WHAT AFFECTS UNEMPLOYMENT?

A Statistical Analysis

Abstract

In this report, Multiple Regression and relevant statistical procedures are applied to build an appropriate model, attempting to explain unemployment.

GUO Siyue, ZHANG Yuejia, Liao Dongyu, DU Peng, ZHOU Yuhao

P-2E.AB

Table of Contents

INTRODUCTION

Nowadays, the world is facing the threat of COVID-19 as a whole. In order to prevent the spread of the epidemic, many shops are forced to stop business, and many companies have stopped production, thus, lose millions of dollars. In contrast, Chinese have returned to work in recent days due to the end of the epidemic. However, most companies, especially private companies, are now facing huge economic problems and have to lay off employees.

According to experts, the impact of the epidemic on the unemployment rate in the United States is also huge. They predict that the unemployment rate in the United States will rise to about 24%.

This is why we want to launch a research on the factors that influences the unemployment rate, not just during the COVID-19, but in a more common situation.

So, we try to come up with the ideas, “what can influence the unemployment rate?” Later, you can find our following analysis about the common factors in different countries.

We hypothesize that:

- Sex-ratio might influence the unemployment rate due to the discrimination
- For those who have the largest population in capital city might have lower unemployment rate
- The unemployment could be different because of the different geographic locations
- The unemployment rate could be influenced by the working condition (eg.

average working hours, average salary, etc.)

- GDP could also make a difference in unemployment rate

Database construction

As I mentioned before, that the first thing to try to find is the unemployment rate in 60 different countries. As we try to analyze whether the location influence the rate, our chosen countries include all five continents. Apart from the sex discrimination which may exist in some countries, we also consider about the people's attitudes towards their working environment.

We find our data mainly thanks to the OECD, and also some basic information like the language, the religion on ENCYCLOPÆDIA BRITANNICA. Some data are difficult to find, for example, the average working hours in some Africa areas, so we choose the statutory working hours instead.

The variables

- Population
- Inflation (GPI)
- GDP
- Sex ratio
- Education expenditure as a percentage of GDP
- Average lifetime
- Average salary
- Working hours per year
- Retirement age
- Gross pension wealth

We choose these variables because we expected that (for example):

- The smaller the sex ratio it has, we expected there is less sex discrimination in the work, thus, slighter the relationship is.

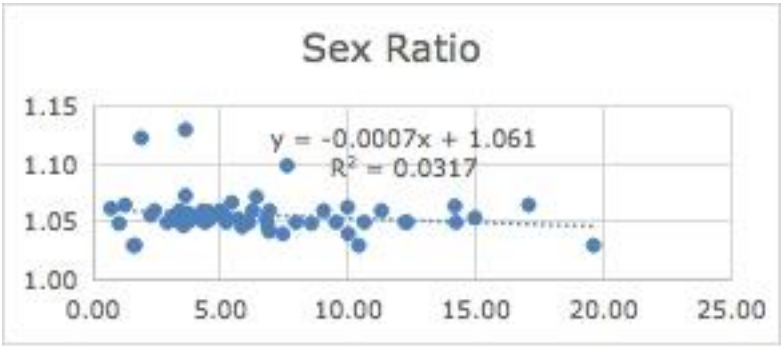
- The larger population it has, leads to a stronger competition in the work, thus a higher unemployment rate. Here, we expected a positive correlation.

- The lower the average salary it has, indicates a worse working welfare, thus, a higher unemployment rate as well. Here, we expected a negative correlation.

SIMPLE LINEAR REGRESSION MODEL

Our goal is to explain dependent variable, the rate of unemployment thanks to the independent variable, the ratio gender/sex whether is strongly correlated to citizens.

First of all, we generated a scatter plot on excel based on our database in order to have a first insight on what type of relationship could exist between sex ration (independent variable) and unemployment rate (dependent variable). At first sight, we noticed that there possibly has a linear relationship between the sex ratio and the ratio of unemployment.



