VIET NAM NATIONAL UNIVERSITY

HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY

-------o0o-------

PROJECT REPORT

PROBABILITY AND STATISTICS

(T-test – Multiple linear regression – Pearson's correlation coefficient)

Instructor: PhD. Nguyễn Tiến Dũng

Members:

	Name:	Student ID
1.	Trần Tiểu Bình	1752010
2.	Nguyễn Đức Minh	1752034
3.	Trần Gia Quốc Bảo	1710611
4.	Vũ Xuân An	1710445

June 20, 2020

Table of content

1. The description of the data

1.1. Data

1.1.1. Air pollution in Vietnam and Philippines

The table below indicates how many people died by air pollution among two ASEAN countries are Vietnam and Philippines by two factors indoor and outdoor.

Table 1.1: Number of deaths by air pollution in Philippines and Vietnam from 1990 to 2017

Year	Philippines		Vietnam	
	Indoor air pollution (deaths) (1)	**Outdoor air pollution** (deaths) (2)	**Indoor air pollution** (deaths) (3)	**Outdoor air pollution** (deaths) (4)
1990	29469.1513	11311.38	33595.18165	10635.24
1991	28168.25213	11446.75	33210.52945	10855.21
1992	28048.5887	11746.60	32673.3821	11154.82
1993	26610.32954	11769.58	32189.42073	11407.13
1994	25835.4167	11909.94	31500.74543	11736.37
1995	25599.15098	12310.74	30848.22456	12038.83
1996	25000.10439	12558.01	30225.88759	12456.92
1997	24913.83927	13000.48	29844.38126	12893.41
1998	24883.96326	13375.36	29513.17708	13456.23
1999	25097.28192	13714.29	29430.78796	14064.83
2000	25676.01743	14221.93	29539.65206	14663.06
2001	26740.45729	14930.00	29384.36383	15570.95
2002	27672.6466	15485.04	29231.4519	16356.56
2003	28694.61624	15884.78	28904.95218	17288.27
2004	30061.11252	16421.96	28770.30129	18094.41
2005	31906.84217	17124.96	28130.05495	18975.7
2006	33472.95888	17835.65	27618.65539	20052.49

2007	34860.35902	18473.23	26751.38288	21259.62
2008	36783.80853	19435.27	26069.05961	22372.28
2009	38765.89959	20313.10	25402.52216	23447.26
2010	40094.0478	21028.55	24310.88284	24595.28
2011	41522.34886	21273.60	23683.83174	25482.83
2012	41772.43564	22372.07	22980.68621	26288.69
2013	42347.97254	22215.84	22258.51688	27516.07
2014	42736.90032	22206.59	21956.85479	28077.13
2015	43207.61197	22722.93	21773.23976	28207.42
2016	42603.98461	22242.41	21783.12759	28431.96
2017	42000.06841	22638.53	21274.04751	29549.67

1.2. GPD, working hours and unemployment rate

The table below illustrates the relation among working hours (hours per person engaged), unemployed (rates) and GDP (dollars per capita). In this table we use data from 2000 to 2017.

Table 1.2. GDP, working hours and unemployment rate in Vietnam from 1990 to 2018

Year	**Working hours** (hours per person engaged)	**GDP per capita** (int.-$) (constant 2011 international $)	**Unemployment rate** (%)
1990	2643.469	1452.877479	N/A
1991	2644.7563	1507.191728	N/A
1992	2646.5408	1603.903822	N/A
1993	2647.239	1699.222153	N/A
1994	2648.5125	1815.275905	N/A
1995	2648.729	1954.776754	N/A
1996	2648.8669	2104.505112	N/A
1997	2317.3438	2244.310844	N/A
1998	2362.4314	2343.440239	N/A
1999	2362.1404	2426.282479	N/A

