 282.8 on 1 and 8 DF,  p-value: 1.583e-07
```

- Hypotheses:
  Null Hypothesis: H0: β1=0
  Alternate Hypothesis: Ha: β1 not equal to 0
- T test Value (ts): -16.82
- p value: Almost 0
- With a p value of almost 0 we have very strong evidence that the true value of unemployment rate for each year is different from 0.

- **Confidence Interval for slope**

```
> ##finding the confidence interval for the slope
> confint(usaobj, level = 0.95)
                          2.5 %       97.5 %
(Intercept)      1227.5421238 1615.7115971
nacountries$Time   -0.7989498   -0.6062623
```

  - We are 95% confident that the change in unemployment rate for each year is between -0.8 to 0.6 percent.

- **Confidence Interval for average Y**

```
> predicted = data.frame(x = 2020)
> predict.lm(tofind, predicted , interval = "confidence")
       fit      lwr      upr
1 2.362667 1.764869 2.960464
```

  - We are 95% confident that the average unemployment rate for the United States in the year 2020 is between 1.77% and 2.98%

- **Prediction Interval for Y**

```
> predict.lm(tofind, predicted , interval = "prediction")
       fit      lwr      upr
1 2.362667 1.302885 3.422448
```

  - We are 95% confident that 95% of the unemployment rate for the United States in the year 2020 is between 1.3% and 3.4%

Comparing the change in Unemployment rate for Spain, Germany and United Kingdom from Europe

by, Colin Glen Fernandes

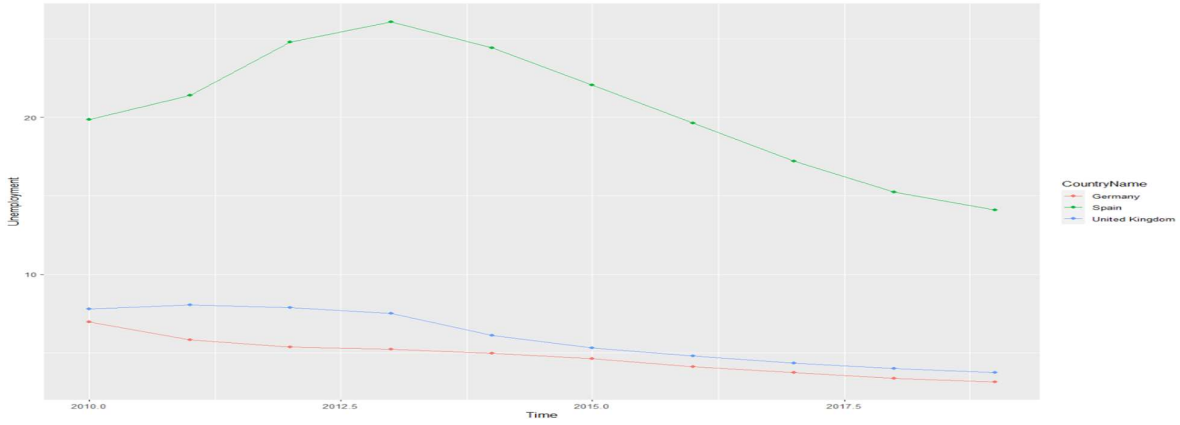
- **Plan:**

a) Region: Europe

b) Countries: United Kingdom, Spain and Germany

- **GG Plot**

  Below is the GG plot for the unemployment percentage of the three countries in relation to the time



a) The GG plot for Spain initially shows a positive increase but as time increases the unemployment value for Spain decreases. Therefore, we can say the relationship is strong and negative.

b) The GG plot for Germany shows a negative decrease as time increases the unemployment value in Germany decreases. Therefore, we can say the relationship is strong and negative.

c) The GG plot for the United Kingdom shows a negative decrease as time increases the unemployment value in the United Kingdom decreases. Therefore, we can say the relationship is strong and negative.

1) **Pearson's Correlation Coefficient**

```
> cor(P_Data_Extract_From_World_Development_Indicators$Spain, P_Data_Extract_From_World_Development_Indicators$Time)
[1] -0.6967011
```

Pearson's Correlation Coefficient for the country of Spain in relation to time is -697011. So here we can indicate that the relationship for Spain is Moderate and Negative

2) **Simple Linear Regression**

```
> ##To find the fitted line and also find the y-intercept and slope (Regression output)
> lmobject2<- lm(P_Data_Extract_From_World_Development_Indicators$Spain ~ P_Data_Extract_From_World_Development_Indicators$Time)
> summary(lmobject2)

Call:
lm(formula = P_Data_Extract_From_World_Development_Indicators$Spain ~
    P_Data_Extract_From_World_Development_Indicators$Time)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8349 -2.1195 -0.1825  2.0246  4.2024