After, we did the simple linear regression model and according to the result, we can get our simple polynomial regression equation which is:

$$\hat{Y} = 1.061 - 0.0007X$$

- $\hat{Y} = estimated\ the\ ratio\ of\ unemployment$
- $X = ratio\ of\ sex$
- $B0 = 1.061\ estimate\ of\ the\ regression\ intercept$
- $B1 = -0.0007\ estimate\ of\ the\ regression\ slope$

B0 is the estimated average value of $\hat{Y}$ when the value of X is zero.
B1 means that the ratio of sex will increase on average by -0.0007 at each 1 unit increase in the ratio of unemployment.

Coefficient of determination r^2

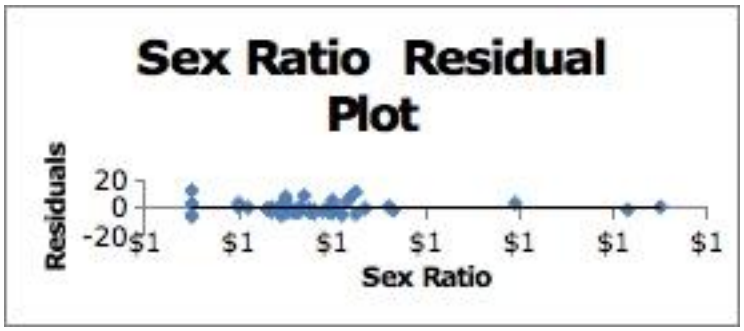
$r^2 = \frac{SSR}{SST} = 0.0317$

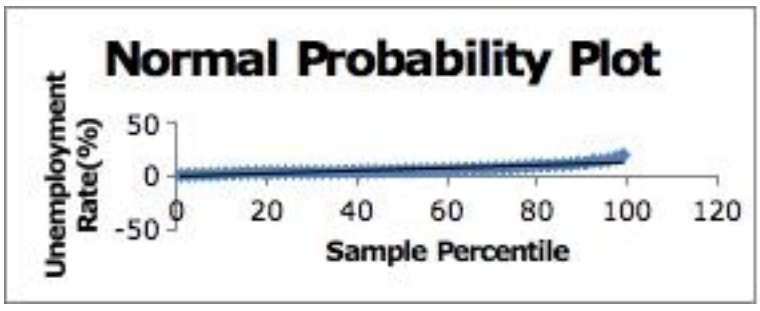
Regression Statistics	
Multiple R	0.1780977
R Square	0.0317188
Adjusted R S	0.0150243
Standard Err	4.088003
Observation	60

According the result that we get, R square is equal to 0.0317 which means that independent variable explains 3.2% of our regression model, so 3.2% employment rate can be explained by the ratio of sex. We expected the R square the higher the better but because $0<r^2<1$, and 3.2% close to 0, far from 1. Therefore, we do not have a good model.

Assumptions of regression: The L.I.N.E

We are now going to check those assumptions by performing a residual analysis.





As we can observe on the residual analysis plot, all residuals are above or below 0 which shows a linear relationship between X and Y. The errors are independent as each point on the graphs does not seem to depend on the position of others and we cannot observe a trend or shape. Meanwhile, our residuals are almost normally distributed, and we do not notice any heteroskedasticity because of constant variance.

Confidence Intervals analysis

Coefficients	tandard Erro	t Stat	P-value	Lower 95%	Upper 95%	Lower 95,0%	Upper 95,0%
51.681908	32.92547	1.5696635	0.1219334	-14.22558	117.5894	-14.22558	117.5894
-42.95962	31.166563	-1.378388	0.1733775	-105.3463	19.427039	-105.3463	19.427039

As we can see from the chart above, the confidence interval is from negative to positive which include 0. As a result, there is no significant relationship between ratio of sex and ratio of employment at the 0.05 level of significance.

