Logo, company name

Description automatically generatedShape, square

Description automatically generated

Ho Chi Minh City, May 2022

**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 4**

**TABLE OF CONTENT**

[WORK ASSIGNMENT TABLE 3](#_Toc104372268)

[I. EXERCISE 4A 4](#_Toc104372269)

[1. R square/Adjusted R square 5](#_Toc104372270)

[2. Residual mean square (RMS) 5](#_Toc104372271)

[3. Using MAE, MSE, RMSE 6](#_Toc104372272)

[4. Mallow’s Cp criterion 6](#_Toc104372273)

[5. Akaike information criterion (AIC) 7](#_Toc104372274)

[6. Bayesian information criterion (BIC) / Schwarz criterion (SC) 7](#_Toc104372275)

[7. Hannan-Quin criterion (HQC) 7](#_Toc104372276)

[II. EXERCISE 4B 7](#_Toc104372277)

[1. Ms Excel 9](#_Toc104372278)

[2. R 11](#_Toc104372279)

[3. Python 14](#_Toc104372280)

[4. Evaluate model accuracy 17](#_Toc104372281)

[III. EXERCISE 4C 20](#_Toc104372282)

[1. Ms Excel 21](#_Toc104372283)

[2. R 25](#_Toc104372284)

[3. Python 28](#_Toc104372285)

[4. Evaluate model accuracy 31](#_Toc104372286)

[IV. EXERCISE 4D 34](#_Toc104372287)

[1. Ms Excel 34](#_Toc104372288)

[2. R 38](#_Toc104372289)

[3. Python 42](#_Toc104372290)

[4. Evaluate model accuracy 46](#_Toc104372291)

[REFERENCES 48](#_Toc104372292)

**TABLE OF CONTENT**

WORK ASSIGNMENT TABLE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Student’s name** | **Student’s ID** | **Assignment** | **Evaluation (%)** |
| 1 | Trần Mẫn Quân | 19520873 | - Done Ex 4A, 4B (4. Evaluate model accuracy)  - Formated final report | 100% |
| 2 | Đặng Nguyễn Phước An | 19521171 | - Done Ex 4D  - Written raw report | 100% |
| 3 | Trần Thị Ngọc An | 19521189 | - Done Ex 4C  - Written raw report | 100% |

1. EXERCISE 4A

***Problem statement:*** *A study of multiple variables will have many different models.*

*For example, we need to predict a dependent variable based on 2 independent variables (k=2) 🡪 the number of possible linear regression models is 3*

*In turn: Y = 0 + 1X1 + 2 X2*

*Y = 0 + 1X1*

*Y = 0 + 2X2*

*🡪 Formula for minimum model number is 2k - 1*

*If we had 100 independent variables, the number of models would be a lot. Therefore, we need to evaluate and choose the most optimal model. In fact, George Box said: “All models are wrong, but some are useful”. What we need to do is find the right model, close to reality, that has a predicted value closest to the most realistic value. That model must meet the parsimonious rule - few parameters but maximum interpretation of the data, avoiding overfitting and underfitting.*

*In summary, the attributes of a good model include:*

1. *Savings: the simplest possible model.*
2. *Relevance: dependent variables are explained as much by modeling as possible.*
3. *Uniformity: the same set of metrics, the estimated parameters have a uniform value.*
4. *Theoretical sustainability: in accordance with theory, i.e. the theoretical basis for building the model.*
5. *Good predictability: inferences for values beyond the sample.*

Currently, we can evaluate and choose the useful forecasting/regression equation model through several criteria:

* Determining coefficient R square, adjusted R square
* Residual mean square (RMS)
* Using MAE, MSE, RMSE
* Mallow’s Cp
* Deviance
* Akaike information criterion (AIC)
* Bayesian information criterion (BIC) / Schwarz criterion (SC)
* Hannan-Quin criterion (HQC)

First, we have: n = number of observations; k = number of estimated parameters

SSR (sum of squares regression) =

SST (sum of squares total) =

SSE (sum of squares error) = RSS (residual sum of squares) = =

🡪 SST = SSE + SSR

Let's talk about each criterion one by one.

* 1. R square/Adjusted R square

These are two criteria that measure the suitability of the model (Measures of Fit).

R2 = Adjusted R2 = 1 - (1 - R2)

Basically, the model has a higher adjusted R2 or R2 is the better model.

