Logo, company name

Description automatically generatedShape, square

Description automatically generated

**CLASS: IS403.M21.HTCL**

**TEAM: 3**

**STUDENTS:**

TRẦN MẪN QUÂN 19520873

ĐẶNG NGUYỄN PHƯỚC AN 19521171

TRẦN THỊ NGỌC AN 19521189

**VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY**

**UNIVERSITY OF INFORMATION TECHNOLOGY**

**FACULTY OF INFORMATION SYSTEM**

———h&g———

**DATA ANALYSIS IN BUSINESS**

**CONTENT: REPORT LAB 3**

Ho Chi Minh City, April 2022

**TABLE OF CONTENT**

[WORK ASSIGNMENT TABLE 3](#_Toc101949207)

[I. EXERCISE 3A 4](#_Toc101949208)

[II. EXERCISE 3B 7](#_Toc101949209)

[III. EXERCISE 3C 10](#_Toc101949210)

[IV. EXERCISE 4D 13](#_Toc101949211)

[1. Ms Excel 13](#_Toc101949212)

[2. R 15](#_Toc101949213)

[3. Python 18](#_Toc101949214)

[V. EXERCISE 4E 21](#_Toc101949215)

[1. Ms Excel 23](#_Toc101949216)

[2. R 25](#_Toc101949217)

[3. Python 27](#_Toc101949218)

[VI. EXERCISE 4F 30](#_Toc101949219)

[1. Ms Excel 31](#_Toc101949220)

[2. R 34](#_Toc101949221)

[3. Python 38](#_Toc101949222)

[REFERENCES 40](#_Toc101949223)

WORK ASSIGNMENT TABLE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Student’name** | **Student’s ID** | **Assignment** | **Evaluation(%)** |
| 1 | Trần Mẫn Quân | 19520873 | - Done Ex 3B, 4E  - Found datasets for 4E  - Written problem statements and formated final report | 100% |
| 2 | Đặng Nguyễn Phước An | 19521171 | - Done Ex 3A, 4D  - Written raw report  - Present | 100% |
| 3 | Trần Thị Ngọc An | 19521189 | - Done Ex 3C, 4F  - Found datasets for 4F  - Searched for reference resources | 100% |

1. EXERCISE 3A

**Explanation:** The multiple linear regression model is the equation that describes the relationship between dependent variables **Y** and independent variables **X1, X2,… Xp**and the wrong number

Y = 0 + 1X1 + 2 X2 + . . . + + p Xp +

With:

* **Y** is the dependent variable
* **X1, X2, . . . Xp** are independent variables.
* **0, 1, 2 . . . , p** are parameters
* is a random variable called the wrong number of orders.

Through multiple linear regression analysis, we can predict the value of Y when we know the value of independent variables X.

In multiple regression analysis we need to pay attention to some of the following criteria:

* **R square, Adjusted R square**: this value reflects the degree of interpretation of dependent variables of independent variables in the regression model. Adjusted R square reflects more closely than R square. The oscillation of these two values is from 0 to 1, from 0.5 or higher, the pattern is good, and vice versa. This value is usually in the Regression Statistics table.
* **Sig value of test F:** used to test the suitability of the regression model through the rejection or acceptance of the hollow hypothesis:

A screenshot of a computer

Description automatically generated with low confidence

If Sig < 0.05, we reject H0 and conclude at least one Bj is not 0 🡪 multiple linear regression model that fits the dataset and is usable. This value is usually in the ANOVA table.

* **The t-test:** is used to determine if each individual independent variable makes sense. The t-test testing is performed separately for each independent variable in the model. The t-test is considered an individual meaning test.

**Example:** Programmer salary survey

A software company collects the data of a sample of 20 programmers. It is recommended to use regression analysis to determine if the salary is related to the number of years of experience and programming aptitude test scores organized by the company. The number of years of experience, aptitude test scores, and annual salaries ($1000s) of 20 programmers is presented in the following table:

Table

Description automatically generated with medium confidence

Suppose we believe that the annual **salary** (Y) is associated with the number of years of experience (X1) and the aptitude test score (X2) according to the following regression model:

Y = 0 + 1X1 + 2 X2

With:

* *Y = The annual salary ($1000)*
* *X1* ***=*** *The number of years of experience.*
* *X2**= The aptitude test score.*

After performing the calculation of the multiple linear regression equation, we get the result as shown below:

Table

Description automatically generated

**In the Regression Statistics table**: We see an **Adjusted R Square** value of approximately **0.8146** which means that the relevance **/Re le vant/** of the linear regression model to this dataset is approximately 81%.