Mean	1.0563
SD	0.017076
XI	1.87
Predicted Y	-28.6526
T	2.001717
SE	25.36572
ME	50.77501
Upper	22.12244
Lower	-79.4276

Also, suppose a sex ratio of 1.87, we can be 95% confident that the average unemployment rate is between -79.43 and 22.12.

DUMMY VARIABLE ANALYSIS

In this case, we do not only consider the influence of quantitative variables but also include some qualitative variables. Here, we plan to use dummy variables to add qualitative variables in our regression analysis.

First dummy variable analysis

First, we did an analysis with 2 categorical dummy variables. The independent variable is whether the capital city has the largest population, expressing in yes and no. Meanwhile, we set "yes" as dummy variable which equals to 1. Besides, we use unemployment rate as dependent variable. Here are our results **(Appendix Table 1)**.

Under zero hypothesis (H_0): There is no significant relationship between X_1 and Y ($\beta_1 = 0$) Under alternative hypothesis (H_1): There is a significant relationship between X_1 and Y ($\beta_1 \neq 0$) The output indicates that the p-value equals to 0.79, which is larger

than 5 %, so we decided to reject H1. There is no significant relationship between unemployment rate and the population of the capital city.

Second dummy variable analysis

Second, we did an analysis with 6 categorical dummy variables. In this regression analysis, we still use unemployment rate as dependent variable. Besides, we define the following dummies:

X1=1, if continent is Europe, 0 otherwise

X2=1, if continent is Oceania, 0 otherwise

X3=1, if continent is Asia, 0 otherwise

X4=1, if continent is North America, 0 otherwise

X5=1, if continent is Africa, 0 otherwise

The output **(Appendix Table 2)** shows that the p-value of Asia equals to 0.04, which is the only one smaller than 5 %, so there is a significant relationship between unemployment rate and Asia.

Therefore, the multiple regression equation is:

$$Unemployment\ Rate = 7.63 - 3.95(Asia)$$

On average, unemployment rate was 3.59% less in Asia than continents which is not Asia, given the same GDP, average salary, working hours and gross pension wealth.

Interaction Effect

Furthermore, we hypothesize interaction between pairs of X variables. As response to one X variable may vary at different levels of another X variable, we can expect variables to interact to change the unemployment rate. Besides, we defined the following independent variables and dependent variables:

Y= unemployment rate

X1=Average Salary

X2=Asia

X3= Average Salary*Asia

The result **(Appendix Table 3)** indicates that the P-value of average salary is 4.32%（<5%), that of dummy variable, Asia, is 0.3% (<5%). However, the P-value of X3 which equals to average salary multiply Asia is 38% (>5%).

Therefore, our multiple regression equation is:

$$Unemployment\ Rate = 8.62 - 5.4 \times 10^{-5}(X1) - 4.47(X2)$$

As a result, whether without interaction term or with interaction term, the effect of average salary(X1) is always measured by β1.

MULTIPLE REGRESSION MODEL

Firstly, we run a multiple regression analysis with all the independent variables which are population, inflation, GDP, sex ratio, education expenditure, average lifetime, average salary, retainment age, working hours and gross pension wealth, and using unemployment rate as our dependent variable. However, we are not satisfied with the result.

Regression Statistics	
Multiple R	0.489033
R Square	0.239153
Adjusted R Square	0.083879
Standard Error	3.942529
Observations	60

	df	*SS*	*MS*	*F*	*Significance F*
Regression	10	239.4007	23.94007	1.540195	0.153825877
Residual	49	761.6334	15.54354		
Total	59	1001.034			

As we can see, R^2 is equal to 0.239 which means only 23.9% of the variation in unemployment rate is explained by the variation in variables mentioned above. Adjusted R^2 is equal to 0.084 so that only 8.4% of the variation in unemployment rate is explained by the variation in variables mentioned above, taking into account the sample size and number of independent variables. These are quite low and means only a few of variation in unemployment rate can be explained by the collecting variables.