R2 increases when there are additional variables in the model. The model with the highest R2 is the model with all the variables. However, this may be the case of overfitting, which does not guarantee the *parsimonious* *rule.* Adjusted R2 does not necessarily increase when more independent variables are added to the regression, so the **adjusted R2** reflects the model's suitability more accurately than the R2 coefficient.

The mechanism of this criterion is to compare and evaluate only models that have:

(1) The dependent variable is the same

(2) Have the same sample size.

Therefore, it is not possible to take R2 as the standard for evaluating logistic regression models. So we will use criteria from 5 onwards to overcome the disadvantages of R2

* 1. Residual mean square (RMS)

The model with the lowest RMS is the best. RMS is equivalent to R2.

RMS is calculated using the formula , with p being the number of prognostic variables.

* 1. Using MAE, MSE, RMSE

This is a method of checking the degree of deviation between the predicted value and the actual value, thereby evaluating whether this model is accurately forecasted or not. The MSE, MAE, and RMSE metrics are mainly used to evaluate the prediction error rates and model performance in regression analysis.

* **MAE** (Mean absolute error): The average of the absolute value of the error between the actual value yi and the predicted value i

MAE =

With n is the number of predictive data samples

The smaller this indicator, the more accurate the model.

* **MSE** (Mean Squared Error): The average of the variance of the error between the actual valuei and the predicted value i

MSE =

With n is the number of predictive data samples

The smaller this indicator (i.e., the closer the estimated values of Y will be to the real value), the more accurate the model is.

* **RMSE** (Root Mean Squared Error): is the error rate by the square root of MSE.

RMSE = =

With n is the number of predictive data samples.

The smaller this indicator, the more accurate the model.

* 1. Mallow’s Cp criterion

If we have a model with p variable prognosis, Cp is calculated by the following formula:

Cp =

With RMSfull being the model with all variables.

The smaller Cp, the better model.

* 1. Akaike information criterion (AIC)

The Akaike information criterion is named after the Japanese statistician Hirotugu Akaike, who built it.

AIC can be calculated in many ways:

AIC =

Or AIC = 2(k – LogLikelihood)

If the sample size is small, we can calculate AICc = AIC +

With n and k being constants, we want RSS as small as possible. Therefore, the lower the AIC, the better the model. We often use AIC for time series forecasting.

* 1. Bayesian information criterion (BIC) / Schwarz criterion (SC)

BIC was developed by Gideon E. Schwarz and published in a 1978 paper where he made a Bayes argument for its application.

We can calculate BIC according to the following formula:

BIC =

Or BIC (= SC) = k\*ln(n) - 2\*LogLikelihood

Similar to AIC, the smaller the BIC, the better the model.

* 1. Hannan-Quin criterion (HQC)

This formula is almost the same as BIC is also for choosing the useful model. It given by:

HQ =

Or HQ = 2[k\*ln(ln(n))- LogLikelihood]

The smaller the HQ, the better the model.

1. EXERCISE 4B

**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 multivariate 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(Graduation..~Median.SAT+Acceptance.Rate+Expenditures.Student+Top.10..HS)

>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 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. 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

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

**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. Evaluate model accuracy

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

As can be seen, the regression equation that we find is full of all independent variables. Therefore, we need to assess whether this case is overfitting or not. Suppose that when we consider only one independent variable (i**nvert of humidity or pressure),** we have two other regression equations as follows:

Graphical user interface, application, table, Excel

Description automatically generated

🡪 **Temperature = 23,6 + 403,19 (Invert of Humidity)**

Application, table

Description automatically generated

🡪 **Temperature = 394,044 – 0,36 (Pressure)**

This model has R2 = 0.10. This value is so low that this model is not suitable. We ignored it and looked at the other two models.