**In the ANOVA table**: We need to assess model fit accurately through ANOVA testing. The result is **Sig F =2.32774E-07**. Because the Sig F value is **less than 0.05**, we reject the H0 hypothesis and conclude that the regression model is appropriate.

**In the Coefficients /Cô e fi tion/ table:** We will evaluate the regression coefficient of each meaningful independent variable in the model based on testing t. Once tested, we see that all the toxic variables in the table have **a p-value of less than 0.05**. So we reject H0, which means that both independent variables make sense.

Based on the above results, we have the following equation:

**Salary = 3.174 + 1.404(Experience) + 0.251(Score)**

So:

Salaries are expected to increase by $1,404 for each additional year of experience (when aptitude scores are kept unchanged). Or the salary is expected to increase by $251 for every year of additional experience (when the number of years of experience is kept constant).

1. EXERCISE 3B

**Explanation:** Multiple nonlinear regression is a correlation relationship between criteria that is not represented by a straight line but by curves of different shapes. Some common curves:

- Parabolic equation (2nd degree): apply when the cause criterion increases or decreases by an equal amount, the effect criterion fluctuates by an unequal amount (faster or slower). For example: age and labor productivity.

🡪 Regression equation: Y = 0 + 1Xi2 + 2Xi + … +

- Hyperbolic equation: apply when the cause criterion increases, the effect criterion decreases at a rate erratic.

🡪 Regression equation: Y = 0 + + … +

- Exponential equation: apply when the value of the result criterion changes exponentially.

🡪 Regression equation: Y = 0 1X +…+

or ln (Y) = ln 0 + X.ln 1 + … +

This model is linear in the β parameters so we can use linear regression methods.

**Example:** Temperature forecast in Ho Chi Minh City

The dataset used below captures 2501 lines of data from air traffic control stations in Vietnam in 2021. This data has a lot of different attribute columns, but our team only selected the incoming data from *Ho Chi Minh City* and divided it into three attribute columns as follows: *Temperature, Humidity and Pressure*. Use this data to make temperature predictions for manufacturing operations. (Link of dataset: [[Vietnamese] Dataset on air quality in Vietnam in 2021 - [Vietnamese] Dataset on air quality in Vietnam in 2021 OD Mekong Datahub (opendevelopmentmekong.net)](https://data.vietnam.opendevelopmentmekong.net/dataset/dataset-on-air-quality-in-vietnam-in-2021/resource/6afcacb0-2e4e-41e0-84f3-d0f632b40ea5) )

Table, Excel

Description automatically generated

Suppose we believe that the annual **Temperature** (Y) is associated with **Humidity** (X1), **Pressure** (X2) according to the following regression model:

Y = 0 + 1X1 + 2 X2 +

With:

* *Y = Temperature*
* *X1 = Humidity*
* *X2 = Pressure*

But in fact, **Temperature** will be inversely proportional to **Humidity** and directly proportional to **Pressure**. So we will have the complete equation as follows.

Y = 0 + 1 + 2 X2 +

After performing the calculation of the multiple nonlinear regression equation, we get the result as shown below:

Table, Excel

Description automatically generated

**In the Regression Statistics table**: We see an **Adjusted R Square** value of approximately **0.58** which means that the relevance of the linear regression model to this dataset is approximately 58%.

**In the ANOVA table**: We need to assess model fit accurately through ANOVA testing. The result is **Sig F =1.187E-21**. Because the Sig F value is **less than 0.05**, we reject the H0 hypothesis and conclude that the regression model is appropriate.

**In the Coefficients table:** We will evaluate the regression coefficient of each meaningful independent variable in the model based on testing t. Once tested, we see that all the toxic variables in the table have **a p-value of less than 0.05**. So we reject H0, which means that both independent variables make sense.