Moreover, the overall p-value is 0.15 which is more than 5% so there is no linear relationship between dependent variable and independent variables.

Based on the above analysis, we delate some independent variables which are not significant to the model.

New multiple regression model

After the adjustment, we only maintain 4 independent variables which are GDP, average salary, working hours (per year) and gross pension wealth.

As the chart shows **(Appendix Table 4),** R^2 is equal to 0.223 which is still quite low, but the adjusted R^2 increased a little bit compare to the previous one which is equal to 0.166. The overall p-value declined a lot from 0.15 to 0.0069 which shows a linear

relationship between dependent variable and independent variables. Meanwhile, the standard error decreased to 3.76. Compare to the previous one, although it doesn't change a lot which only decrease 0.18, it's a good result to become smaller.

According to the Excel, our multiple regression equation is

$$\begin{aligned} Unemployment\ Rate \\ = 11.9171 - 1 \times 10^{-13}(GDP) - 5.6 \times 10^{-5}(Average\ Salary) \\ - 0.0046(Working\ hours) + 0.4122(Gross\ Pension\ Wealth) \end{aligned}$$

The estimated intercept (Y) which is unemployment rate, is equal to 11.92%. This is the unemployment rate a country would have had if all the independent variables are equal to 0.

L.I.N.E Assumption

L.I.N.E assumption is important for us to understand the quality of each variable in the model. **(Appendix Table 5)**

Linearity: As we can see from the four residual plots, the residuals are randomly distributed above and below 0, so the relationship between x and y is linear.

Independence of errors: We cannot see a shape or a trend in four residual plots so we can say that residuals are following the independence of errors.

Normality of Error: According to the normal probability plot, there are some dots are not close to the regression line so that the distribution of residuals is not normal.

Equal Variance: Through the observation of the residual plots, we can find that the variance of GDP and average salary are not equal.

Individual Variables Effect

To further study the relationship between dependent variable and each independent variable, we performed t-test regarding each independent variable and looked at the p-value **(Appendix Table 4)**. We found that except GDP, all the other independent variables have a p-value lower than 5% which means there is a linear relationship between them and unemployment rate. As we can observe from the table, the working hours has a smallest p-value (0.009) which shows a significant impact to unemployment rate. The p-value of GDP is equal to 0.409 which is much higher than 5% so we can say that GDP doesn't have a huge impact on unemployment rate.

Then, looking at the confidence interval **(Appendix Table 4)**, we can also find that the CI of GDP is from negative to positive which contains 0. Based on this result we can confirm that GDP doesn't have a significant influence on unemployment rate. On the contrary, the confidence interval of other 3 variables show a good result because they don't include 0. For instance, we are 95% confident that with an increase of

gross pension wealth, the unemployment rate will increase between 0.0971 and 0.7272 percentage.

Variance Inflationary Factor (VIF) analysis

To test whether there is a correlation among two or more independent variables, we use VIF to measure collinearity. If the VIF is higher than 5, that variable is highly correlated with the other independent variables which shows a redundant information, so that we should only accept independent variable with VIF smaller than 5.

Variable	**VIF Value**
GDP	1.2
Average Salary	1.15
Working hours	1.16
Gross Pension Wealth	1.19

According to our calculation, all the VIFs are lower than 5 and they are all around 1 which is quite good because this means our independent variables do not contribute redundant information to the multiple regression model.

COMBINED MULTIPLE REGRESSION MODEL

If all elements are considered together, we can produce the following model **(Appendix Table 7)**:

$$\begin{aligned} Unemployment\ Rate \\ &= 11.4358 - 0.0033(hour) - 1.1 \times 10^{-13}(GDP) \\ &- 6.2 \times 10^{-5}(salary) + 0.3274(pension) - 3.4204(dummy\ Asia) \\ &+ 0.1391(interaction\ Asia\ Salary) \end{aligned}$$

From the R squared value of 0.27, more variation in unemployment is explained by all the considered elements. These variables together show significance (p value=0.006<0.5) in explaining unemployment.