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                   1905.5677   686.2426   2.777   0.0240 *
P_Data_Extract_From_World_Development_Indicators$Time   -0.9358     0.3407  -2.747   0.0252 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.094 on 8 degrees of freedom
Multiple R-squared:  0.4854,    Adjusted R-squared:  0.4211
F-statistic: 7.546 on 1 and 8 DF,  p-value: 0.02518
```

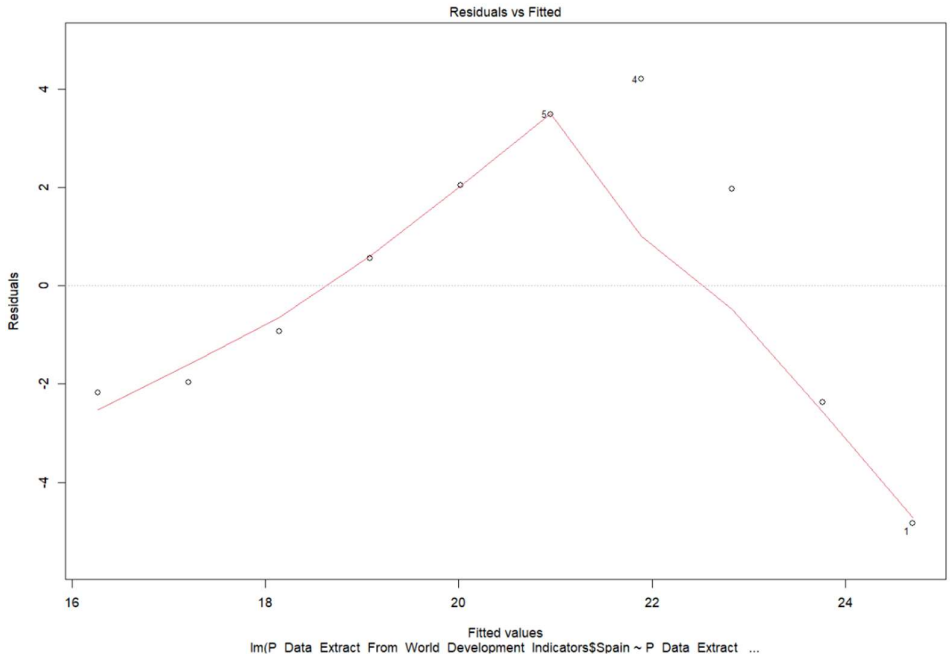
A) **The y-intercept is 1905.5677.**
   a. Since we do not have any plots near x = 0 on the graph we do not need to interpret the y intercept

B) **The slope is -0.9358**
   b. The slope of -0.9358 indicates the average change per point on the unemployment percentage to the increase of the time on the x-axis

C) **T test value:**
   c. For Intercept: 2.777
   d. For Slope: -2.747

D) **P-Value:**
   e. For Intercept: Almost Zero
   f. For Slope: Almost Zero

3) **Assumptions**: Residuals vs Fitted Plot

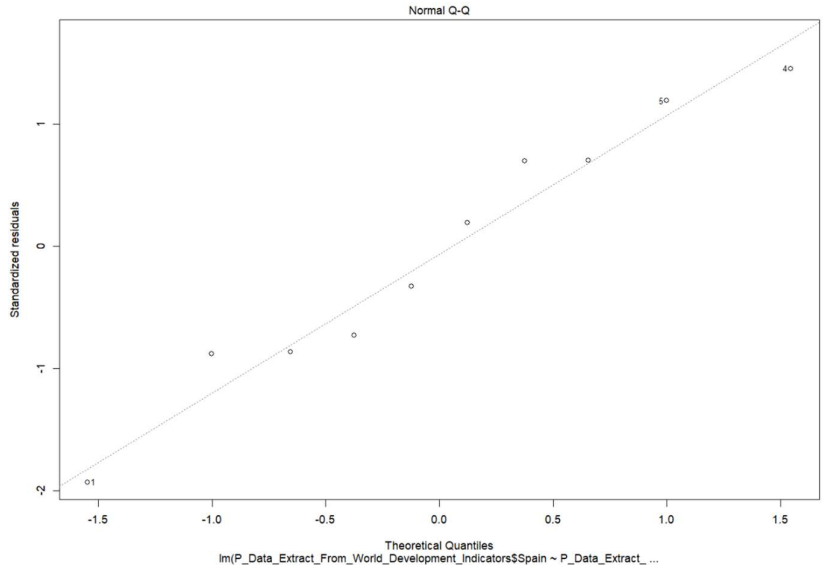


Assumptions:

- **Linearity**: There is no linearity. There is a cone shaped pattern
- **Zero Mean:** Yes, the residual mean is 0