Based on the above results, we have the following equation:

**Temperature =357.34 + 396.47 () – 0.33(Pressure)**

So:

Multiple R = 0.77 Shows that the relationship between variables is relatively close.

R2 = 0.59 shows that in 100% of graduation fluctuations, 59% of volatility is due to Invert of Humidity and Pressure. The remaining 41% is due to random factors and other factors not present in the model.

1. EXERCISE 3C

**Explanation:** In statistics, multiple logistic regression is one of the most commonly used forms of nonlinear regression. It is used to estimate the probability of an event based on more than one independent variable. We use it to predict a binary outcome (*1/0, Yes/ No, True/ False*) given as a set of independent variables. Moreover, it helps to represent binary/categorical outcomes by using dummy variables.

Multiple logistic regression identifies the relationships between the *enumerated variables* and *independent variables* using the probability theory.:

If p is the probability that an event occurs (e.g. death), then 1-p is the probability that the event does not occur (e.g. survival). We have Odds =

If Odds > 1: The probability that event A occurs is more likely than its opposite event.

If Odds < 1: The probability that event A occurs is less likely than its opposite event.

If Odds = 1: The probability that event A occurs is equal to its opposite event.

From Odds, we convert Y in the linear regression equation to log (Odds).

The logistic regression equation states the following:

Log(Odds) = log() = α + β1X1 + β2X2 + … + ε

So, we can calculate the probability of predicting death by the value of X as follow:

=

p =

To test the performance of this model, we must consider a few metrics:

- Null Deviance: the response that is predicted by the model is just an intercept.

- Residual Deviance: It indicates the response predicted by a model of adding independent variables.

**Example:** Study on the relationship between blood procalcitonin levels and mortality in sepsis

There are various factors that may increase the risk of death in sepsis, such as patient age, hypotension, underlying disease or immunodeficiency, etc. In this example, we will consider only up to 3 factors: procalcitonin (variable), hypotension (binary variable), and patient age (variable).

The data of 30 patients with sepsis is as follows:

Table

Description automatically generated with low confidence

There are 5 columns: ID, Age, Hypotension (low blood pressure), Procalcitonin (ng/ml), and Death. Suppose we believe that the risk of death in sepsis is associated with the patient's age (X1), hypotension (X2), and procalcitonin (X3) according to the following regression model:

log()= α + β1X1 + β2X2 + β3X3

After performing the calculation of the multiple logistic regression equation, we get the result as shown below:

Graphical user interface, application, table, Excel

Description automatically generated

The residual deviance value indicates the fit of the model. The smaller this value, the better the model fits. In this model with Residual Deviance = 14,627 is not very high, so it has a good fit with the overall model.

With this result, the logistic regression model is written as:

log()= -3,232 + 0,013. (age) + 4.915. (hypotension) + 0,137. (procalcitonin)

or = -3,232 + 0,013. (age) + 4.915. (hypotension) + 0,137. (procalcitonin)

As paraphrased above:

* Odds ratio of age = e0,013 = 1,013
* Odds ratio of hypotension = 𝑒4,915= 136
* Odds ratio of procalcitonin = 𝑒0,137 = 1,147

🡪 Conclusion: Only hypotension was an independent factor in mortality in patients with sepsis with odds ratio = 136 (95% CI: 3 – 6447; p=0.01).

1. EXERCISE 4D

**Problem statement:** This is a sample of 49 schools in the US. It is recommended to use regression analysis to determine if the annual **percentage of graduation** is associated with the median SAT, the acceptance rate, expenditures/student, and the top 10% HS of each school, according to the following regression model:

Y = 0 + 1X1 + 2 X2 + 3 X3 + 4 X4

With:

* *Y: the annual percentage of graduation (%)*
* *X1 :**the median SAT*
* *X2**: the acceptance rate*
* *X3**:**expenditures/student*
* *X4**: top 10% HS*
* *0, 1, 2, 3, 4 : parameters*
  1. Ms Excel

**Step 1:** Open file “Colleges and Universities.xlsx” to calculate.

Table

Description automatically generated

**Step 2:** Regression

* Go to Data 🡪 Data Analysis 🡪 Choose Regression.