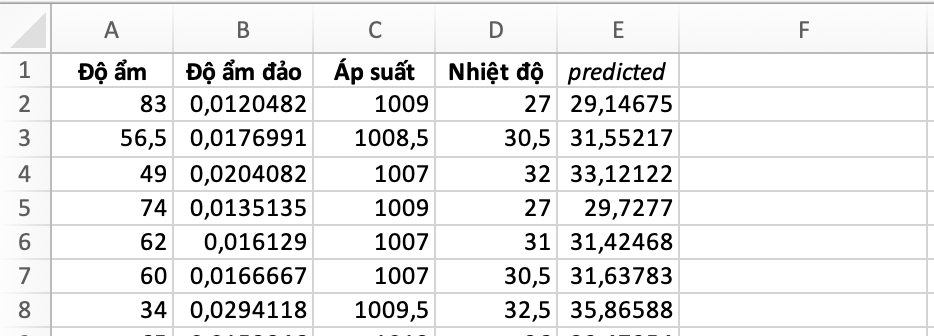
Now, we calculate the criteria in exercise 4a for all two models in turn and compare which models are best suited. Our team will use R2 or Adjusted R2 , MSE and AIC value. We'll choose which model has a higher R2 / Adjusted R2, and a smaller MSE, AIC.

**Model 1:** **Temperature = 357.34 + 396.47 (Invert of Humidity) – 0.33(Pressure)**

As calculated above, this model has R2 = **0,597** and Adjusted R2 = **0,58**

We continue to calculate MSE using the formula: MSE = =

We use this regression model to calculate the predicted values for all 109 samples



Next, we take the observed value minus the predicted value.

Table

Description automatically generated

Summing up all the values in column F, we get RSS = = 367,8956

Then MSE = = = **3,375** ; AIC **=**  = = **3,565**

**Model 2:** **Temperature = 23,6 + 403,19 (Invert of Humidity)**

As calculated above, this model has R2 = **0,506** and Adjusted R2 = **0,502**

We continue to calculate MSE using the formula: MSE = =

We use this regression model to calculate the predicted values for all 109 samples

Graphical user interface, table, Excel

Description automatically generated

Next, we take the observed value minus the predicted value.

Graphical user interface, application, table, Excel

Description automatically generated

Summing up all the values in column F, we get RSS = = 361,033

Then MSE = = = **3,312** ;AIC **=**  = = **3,499**

After all, we got the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | R2 | Adjusted R2 | MSE | AIC |
| Model 1 | 0,597 | 0,589 | 3,375 | 3,565 |
| Model 2 | 0,506 | 0,502 | 3,312 | 3,499 |

Noticed that the two models have MSE and AIC approximately each other, however, model 2 is slightly lower. We know the model has the lower MSE and AIC as possible but considering Adjusted R2, model 1 is higher, explaining 58.9% of the dependent variable. Therefore, we will consider and choose model 1: **Temperature = 357.34 + 396.47 (Invert of Humidity) – 0.33 (Pressure)**

1. EXERCISE 4C

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 from males over 60 years old 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 “fracture.txt”

Description of all the columns:

1. id: patient number

2. fx: fracture or not (0=no fracture, 1=fracture)

3. age

4. bmi: body mass index, calculated by weight divided by square height

5. bmd: (bone mineral density) mineral density in the femur.

6. ictp: a biochemical index that measures bone destruction activity

7. pinp: a biochemical index that measures bone-forming activity

Suppose we believe that the risk of fracture is associated with age (X1), bmi (X2), bmd (X3), ictp (X4) and pinp (X5) according to the following regression model:

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

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

* 1. Ms Excel

**Step 1:** Open file “fracture.txt”. To calculate logistic regression.

Graphical user interface, application, table, Excel

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

Application, table

Description automatically generated with medium confidence

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

**log()= -1.9 + 0.02(age) – 0.05(bmi) – 2.333(bmd) – 0.538(ictp) – 0.01(pinp)**

or **= 1.9 + 0.02(age) – 0.045(bmi) – 2.333(bmd) – 0.538(ictp) – 0.007(pinp)**

As paraphrased above:

* Odds ratio of age = 𝑒0,02 = 1,02
* Odds ratio of bmi = 𝑒-0.05 = 0,95
* Odds ratio of bmd = 𝑒-2.333 = 0,097
* Odds ratio of ictp = 𝑒-0.538 = 0,584
* Odds ratio of pinp = 𝑒-0,01 = 0,99
  1. R

**Step 1:**

* Convert **fracture**.**xlsx** file to **fracture**.**txt**. 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

f <-read.csv('Q4F.csv', header = TRUE)

f

Logo, company name

Description automatically generated with medium confidence





* Attach this file

attach(dat)