```
> lmobjectres <- resid(lmobject2)
> mean(lmobjectres)
```

- `[1] 0`
- **Uniform Spread:** We have a cone-shaped pattern on the residual vs fitted plot so there is no uniform spread.
- **Independence:** Yes, the data has been chosen from the random samples
- **Randomness**: Yes, the samples are randomly chosen
- **Normality**:



4) Confidence Interval and prediction interval:

Hypothesis:
Null Hypothesis: H_0: $\beta_1=0$
Alternate Hypothesis: H_a: β_1 not equal to 0

With a p-value of almost 0, we have very strong evidence that the true value of unemployment rate for each year is different from zero.

a) Confidence Interval

```
> confint(lmobject2,level = 0.95)
                                                      2.5 %       97.5 %
(Intercept)                                         323.089416 3488.0460799
P_Data_Extract_From_World_Development_Indicators$Time  -1.721301   -0.1502145
```

- We are 95% confident that the change in unemployment rate per year is between -1.721 to -0.150 percent

a) Confidence Interval for average y and Prediction Interval for average y

Prediction Interval

```
> predict.lm(PredSpain, latestx, interval = "prediction")
       fit      lwr      upr
1 15.33733 6.696358 23.97831
```

- We are 95% confident that 95% of the unemployment rate for United States in the year 2020 is between 6.69% and 23.97%

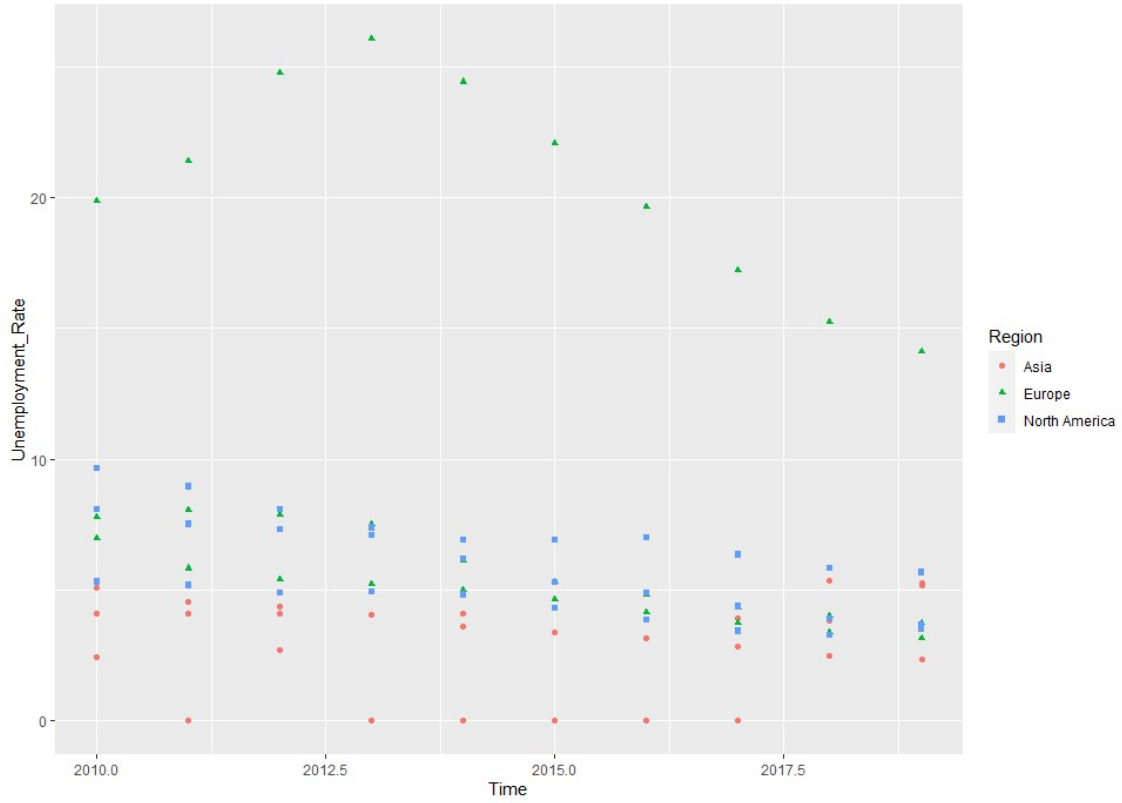
Confidence Interval