Graphical user interface, application

Description automatically generated

* For **Input Y Range**, select property columns **Graduation**. For **Input X Range,** select the remaining four property columns as (**Median SAT**, **Acceptance Rate,** **Expenditures/Student, Top 10% HS).** The rest of the settings are seen in the image below.

Graphical user interface, application

Description automatically generated

**Step 3:** View the result and conclude

**Table

Description automatically generated**

**In the Regression Statistics table**: We see an **Adjusted R Square** value of approximately **0.492** which means that the relevance of the linear regression model to this dataset is approximately 49%.

**In the ANOVA table**: We need to assess model fit accurately through ANOVA testing. The result is **Sig F =6.33158E-07**. Because the Sig F value is **less than 0.05**, we reject the H0 hypothesis and conclude that the regression model is appropriate.

**In the Coefficients table:** We will evaluate the regression coefficient of each meaningful independent variable in the model based on testing t. Once tested, we see that all the toxic variables in the table have **a p-value of less than 0.05**. So we reject H0, which means that both independent variables make sense.

Based on the above results, we have the following equation:

**Graduation% = 17.92 +0.072(SAT) – 24.859(ACCEPTANCE) – 0.000136(EXPENDITURES) – 0.163 TOP10% HS**

* 1. R

**Step 1:**

* Convert the **Colleges and Universities**.**xlsx** file to **Question4d\_University**.**csv**. Select \*CSV format like the picture below

Graphical user interface, text, application, email

Description automatically generated

**Step2:**

* Import file \*csv in step 1 into RStudio Application

uni<-read.csv('Question4d\_University.csv', header = TRUE)



* Attach this file

attach(uni)



**Step 3:** Calculate

>reg1=lm(Graduation..~Median.SAT+Acceptance.Rate+Expenditures.Student+Top.10..HS)

>summary(reg1)

The results are as follows:

Table

Description automatically generated with medium confidence

**In the Regression Statistics table**: We see an **Adjusted R Square** value of approximately **0.492** which means that the relevance of the linear regression model to this dataset is approximately 49%.

**In the ANOVA table**: We need to assess model fit accurately through ANOVA testing. The result is **Sig F =6.33158E-07**. Because the Sig F value is **less than 0.05**, we reject the H0 hypothesis and conclude that the regression model is appropriate.

**In the Coefficients table:** We will evaluate the regression coefficient of each meaningful independent variable in the model based on testing t. Once tested, we see that all the toxic variables in the table have **a p-value of less than 0.05**. So we reject H0, which means that both independent variables make sense.

Based on the above results, we have the following equation:

**Graduation% = 17.92 +0.072(SAT) – 24.859(ACCEPTANCE) – 0.000136(EXPENDITURES) – 0.163 TOP10% HS**

* 1. Python

**Step 1:** Open Google Colab to perform the calculation.

A screenshot of a computer

Description automatically generated

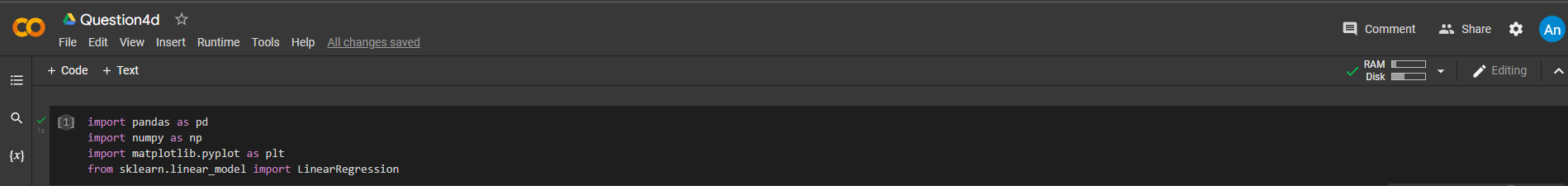
**Step 2**: Import the library

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.linear\_model import LinearRegression



**Step 3:** Import file **Question4d\_University.csv** into Google Colab. Then read this file csv.

from google.colab import files

uploaded = files.upload()

import io

df=pd.read\_csv(io.BytesIO(uploaded  
['Question4d\_University.csv']))

print(df)