2000	**2393.6313**	**2562.104865**	**2.33766**
2001	**2246.991**	**2692.125097**	**2.53165**
2002	**2230.7598**	**2833.770963**	**2.22772**
2003	**2254.8853**	**3000.311156**	**2.17391**
2004	**2207.4919**	**3196.297215**	**2.11765**
2005	**2191.7068**	**3405.679195**	**4.73296**
2006	**2240.8411**	**3609.683037**	**4.88424**
2007	**2241.1814**	**3831.243217**	**4.13978**
2008	**2336.6824**	**4009.959321**	**3.62748**
2009	**2405.3545**	**4185.019792**	**2.60939**
2010	**2300.02**	**4408.168612**	**2.66792**
2011	**2331.2832**	**4632.765965**	**2.03578**
2012	**2310.541**	**4821.137231**	**1.77658**
2013	**2267.4883**	**5024.438902**	**1.70906**
2014	**2150.6357**	**5264.8281**	**1.8622**
2015	**2169.5916**	**5554.858056**	**2.11189**
2016	**2169.5916**	**5837.628704**	**3.02**
2017	**2169.5916**	**6171.884192**	**2.0073**
2018	N/A	N/A	1.995871198

2. The aim of this project:

We use some methods to analyze the data and to answer these questions:

- *Is Vietnam more polluted than Philippines?*
- *The relation among working hours (hours), unemployed (rates) and GDP (dollars per capita)*

3. Statistical methods

3.1. T-test

3.1.1. Definition

A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features.

It is mostly used when the data sets, like the data set recorded as the outcome from flipping a coin 100 times, would follow a normal distribution and may have unknown variances. A t-test is used as a hypothesis testing tool, which allows testing of an assumption applicable to a population. A t-test looks at the t-statistic, the t-distribution values, and the degrees of freedom to determine the probability of difference between two sets of data. To conduct a test with three or more variables, one must use an analysis of variance.

3.1.2. Types of T-test

- **1-Sample t**: Tests whether the mean of a single population is equal to a target value
- **2-Sample t**: Tests whether the difference between the means of two independent populations is equal to a target value
- **Independent samples t-test**: Equal variance and unequal variance (Welch's test)
- **Paired t:** Tests whether the mean of the differences between dependent or paired observations is equal to a target value
- **T-Test in regression output**: Tests whether the values of coefficients in the regression equation differ significantly from zero.

3.1.3. Formula

❖ **General:**

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}}$$

Which t: a Student t quantile with n-1 degrees of freedom

$\bar{x}$: the sample mean

μ: the specified population mean s^2: the sample variance n: the sample size

❖ **One sample t-test:**

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

To test the null hypothesis that the population mean is equal to a specified value μ_0 Degrees of freedom: $n - 1$

❖ **Independent sample t-test:**

- **Equal variance:**

  - Equal sample sizes

$$t = \frac{\overline{X_1} - \overline{X_2}}{s_p\sqrt{\frac{2}{n}}}$$

$$s_p = \sqrt{\frac{s_{x_1}^2 + s_{x_2}^2}{2}}$$

  - Degree of freedom: $2n - 1$

  - Equal or unequal sample sizes

$$t = \frac{\overline{X_1} - \overline{X_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_{x_1}^2 + (n_2 - 1)\, s_{x_2}^2}{n_1 + n_2 - 2}}$$

  - Degree of freedom: $n_1 + n_2 - 1$

- **Unequal variance:**

  - Equal or unequal sample sizes

$$t = \frac{\overline{X_1} - \overline{X_2}}{s_{\bar{\Delta}}}$$

$$s_\Delta = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

  - Welch's t-test:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

❖ **Paired t-test:**

$$t = \frac{\overline{X_D} - \mu_0}{\frac{s_D}{\sqrt{n}}}$$

- Used when the samples are dependent

- Degree of freedom: $n - 1$

3.1.4. Applications:

A one-sample location test of whether the mean of a population has a value specified in a null hypothesis.

A two-sample location test of the null hypothesis such that the means of two populations are equal. All such tests are usually called Student's t-tests, though strictly speaking that name should only be used if the variances of the two populations are also assumed to be equal; the form of the test used when this assumption is dropped is sometimes called Welch's t-test. These tests are often referred to as "unpaired" or "independent samples" t-tests, as they are typically applied when the statistical units underlying the two samples being compared are non- overlapping.

3.2. Multiple linear regression

Multiple linear regression, also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the independent variables and dependent variable. In essence, multiple linear regression is the extension of ordinary least-squares regression that involves more than one explanatory variable.