Among all the variables, we can see that the Average Salary has the best quality in predicting Unemployment (p value<0.05).

From the coefficients, we can observe that the dummy variable of Asia produces significant contribution to the unemployment rate while the interaction between Asia and salary actually produces minimal effect with a p-value approaching 1.

CONCLUSION

To conclude what we have analyzed, we can find something out of expectations. For example, in our first model, we find that there is no relationship between the sex ratio and the unemployment rate. It would be interesting to find this, because actually in some Asian countries like Japan and Korean, when company has to lay off employees, they will first consider the females, which is actually exists. According to the graph, we come up with the conclusion that sex ratio does have influence, but just in some particular countries. So, it wouldn't be regarded as a common factor.

However, we can also find something that reach our expectations. For example, unemployment rate is 3.59% less in Asia than continents which is not Asia, when other factors keep the same. And we are glad to find that gross pension wealth actually has a positive correlation with the unemployment.

In conclusion, there might be some personal ideas when choosing the variables as the factor only influence in particular areas. It would be preferable to take into account more common factors.

REFERENCE

Gross pension wealth. (n.d.). [Text]. Retrieved 27 April 2020, from https://www.oecd-ilibrary.org/finance-and-investment/gross-pension-wealth/indicator/english_62cdd9d3-en

UNdata | record view | GDP, PPP (current international $). (n.d.). Retrieved 27 April 2020, from http://data.un.org/Data.aspx?q=GDP&d=WDI&f=Indicator_Code%3aNY.GDP.MKTP.PP.CD

UNdata | record view | Sex ratio at birth (male births per female births). (n.d.). Retrieved 27 April 2020, from http://data.un.org/Data.aspx?q=sex+ratio&d=PopDiv&f=variableID%3a52

Field Listing: Education expenditures—The World Factbook—Central Intelligence Agency. (n.d.). Retrieved 27 April 2020, from https://www.cia.gov/library/publications/resources/the-world-factbook/fields/369.html

List of Countries by Continent—StatisticsTimes.com. (n.d.). Retrieved 27 April 2020, from http://statisticstimes.com/geography/countries-by-continents.php?fbclid=IwAR1mrdbiIKcqSL_i3UsbOCPMeOc4VrgpKQxGQPezfiOqEXW5caMa1zd8qT4

International Indicators: Population mid-2019 - PRB. (n.d.). Retrieved 28 April 2020, from https://www.prb.org/international/

Unemployment—Unemployment rate—OECD Data. (n.d.). TheOECD. Retrieved 28 April 2020, from http://data.oecd.org/unemp/unemployment-rate.htm

Prices—Inflation (CPI)—OECD Data. (n.d.). TheOECD. Retrieved 28 April 2020, from http://data.oecd.org/price/inflation-cpi.htm

APPENDIX

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.033786
R Square	0.001142
Adjusted R Square	-0.01608
Standard Error	4.152049
Observations	60

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.142701	1.142701	0.066284	0.797738
Residual	58	999.8914	17.23951		
Total	59	1001.034			

	Coefficients	*Standard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	6.1085	0.928426	6.579413	1.49E-08	4.250053	7.966947	4.250053	7.966947
Dummy(yes)	0.29275	1.137085	0.257456	0.797738	-1.98337	2.568874	-1.98337	2.568874

Table 1 (first dummy variable analysis)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.425478
R Square	0.181032
Adjusted R Square	0.105201
Standard Error	3.896378
Observations	60

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	181.219	36.2438	2.387325	0.04989813
Residual	54	819.8151	15.18176		
Total	59	1001.034			