Text

Description automatically generated

**Step 4**: Enter the values X, Y to calculate regression.

from sklearn import linear\_model

reg = linear\_model.LinearRegression()

x = np.array(df[['Median SAT','Acceptance Rate','Expenditures/Student','Top 10% HS']]).reshape((-1, 4))

y = np.array(df['Graduation %']).reshape((-1, 1))

model = LinearRegression()

model.fit(x,y)

A screenshot of a computer

Description automatically generated

**Step 5:** Find parameter values and R2

|  |
| --- |
| model.intercept\_  model.coef\_  model.score(x, y) |

Graphical user interface, text, application

Description automatically generated

**Conclusion**

* If the Acceptance rate, expenditures/student and top 10% HS are unchanged, increasing the Median SAT to 1 unit will cause Graduation% to increase to 0.072%.
* If Median SAT, Expenditures/Student and Top 10% HS are unchanged, increasing the acceptance rate to 1 unit will cause Graduation% to decrease by 24,859%.
* If Median SAT, ACCEPTANCE RATE and Top 10% HS remain unchanged, increasing expenditures/students to 1 unit will cause Graduation% to decrease by 0.000136%.
* If the Median SAT, ACCEPTANCE RATE and Expenditures/Student remain unchanged, a 10% increase in HS to 1 unit will cause Graduation% to decrease by 0.163**%.**
* Multiple R = 0.73 Shows that the relationship between variables is relatively close.
* R2 = 0.53 shows that in 100% of graduation fluctuations, 53% of volatility is due to acceptance rate, expenditures/student, Top 10% HS, and median SAT. The remaining 47% is due to random factors and other factors not present in the model.

1. EXERCISE 4E

**Problem statement:**

The dataset used below captures 2501 lines of data from air traffic control stations in Vietnam in 2021. This data has a lot of different attribute columns, but our team only selected the incoming data from *Ho Chi Minh City* and divided it into three attribute columns as follows: *Temperature, Humidity and Pressure*. Use this data to make temperature predictions for manufacturing operations. (Link of dataset: [[Vietnamese] Dataset on air quality in Vietnam in 2021 - [Vietnamese] Dataset on air quality in Vietnam in 2021 OD Mekong Datahub (opendevelopmentmekong.net)](https://data.vietnam.opendevelopmentmekong.net/dataset/dataset-on-air-quality-in-vietnam-in-2021/resource/6afcacb0-2e4e-41e0-84f3-d0f632b40ea5) )

Table, Excel

Description automatically generated

Suppose we believe that the annual **Temperature** (Y) is associated with **Humidity** (X1), **Pressure** (X2) according to the following regression model:

Y = 0 + 1X1 + 2 X2

With:

* *Y = Temperature*
* *X1 = Humidity*
* *X2 = Pressure*

From the equation above, we can conclude that temperature is proportional to pressure and humidity. This means that if the pressure and humidity increase, the higher the temperature will be. However, in reality, the temperature will be inversely proportional to humidity and proportional to pressure.

So, we have the exact equation that will be:

Y = 0 + 1 + 2 X2

It is recommended to use multiple nonlinear regression analysis to determine whether the above model is appropriate.

* 1. Ms Excel

**Step 1:** Open file “nhietdo.xlsx” and change the value of the Humidity property to Invert of Humidity, like the picture below.

To calculate, we only select 3 columns of properties: **Invert of Humidity; Pressure and Temperature.**

Table

Description automatically generated

**Step 2:** Regression

* Go to Data 🡪 Data Analysis 🡪 Choose Regression.

Graphical user interface, application

Description automatically generated

* For **Input Y Range**, select property columns **Temperature**. For **Input X Range,** select the remaining four property columns as (**Invert of Humidity and Pressure).** The rest of the settings are seen in the image below.

Graphical user interface, application

Description automatically generated

**Step 3:** View the result and conclude

**In the Regression Statistics table**: We see an **Adjusted R Square** value of approximately **0.58** which means that the relevance of the linear regression model to this dataset is approximately 58%.

Table, Excel

Description automatically generated

**In the Regression Statistics table**: We see an **Adjusted R Square** value of approximately **0.58** which means that the relevance of the linear regression model to this dataset is approximately 58%.