3.2.1. Formula

The formula for multiple linear regression is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \ (*)$$

where:

$i = 1,2, \dots, n$ observations

and $n > k$

y_i: dependent variable

x_{ij}: the i-th observation of variable x_j

β_0: the y intercept term (constant term)

β_j: the regression coefficient

ϵ_i: model's error term (or residual)

Each observation $(x_{i1}, x_{i2}, \dots, x_{ik}, y_i)$ satisfies the model in equation (*)

3.3. Pearson's correlation test

3.3.1. Definition:

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name. The Pearson's correlation coefficient is a quantitative measure of the strength of the linear relationship between two random variables x and y.

3.3.2. Formula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

4. Using R Studio to analyze data

4.1. T-test

4.1.1. Data table

GDP | Method 2.R | Method 1.R* | AirPollution

Filter

	Year	PhiIn	PhiOut	VietIn	VietOut
1	1990	29469.15	11311.38	33595.18	10635.24
2	1991	28168.25	11446.75	33210.53	10855.21
3	1992	28048.59	11746.60	32673.38	11154.82
4	1993	26610.33	11769.58	32189.42	11407.13
5	1994	25835.42	11909.94	31500.75	11736.37
6	1995	25599.15	12310.74	30848.22	12038.83
7	1996	25000.10	12558.01	30225.89	12456.92
8	1997	24913.84	13000.48	29844.38	12893.41
9	1998	24883.96	13375.36	29513.18	13456.23
10	1999	25097.28	13714.29	29430.79	14064.83
11	2000	25676.02	14221.93	29539.65	14663.06
12	2001	26740.46	14930.00	29384.36	15570.95
13	2002	27672.65	15485.04	29231.45	16356.56
14	2003	28694.62	15884.78	28904.95	17288.27
15	2004	30061.11	16421.96	28770.30	18094.41
16	2005	31906.84	17124.96	28130.05	18975.70
17	2006	33472.96	17835.65	27618.66	20052.49
18	2007	34860.36	18473.23	26751.38	21259.62
19	2008	36783.81	19435.27	26069.06	22372.28
20	2009	38765.90	20313.10	25402.52	23447.26
21	2010	40094.05	21028.55	24310.88	24595.28
22	2011	41522.35	21273.60	23683.83	25482.83
23	2012	41772.44	22372.07	22980.69	26288.69
24	2013	42347.97	22215.84	22258.52	27516.07
25	2014	42736.90	22206.59	21956.85	28077.13
26	2015	43207.61	22722.93	21773.24	28207.42
27	2016	42603.98	22242.41	21783.13	28431.96
28	2017	42000.07	22638.53	21274.05	29549.67