```
> predict.lm(PredSpain, latestx, interval = "confidence")
       fit      lwr     upr
1 15.33733 10.46317 20.2115
```

- We are 95% confident that the average unemployment rate for the United States in the year 2020 is between 10.46 % and 20.21%

Comparing the unemployment rate trends between three regions (Asia, Europe, North America)

- **GGplot**



1) Interpretation:
   - Europe has an increase in its unemployment rate in the first 3 years, following which there is a steep decrease following a steady trend till 2019. Its shows a negative trend with a moderately strong relationship.
   - North America shows a steady decline in its unemployment rate in the given period with no deviations. It shows a negative trend with a strong relationship.
   - Asia shows a steady decline in the initial years, following which there is a steep increase till 2019. Its shows a positive trend and a strong relationship
   - Asia performed the worst with an overall negative trend, whereas North America and Europe had an overall positive trend. Europe stands out as the best performer for reduction in unemployment rate among the chosen regions.

Code

Asian Region

```
#Importing tidyverse
library(tidyverse)
summary(UnemploymentData)

## plotting the scatter plot for 3 countries in Asia
ggplot(data = UnemploymentData) +
  geom_point(mapping = aes(x=Time, y = Unemployment_Rate, color = Country)) +
  geom_line(mapping = aes(x=Time, y = Unemployment_Rate, color = Country))

## Finding the Pearsons coefficiant
cor(ThreeCountriesSus$Japan, ThreeCountriesSus$Time)

## The SLR model
sustainableModel <- lm(formula = ThreeCountriesSus$Japan ~ ThreeCountriesSus$Time)
summary(sustainableModel)

#Residuals and Plot
SustainableJapanResids<-resid(sustainableModel)
plot(sustainableModel)

## Residual mean
mean(SustainableJapanResids)

## Confidence intervals for slope
confint(sustainableModel, level = 0.95)

## Prediction and Confidence intervals for y intercept
```

```
y <-ThreeCountriesSus$Japan

x <- ThreeCountriesSus$Time

susJaplm <- lm(y ~ x)

summary(susJaplm)

latestx = data.frame(x = 2020)

predict.lm(susJaplm, latestx, interval = "prediction")

predict.lm(susJaplm, latestx, interval = "confidence")
```

```
## Scatter plot to compare the regions

ggplot(data = RegionData) +

  geom_point(mapping = aes(x=Time, y = Unemployment_Rate, color = Region,
shape=Region, size=Region))
```

European Region

```
#invoking Tidyverse

library(tidyverse)
```

```
#Creating summary

summary(Data_Extract_World)
```

```
#Creating a ggplot for the data

ggplot(data = Data_Extract_World) +

  geom_point(mapping = aes(x=Time, y=Unemployment, color= CountryName)) +

  geom_line(mapping = aes(x=Time, y=Unemployment, color= CountryName))
```

```
##Pearsons Correlation coeficients for the selected country Spain

cor(P_Data_Extract_From_World_Development_Indicators$Spain,
P_Data_Extract_From_World_Development_Indicators$Time)
```

```
##To find the fitted line and also find the y-intercept and slope (Regression output)

lmobject2<- lm(P_Data_Extract_From_World_Development_Indicators$Spain ~
P_Data_Extract_From_World_Development_Indicators$Time)

summary(lmobject2)

##for assumptions use residual plot

plot(lmobject2)

## Finding mean

lmobjectres <- resid(lmobject2)

mean(lmobjectres)

##Confidence interval for slope

confint(lmobject2,level = 0.95)

##Finding the Prediction and Confidence values

x<-P_Data_Extract_From_World_Development_Indicators$Time

y<-P_Data_Extract_From_World_Development_Indicators$Spain

PredSpain <- lm(y~x)

summary(PredSpain)

latestx= data.frame(x=2020)

#Prediction

predict.lm(PredSpain, latestx, interval = "prediction")

#Confidence

predict.lm(PredSpain, latestx, interval = "confidence")
```

North American Region

```
#Importing tidyverse
library(tidyverse)

##Viewing the data
summary(nacountries)

##plotting graphs for threee countries
ggplot(data = nacountries ) +
  geom_point(mapping = aes(x=Time, y = Unemployment, color=Country))+
  geom_line(mapping = aes(x=Time, y = Unemployment, color=Country))

##finding the Pearson's coefficent
cor(nacountries$USA, nacountries$Time)

##performing SLR
usaobj <- lm(formula = nacountries$USA ~ nacountries$Time)
summary(usaobj)

##checking assumptions
usaresid<-resid(usaobj)
plot(usaobj)
mean(usaresid)

##finding the confidence interval for the slope
confint(usaobj, level = 0.95)

##finding confidence interval for avg y and prediction interval for y
y <-nacountries$USA
```

```
x <- nacountries$Time

tofind <- lm(y ~ x)

summary(tofind)

predicted = data.frame(x = 2020)

predict.lm(tofind, predicted , interval = "confidence")

predict.lm(tofind, predicted , interval = "prediction")
```

Sources

- https://databank.worldbank.org/source/world-development-indicators# (The World Bank DB)
- https://sdgs.un.org/goals/goal8 (United Nations Goals)