**In the ANOVA table**: We need to assess model fit accurately through ANOVA testing. The result is **Sig F =1.187E-21**. Because the Sig F value is **less than 0.05**, we reject the H0 hypothesis and conclude that the regression model is appropriate.

**In the Coefficients table:** We will evaluate the regression coefficient of each meaningful independent variable in the model based on testing t. Once tested, we see that all the toxic variables in the table have **a p-value of less than 0.05**. So we reject H0, which means that both independent variables make sense.

Based on the above results, we have the following equation:

**Temperature =357.34 + 396.47 (Invert of Humidity) – 0.33(Pressure)**

* 1. R

**Step 1:**

* Convert **nhietdo**.**xlsx** file to **nhietdo**.**csv**. Select \*CSV format like the picture below

Graphical user interface, text, application, email

Description automatically generated

**Step2:**

* Import file \*csv in step 1 into RStudio Application

temp<-read.csv('nhietdo.csv', header = TRUE)



* Attach this file

attach(temp)



**Step 3:** Calculate

>reg1=lm(NhietDo~ DoAmDao +ApSuat)

>summary(reg1)



The results are as follows:

A screenshot of a computer

Description automatically generated with medium confidence

**In the Regression Statistics table**: We see an **Adjusted R Square** value of approximately **0.58** which means that the relevance of the linear regression model to this dataset is approximately 58%.

**In the ANOVA table**: We need to assess model fit accurately through ANOVA testing. The result is **Sig F =1.187E-21**. Because the Sig F value is **less than 0.05**, we reject the H0 hypothesis and conclude that the regression model is appropriate.

**In the Coeficients table:** We will evaluate the regression coefficient of each meaningful independent variable in the model based on testing t. Once tested, we see that all the toxic variables in the table have **a p-value of less than 0.05**. So we reject H0, which means that both independent variables make sense.

Based on the above results, we have the following equation:

**Temperature =357.34 + 396.47 (Invert of Humidity) – 0.33(Pressure)**

* 1. Python

**Step 1:** Open Google Colab to perform the calculation.

A screenshot of a computer

Description automatically generated

**Step 2**: Import the library

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.linear\_model import LinearRegression

A screenshot of a computer

Description automatically generated with medium confidence

**Step 3:** Import file **nhietdo.csv** into Google Colab. Then read this file csv.

from google.colab import files

uploaded = files.upload()

import io

df=pd.read\_csv(io.BytesIO(uploaded  
['nhietdo.csv']))

print(df)

Graphical user interface, text, application

Description automatically generated

**Step 4**: Enter the values X, Y to calculate regression.

from sklearn import linear\_model

reg = linear\_model.LinearRegression()

x = np.array(df[['DoAmDao','ApSuat']]).reshape((-1, 2))

y = np.array(df['NhietDo']).reshape((-1, 1))

model = LinearRegression()

model.fit(x,y)

Text, chat or text message

Description automatically generated

**Step 5:** Find parameter values and R2

|  |
| --- |
| model.intercept\_  model.coef\_  model.score(x, y) |

Graphical user interface, text, application

Description automatically generated

**Conclusion**

1. Multiple R = 0.77 Shows that the relationship between variables is relatively close.
2. R2 = 0.59 shows that in 100% of graduation fluctuations, 59% of volatility is due to Invert of Humidity and Pressure. The remaining 41% is due to random factors and other factors not present in the model.
3. Based on these figures, users can predict the previous stage of future temperatures. To be able to give the necessary warnings for productive and living activities.

**For example**: If the temperature is too low, it may rain today, from which appropriate measures will be taken to deal with the weather.

1. EXERCISE 4F

Problem statement: A study was conducted to investigate the relationship between fracture risk and bone mineral density and several other biochemical indicators. In 1990, the following data were collected for each subject:

Table

Description automatically generated

We collect this dataset from <https://github.com/tuanvnguyen/R-book> . This site contains all datasets for the book 'Phân tích dữ liệu với R' , author: prof. Nguyễn Văn Tuấn, publisher: Tổng Hợp, 2016. The file’s name is “does\_vn07.csv”

Suppose we believe that the risk of fracture is associated with the duration (X1), femoral neck bone mineral density - fnbmd (X2), lumbar spine bone mineral density - lsbms (X3), age (X4), weight (X5) and height (X6) according to the following regression model:

log()= α + β1X1 + β2X2 + β3X3 + β4X4 + β5X5 + β6X6

Here, since the dependent variable (fracture) is not measured on continuity (but just yes-1 or no-0), it is recommended to use multiple logistic regression analysis to analyze the relationship between the dependent variable and the independent variable.