*Figure **4.1**: Data imported in R*

4.1.2. R Code

```r
#AP: Air Pollution
AP.data <- read.csv('G:/Shared drives/Probability and Statistics/Project/AirPollution.csv')
PhiIn=AP.data$PhiIn
VieIn=AP.data$VieIn
PhiOut=AP.data$PhiOut
VieOut=AP.data$VieOut

ttestIn=t.test(PhiIn,VieIn,alternative = "greater",paired = FALSE, var.equal = FALSE, conf.level = 0.95)
ttestOut=t.test(PhiOut,VieOut,alternative = "greater",paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

Figure 4.2: T-test in R code

We want to compare the means of death in air pollution. We use 2 Welch (two samples) T-tests with significant level $\alpha = 5\% = 0.05$

1. Testing indoor air pollution death between Philippines and Vietnam

   Denote:

   μ_1: mean of indoor air pollution death in Philippines

   μ_3: mean of indoor air pollution death in Vietnam

   Hypotheses:

   $H_0: \mu_1 \leq \mu_3$ $H_1: \mu_1 > \mu_3$

2. Testing outdoor air pollution death between Philippines and Vietnam

   Denote:

   μ_2: mean of outdoor air pollution death in Philippines

   μ_4: mean of outdoor air pollution death in Vietnam

   Hypothesis:

   $H_0: \mu_2 \leq \mu_4$ $H_1: \mu_2 > \mu_4$

4.1.3. Result

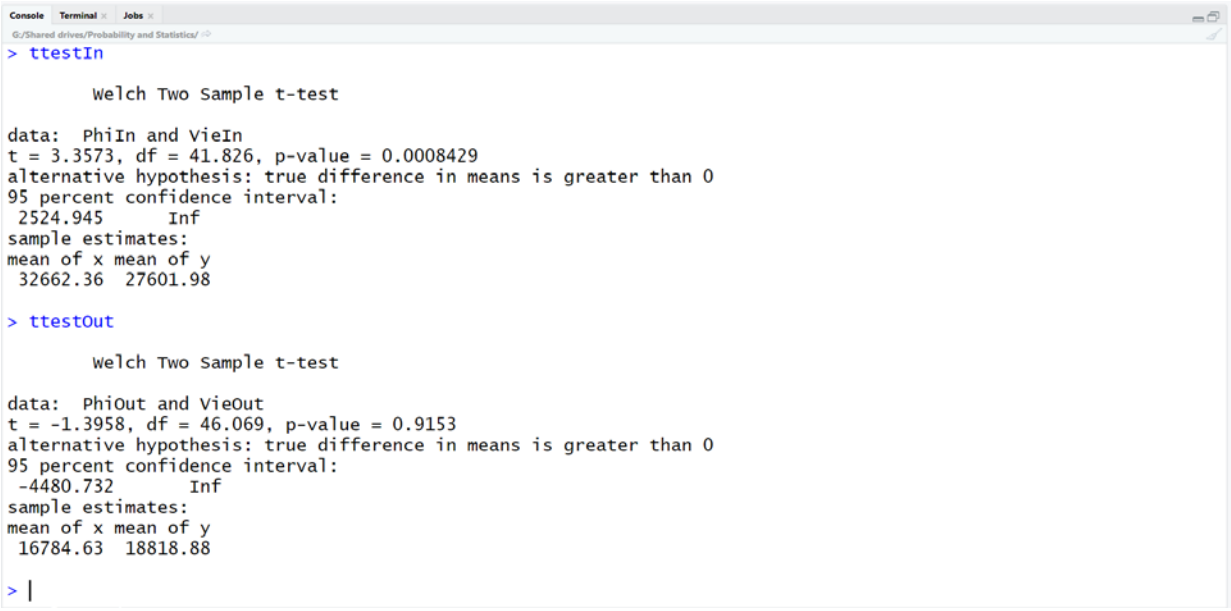

Figure 4.3: Result of hypothesis test

In test 1:

Degree of freedom: $df = 41.826$

The test statistic $t = 3.3573$ is in the rejection range (the 5% range) $[1.6821, +\infty)$

The p-value is

$$p_{v1} = 0.0008429 < \alpha$$

So, we can reject the null hypothesis H_0 and the chance of type I error (rejecting a correct H_0) is small: 0.0008429 (0.084%). We can conclude, the average of indoor air pollution death in Philippines is considered to be greater than one in Vietnam.