	Coefficients	*Standard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	7.631	1.472692	5.181666	3.36E-06	4.67842938	10.58357	4.678429	10.58357
dummy(Europe)	-0.92783	1.681933	-0.55164	0.583469	-4.29989937	2.444247	-4.2999	2.444247
dummy(Oceania)	-3.041	3.124052	-0.97342	0.334687	-9.30434813	3.222348	-9.30435	3.222348
dummy(Asia)	-3.59059	1.749819	-2.05198	0.045034	-7.09876489	-0.08241	-7.09876	-0.08241
dummy(North Ameri	-1.08525	2.442184	-0.44438	0.658546	-5.98153446	3.811034	-5.98153	3.811034
dummy(Africa)	1.880429	2.082702	0.902879	0.3706	-2.29513685	6.055994	-2.29514	6.055994

Table 2 (second dummy variable analysis)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.428949
R Square	0.183997
Adjusted R Square	0.140282
Standard Error	3.819235
Observations	60

ANOVA

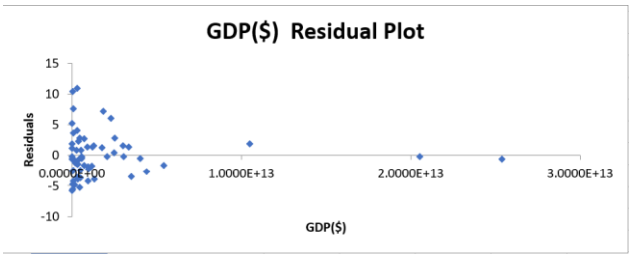
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	184.1872	61.39572	4.209063	0.009368
Residual	56	816.847	14.58655		
Total	59	1001.034			

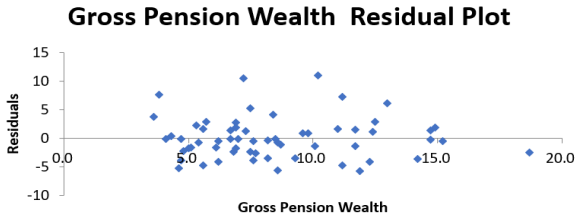
	Coefficients	*Standard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	8.620641	0.901273	9.564961	2.23E-13	6.815174	10.42611	6.815174	10.42611
Average Salary	-5.4E-05	2.62E-05	-2.06775	0.043296	-0.00011	-1.7E-06	-0.00011	-1.7E-06
dummy(Asia)	-4.47606	1.474056	-3.03656	0.003629	-7.42895	-1.52317	-7.42895	-1.52317
Interaction Asia*Average	4.73E-05	5.36E-05	0.883308	0.380846	-6E-05	0.000155	-6E-05	0.000155

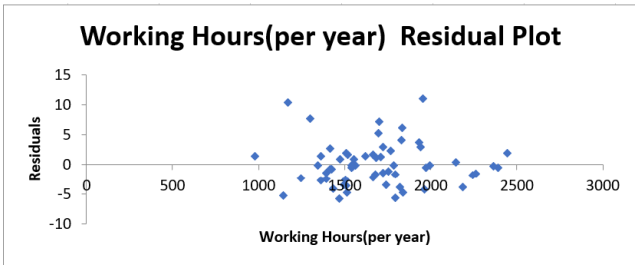
Table 3 (interaction effect)