* 1. Ms Excel

**Step 1:** Open file “Question4f.xlsx”. To calculate logistic regression.

Table

Description automatically generated

**Step 2: Insert Add-in**

To perform logistic regression calculations, we must insert Add-in called **XLMiner Analysis ToolPak.**

**Graphical user interface, application

Description automatically generated**

**Step 3: Open Logistic Regression**

On the **Home** screen of the MS Excel tool. Select the **Show ToolPak function.**

**Graphical user interface, application

Description automatically generated**

Then we select the **Logistic Regression** tool at XLMiner Analysis ToolPak.

Graphical user interface, application

Description automatically generated

**Step 4:**

After inserted and successfully opened **XLMINER ANALYSIS TOOLPAK.** We begin to do the following calculation:

For **Input Y Range**, select property columns **Fracture**. For **Input X Range,** select the remaining four property columns as **(duration, fnbmd, lsbmd, age, wt, ht).** The rest of the settings are seen in the image below.

Graphical user interface, text, application

Description automatically generated

**Step 5:** View the result and conclude

Graphical user interface, application, table, Excel

Description automatically generated

With this result, the logistic regression model is written as:

log()= 6,981 – 0.254(duration) – 1.459(fnbmd) – 1.859 (lsbmd) – 0.035(age) – 0.001(wt) - 0.002(ht)

or = 6,981 – 0.254(duration) – 1.459(fnbmd) – 1.859 (lsbmd) – 0.035(age) – 0.001(wt) - 0.002(ht)

As paraphrased above:

* Odds ratio of duration = 𝑒0,254 = 0,775
* Odds ratio of fnbmd = 𝑒1,459 = 0,232
* Odds ratio of lsbmd = 𝑒1,859 = 0,155
* Odds ratio of age = 𝑒0.035 = 0,965
* Odds ratio of wt = 𝑒0,001 = 1,001
* Odds ratio of ht = 𝑒0,002 = 0,998
  1. R

**Step 1:**

* Convert **does\_vn07**.**xlsx** file to **does\_vn07**.**csv**. Select \*CSV format like the picture below

Graphical user interface, text, application, email

Description automatically generated

**Step2:**

* Import file \*csv in step 1 into RStudio Application

Df <-read.csv('does\_vn07.csv', header = TRUE)

df



* Attach this file

attach(df)



**Step 3:** Then, we have to install and use library **epicalc**

install.packages("epicalc")

****

* In tab Packages, select **Install**

Graphical user interface, application, table

Description automatically generated

* Enter the library name, then click **Install**

**Graphical user interface, text, application, email

Description automatically generated**

* After the installation is complete, we use the **require()** statement to use the library

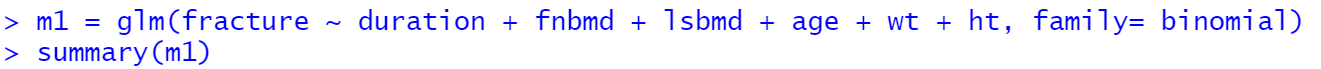
require(epicalc)



**Step 3:** Calculate

M1=glm(fracture ~ duration + fnbmd + lsbmd + age + wt + ht, family = binomial)

Summary(m1)



The results are as follows:

Text

Description automatically generated

With this result, the logistic regression model is written as:

log()= 6,981 – 0.254(duration) – 1.459(fnbmd) – 1.859 (lsbmd) – 0.035(age) – 0.001(wt) - 0.002(ht)

As paraphrased above:

* Odds ratio of duration = 𝑒-0,254 = 0,775
* Odds ratio of fnbmd = 𝑒-1,459 = 0,232
* Odds ratio of lsbmd = 𝑒-1,859 = 0,155
* Odds ratio of age = 𝑒-0,035 = 0,965
* Odds ratio of wt = 𝑒0,001 = 1,001
* Odds ratio of ht = 𝑒-0,002 = 0,998
  1. Python