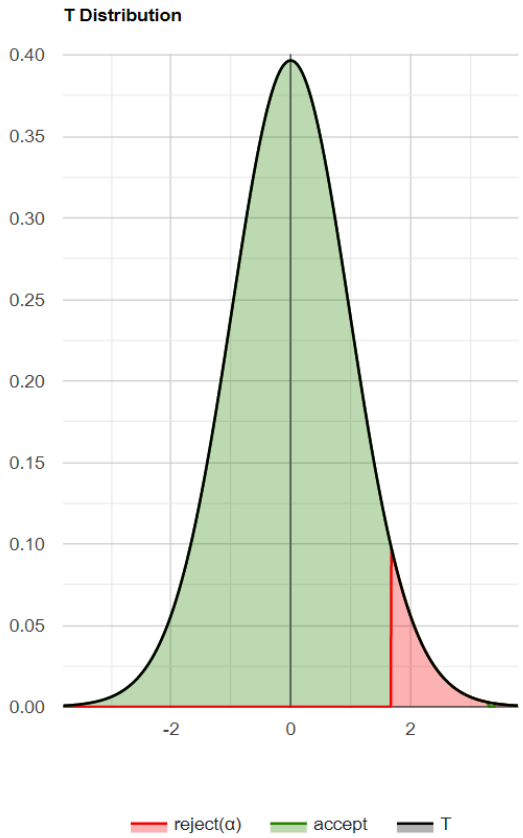

Figure 4.4: T-distribution with rejection area of test 1

In test 2:

Degree of freedom: $df = 46.069$

The test statistic $t = -1.3958$ is in the 95% critical value accepted range $(-\infty, 1.6786]$

The p-value is

$$p_{v2} = 0.9153 > \alpha$$

So, we cannot reject the null hypothesis H_0, in contrast, the result strongly supports H_0. This means that if we would reject H_0, the chance of type I error (rejecting a correct H_0) would be too high: 0.9153 (91.53%). Moreover, the larger the p-value the more it supports H_0.

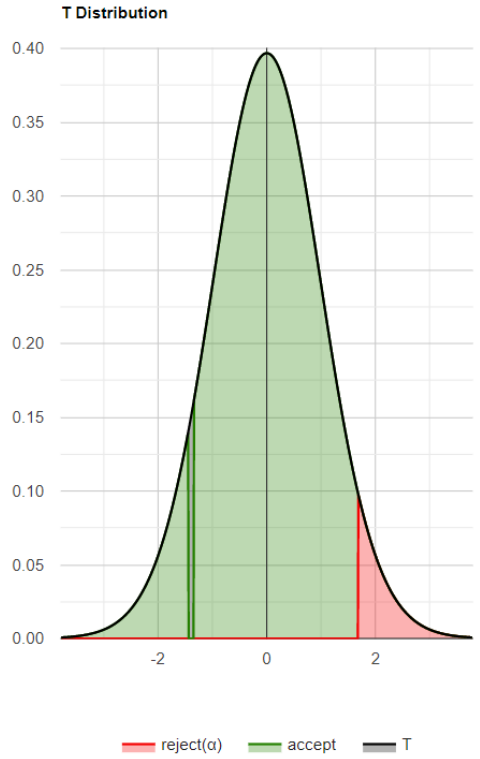

Figure 4.5: T-distribution with rejection area of test 2

We conclude, the average of outdoor air pollution death in Philippines is considered to be less than or equal to one in Vietnam.

4.2. Multiple linear regression

4.2.1. Data table

GDP × | Method 2.R* × | Method 1-T-test.R* × | AirPollution ×

Filter

	Year	WorkHr	GDP	UnEmp
1	2000	2393.631	2562.105	2.33766
2	2001	2246.991	2692.125	2.53165
3	2002	2230.760	2833.771	2.22772
4	2003	2254.885	3000.311	2.17391
5	2004	2207.492	3196.297	2.11765
6	2005	2191.707	3405.679	4.73296
7	2006	2240.841	3609.683	4.88424
8	2007	2241.181	3831.243	4.13978
9	2008	2336.682	4009.959	3.62748
10	2009	2405.354	4185.020	2.60939
11	2010	2300.020	4408.169	2.66792
12	2011	2331.283	4632.766	2.03578
13	2012	2310.541	4821.137	1.77658
14	2013	2267.488	5024.439	1.70906
15	2014	2150.636	5264.828	1.86220
16	2015	2169.592	5554.858	2.11189
17	2016	2169.592	5837.629	3.02000
18	2017	2169.592	6171.884	2.00730

Figure 4.6: Data imported to R studio

4.2.2. Code