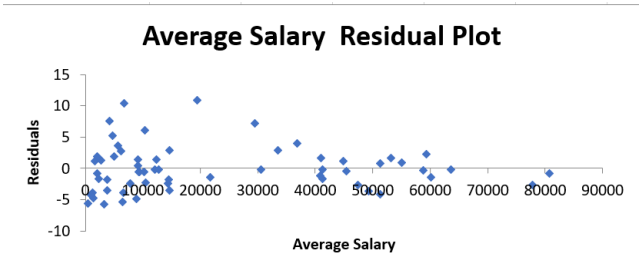
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.472231							
R Square	0.223002							
Adjusted R Square	0.166493							
Standard Error	3.760565							
Observations	60							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	4	223.2324	55.8081	3.946308	0.006915051			
Residual	55	777.8017	14.14185					
Total	59	1001.034						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
Intercept	11.91706	3.018231	3.94836	0.000226	5.868390851	17.96573	5.868391	17.96573
GDP($)	-1E-13	1.24E-13	-0.8312	0.409452	-3.51009E-13	1.45E-13	-3.5E-13	1.45E-13
Average Salary	-5.6E-05	2.35E-05	-2.3672	0.021469	-0.000102916	-8.6E-06	-0.0001	-8.6E-06
Working Hours(per year)	-0.00455	0.001685	-2.69888	0.009223	-0.007925065	-0.00117	-0.00793	-0.00117
Gross Pension Wealth	0.412163	0.157196	2.621973	0.011286	0.097135685	0.72719	0.097136	0.72719

Table 4 (multiple regression analysis)









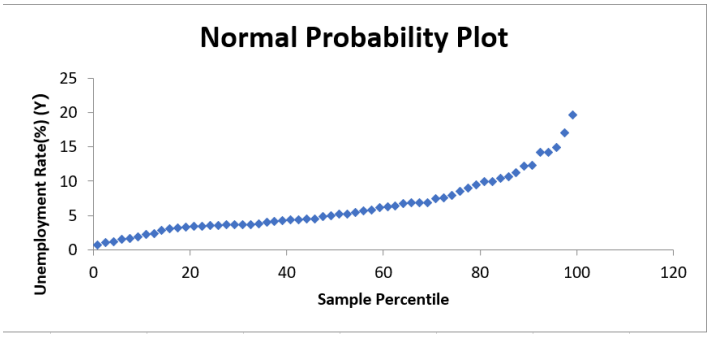


Table 5 (residual plots & normal probobility plot)

Regression Statistics				
Multiple R	0.408426			
R Square	0.166812		VIF(GDP)	1.20021
Adjusted R Square	0.122177			
Standard Error	4.06E+12			
Observations	60			

Regression Statistics				
Multiple R	0.364237			
R Square	0.132668		VIF(average saalary	1.152962
Adjusted R Square	0.086204			
Standard Error	21344.3			
Observations	60			

Regression Statistics				
Multiple R	0.374805			
R Square	0.140479		VIF(working hours)	1.163439
Adjusted R Square	0.094433			
Standard Error	298.2113			
Observations	60			

Regression Statistics				
Multiple R	0.397182			
R Square	0.157754		VIF(gross pension wealth)	1.187301
Adjusted R Square	0.112633			
Standard Error	3.196821			
Observations	60			

Table 6 (VIF result)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.527838
R Square	0.278613
Adjusted R Square	0.196946
Standard Error	3.691227
Observations	60

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	6	278.9007	46.48345	3.411589	0.006394
Residual	53	722.1334	13.62516		
Total	59	1001.034			

	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	11.43575	3.555917	3.215979	0.002217	4.303486	18.56802	4.303486	18.56802
Working Hours(per year)	-0.00334	0.001909	-1.75071	0.085782	-0.00717	0.000487	-0.00717	0.000487
GDP($)	-1.1E-13	1.35E-13	-0.81502	0.41871	-3.8E-13	1.6E-13	-3.8E-13	1.6E-13
Average Salary	-6.2E-05	2.34E-05	-2.65441	0.010466	-0.00011	-1.5E-05	-0.00011	-1.5E-05
Gross Pension Wealth	0.327422	0.177974	1.839721	0.071413	-0.02955	0.684392	-0.02955	0.684392
dummy(Asia)	-3.42037	2.976853	-1.14899	0.255719	-9.39118	2.550437	-9.39118	2.550437
Interaction Asia*Salary	0.139054	0.371476	0.374329	0.709653	-0.60603	0.884141	-0.60603	0.884141

Chart Area

Table 7 (combined multiple regression)