**Step 1:** Open Google Colab to perform the calculation.

A screenshot of a computer

Description automatically generated

**Step 2**: Import the library

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import statsmodels.formula.api as smf

import statsmodels.api.as sm

from sklearn.linear\_model import LogisticRegression

**Text

Description automatically generated**

**Step 3:** Import file **does\_vn07.csv** into Google Colab. Then read this file csv.

from google.colab import files

uploaded = files.upload()

import io

df=pd.read\_csv(io.BytesIO(uploaded  
['does\_vn07.csv']))

print(df)

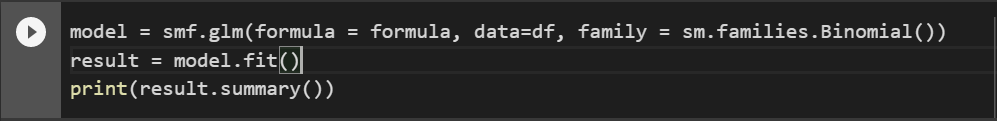
Text

Description automatically generated with low confidence

model = smf.glm(formula = formula, data=df, family = sm.families.Binomial())

result = model.fit()

print(result.summary())



**Result of Python**

Text

Description automatically generated with medium confidence

**Conclusion:**

Each year increases the time of participation until the fracture will reduce the fracture odds by 22% - Odds ratio of fnbmd = exp(-1.46) = 0.232

🡪 For every g/cm2 increases neck bone density, thighs will reduce fractures by 77%

- Odds ratio of lsbmd = 𝑒xp(-1,86) = 0,155

🡪Every g/cm2 increase in spinal bone density will reduce fractures by 85%.

- Odds ratio of age = 𝑒xp(-0.035) = 0,965

🡪 Each year when the age increases, there will be a 1% reduction in fracture odds.

- Odds ratio of wt = 𝑒xp(0,001) = 1,001

🡪 Losing weight will increase the odds of breaking bones by 1 times.

- Odds ratio of ht = 𝑒xp(-0,002) = 0,998

🡪Increasing height will reduce fracture odds by 1%

REFERENCES

<https://fsppm.fulbright.edu.vn/cache/FSLM-10-MultipleRegressionV-2021-03-01-16463454.pdf?fbclid=IwAR31G00z3m3rvOo1LTHLDDA7js8yhxABy45Uu80T0tJvMcpy9pFftnbvIqo>

<https://malouche.github.io/notebooks/regression_model2.html>

[How to Get Regression Model Summary from Scikit-Learn - Statology](https://www.statology.org/sklearn-linear-regression-summary/)

[Hướng Dẫn chạy Multiple Linear Regression trong Python MỚI NHẤT| Học Python TaichinhAcademy - YouTube](https://www.youtube.com/watch?v=pqn8xUEHaNg)

[Phân tích và đọc kết quả hồi quy tuyến tính bội trong SPSS (phamlocblog.com)](https://www.phamlocblog.com/2016/11/cach-chay-hoi-quy-trong-spss.html?fbclid=IwAR0dDPQVjt3dC6LW0SxoWZczIcwLkxKrz8p3svjJ8j1txtXEpuOmitrtqTo)

<https://data-flair.training/blogs/r-nonlinear-regression/>

<http://eldata3.neu.topica.vn/STA302/Giao%20trinh/05_NEU_STA302_Bai5_v1.0013109210.pdf>

<https://cran.r-project.org/doc/contrib/Intro_to_R_Vietnamese.pdf>

<https://bvag.com.vn/wp-content/uploads/2013/01/k2_attachments_CT-Lecture-17.-Introduction-to-logistic-regression.pdf>

<https://bvag.com.vn/wp-content/uploads/2013/01/k2_attachments_PHAN-TICH-HOI-QUY-LOGISTIC-DON-VA-DA-BIEN.pdf>

<https://data.vietnam.opendevelopmentmekong.net/dataset/dataset-on-air-quality-in-vietnam-in-2021/resource/6afcacb0-2e4e-41e0-84f3-d0f632b40ea5>

<https://github.com/tuanvnguyen/R-book/blob/master/does_vn07.csv>