```
1  data<- read.csv('G:/Shared drives/Probability and Statistics/Project/GDP.csv')
2  GDP = data$GDP
3  woHRs=data$WorkHr
4  unEmp=data$UnEmp
5
6  plot (GDP ~ woHRs, pch=16)
7  abline (lm(GDP ~ woHRs))
8
9  plot (GDP ~ unEmp, pch=16)
10 abline (lm(GDP ~ unEmp))
11
12 m1 = lm(GDP ~ woHRs)
13 m2 = lm(GDP ~ unEmp)
14 m3 = lm(GDP ~ woHRs + unEmp)
15 summary (m1); summary (m2); summary (m3)
16
```

Figure 4.7: R code for multiple linear regression

4.2.3. Result

We want to know whether GDP per capita of Vietnam depends on working hours and unemployment rate.

Denote:

GPD per capita = Y

Working hours per person = X_1

Unemployment rate = X_2

First, we create scatter plots (X,Y) of (X_1,Y) and (X_2,Y)

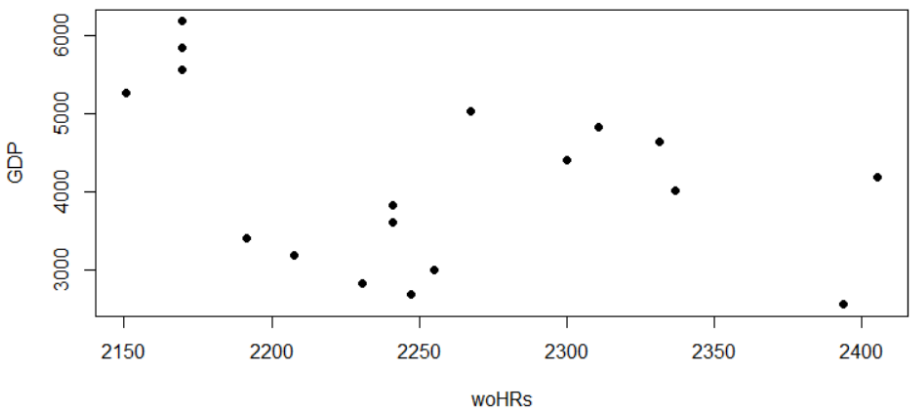

Figure 4.8: Scatter plot of (X_1,Y)

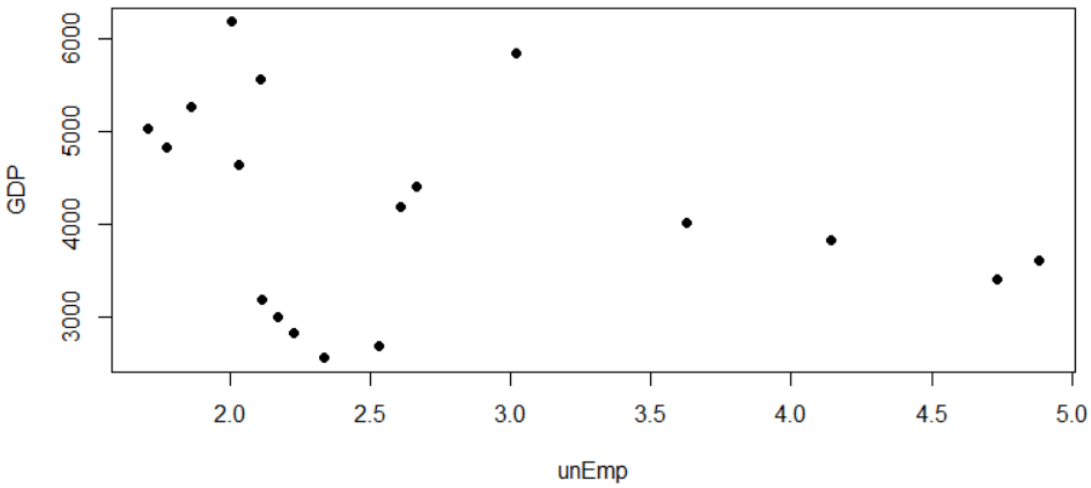

Figure 4.9: Scatter plot of (X_2,Y)

The data is quite scattered, we will use simple linear regression to create the regression lines in two plots. The function in R is $lm(y \sim x)$

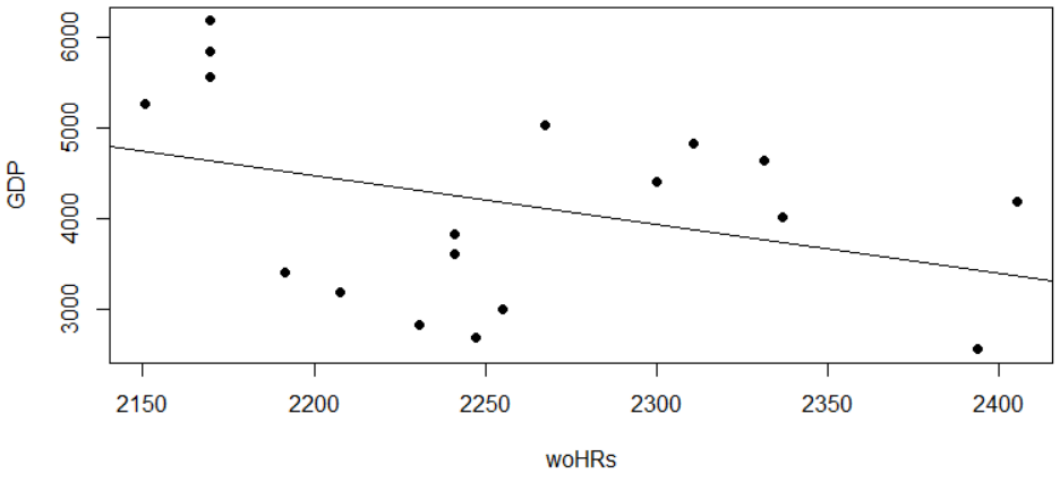

Figure 4.10: Scatter plot of (X_1,Y) with regression line

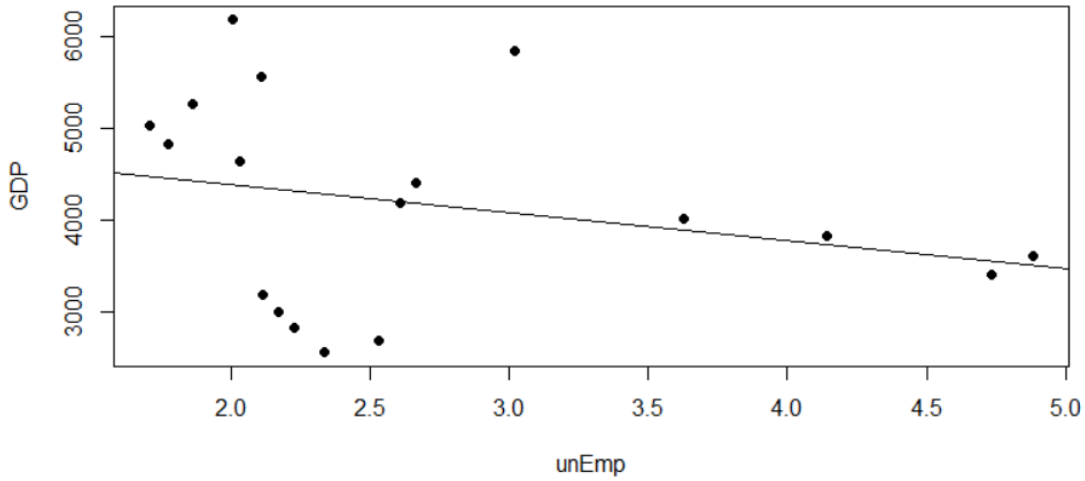

Figure 4.11: Scatter plot of (X_2,Y) with regression line

Now we use multiple linear regression to find out whether the data fits the model. The code in R is $lm(y{\sim}x_1 + x_2)$

```
Console  Terminal  Jobs
G:/Shared drives/Probability and Statistics/
> m3 = lm(GDP ~ woHRs + unEmp)
> summary (m3)

Call:
lm(formula = GDP ~ woHRs + unEmp)

Residuals:
    Min      1Q  Median      3Q     Max
-1635.6  -831.3   349.0   661.1  1286.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17730.309   7595.862   2.334   0.0339 *
woHRs          -5.615      3.333  -1.685   0.1127
unEmp        -330.120    257.026  -1.284   0.2185
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1050 on 15 degrees of freedom
Multiple R-squared:  0.221,     Adjusted R-squared:  0.1172
F-statistic: 2.128 on 2 and 15 DF,  p-value: 0.1536

> |
```

Figure 4.12: Result of multiple linear regression

Y and X relationship

Coefficient of multiple determination is $R^2 = 0.221$. It means that the X explain 22.1% of the variance of Y.

Adjusted R square $R^2_{adj} = 0.1172$.

It means that there is a weak direct relationship between the predicted data ($\hat{y}$) and the observed data (y). Now we use another method to analyze the data.

4.3. Pearson's correlation coefficient

4.3.1. Code

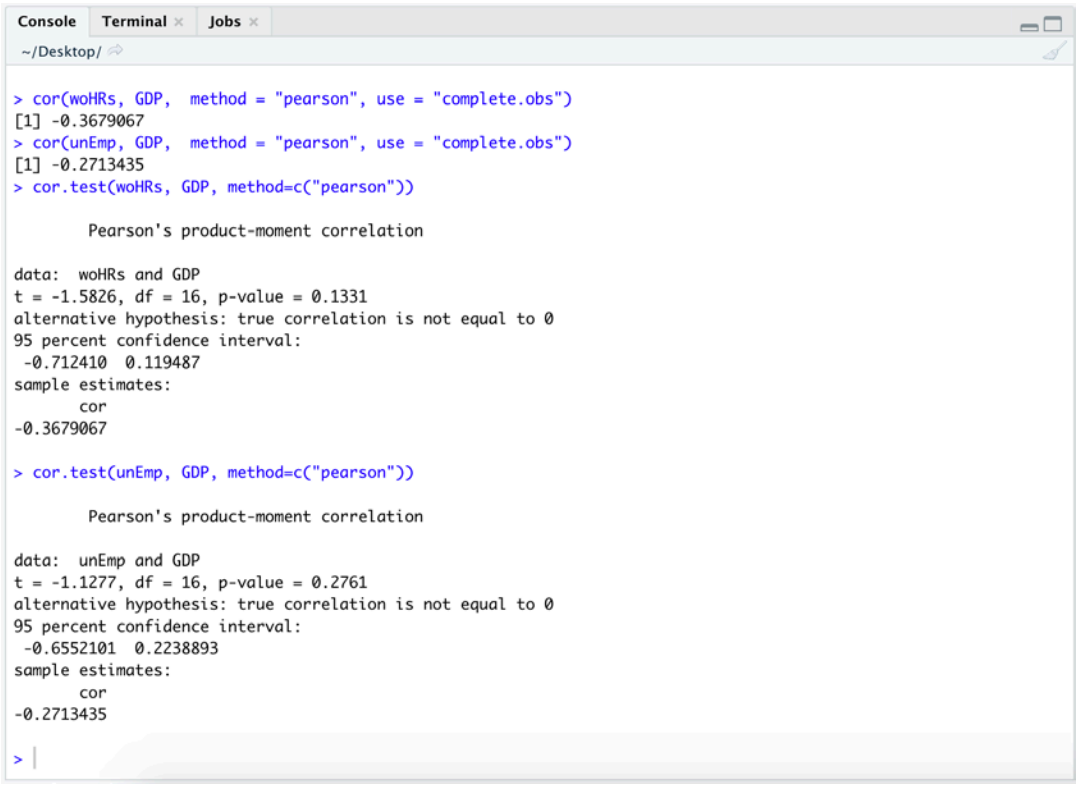

Figure 4.13: Pearson's correlation test

4.3.1. Result

We want to check the relation between GDP and working hours, GDP and unemployment rate, the result is:

Working hours and GDP:

Pearson's correlation coefficient $r_1 = -0.3679067$

Because $r_1 = -0.3679067 < 0$, the increase in working hours causes a decrease in GDP and vice versa.

However, $-0.3 > r_1 > -0.45$ we can conclude there is a normal correlation between working hours and GDP.

Unemployed and GDP:

Pearson's correlation coefficient $r_2 = -0.2713435$

Because $r_2 = -0.2713435 < 0$, the increase in unemployment rate causes a decrease in GDP and vice versa.

$0.29 > r_2 > -0.29$: there is weak correlation between unemployment rate and GDP.

5. Conclusion

From the analyses above, we have some conclusions:

- Philippines has greater average deaths in indoor air pollution than Vietnam does. However, we have enough evidence to say that the average outdoor pollution deaths in Philippines is less than or equal to one in Vietnam.
- Working hours has more influence in the change of GDP than unemployment rate. Working hours and unemployed have the Pearson's correlation coefficient slightly different than zero and lower than zero, so that the functions between GDP and each of them are decreasing function respectively. But in general, the R squared is so small that the multiple linear regression model does not fit the data. We can predict that there are other factors, rather than "working hours" and "unemployment rate", which affect GDP per capita in Vietnam.

6. References

Dũng, N. T. (2019). *Xác suất - thống kê và phân tích số liệu.* VNU Press.

Montgomery, D. C. (2014). *Applied Statistics and Probability for Engineers.* Wiley Press.

Verzani, J. (n.d.). *simpleR - Using R for Introductory Statistics.*