**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

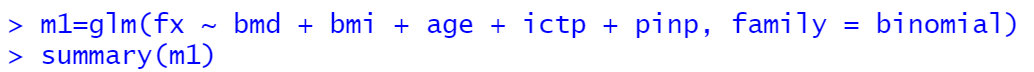
require(epicalc)



**Step 3:** Calculate

m1=glm(fx ~ bmd + bmi + age + ictp + pinp, family = binomial)

Summary(m1)



The results are as follows:

Table

Description automatically generated

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

**log()= -1.9 – 2.4(bmd) – 0.06(bmi) + 0.02(age) + 0.6(ictp) – 0.01(pinp)**

As paraphrased above:

* Odds ratio of bmd = 𝑒-2,4 = 0,09
* Odds ratio of bmi = 𝑒-0,06 = 0,94
* Odds ratio of age = 𝑒0,02 = 1,02
* Odds ratio of ictp = 𝑒0.6 = 1,82
* Odds ratio of pinp = 𝑒-0,01 = 0,99
  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 **fracture.csv** into Google Colab. Then read this file csv.

from google.colab import files

uploaded = files.upload()

import io

df=pd.read\_table(io.BytesIO(uploaded  
['fracture.txt ']))

print(df)

Text

Description automatically generated

formula = 'fx ~ age + bmi + ictp + pinp'

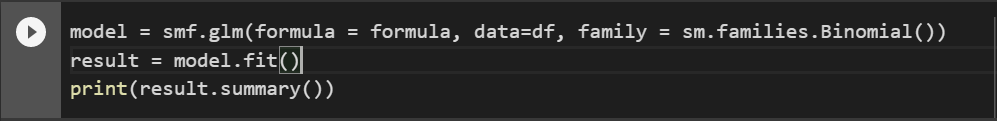
Diagram

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

**Conclusion:**

- Odds ratio of age = exp(0,02) = 1,02

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

- Odds ratio of bmi= exp (-0,06) = 0,94

🡪 Every g/cm2 increase in bmi density will reduce fractures by 6%.

- Odds ratio of bmd = exp (-2,45) = 0,09

🡪 Every g/cm2 increase in bmd will reduce fractures by 99%.

- Odds ratio of ictp = exp (0,6) = 1,82

🡪 Increasing ictp will increase the odds of breaking bones by 1,82 time.

- Odds ratio of pinp = exp (-0,01) = 0,99

🡪 Increasing pinp will reduce fracture odds by 1%

* 1. Evaluate model accuracy

**Step1:**

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

t = file.choose()

t

Dat <-read.table(t, header=T, na.strings = ".")

f

Logo, company name

Description automatically generated with medium confidence





* Attach this file

attach(dat)



**Step 3:** Calculate - Find the optimal model based on AIC

search = glm(fx ~ ., family = "binomial", data = dat)



step(search)



The results are as follows:

A picture containing text, receipt

Description automatically generated

A picture containing text, receipt

Description automatically generated

A picture containing text, receipt

Description automatically generated

Text, letter

Description automatically generated

**Conclusion:**

In the above results, R reports tell us each step in the process of finding the optimal model tomb. The beginning is the model with all 6 variables, and the AIC value = 146.09. The second step consists of only 5 variables (removing pinp) and AIC=144.45, etc. The results can be summarized in the following table:

|  |  |
| --- | --- |
| Model | AIC |
| fx ~ id + age + bmi + bmd + ictp + pinp | 146.09 |
| fx ~ id + age + bmi + bmd + ictp | 144.45 |
| fx ~ age + bmi + bmd + ictp | 142.81 |
| fx ~ bmi + bmd + ictp | 141.33 |
| fx ~ bmd + ictp | 140.34 |

The result of 5 steps to find the model, R stops with the model of 2 variables bmd and ictp because of the lowest AIC value. In short, in this analysis, we conclude that bmd (density of minerals in bones) and ictp (a marker of the bone destruction cycle) are two factors that are associated with or affect the risk of fractures.

So, the original 6-variable model was not the best model. In fact, the fewer independent variables the model, the simpler it is to explain the more dependent variables that are the better. Therefore, the model with 2 variables bmd and ictp is the most useful model.

1. EXERCISE 4D

**Problem statement:** This is the US Department of Agriculture USDA dataset on soybean export sales to Vietnam in 2021. This dataset is very suitable for predicting US export sales to Vietnam in the following years.

* 1. Ms Excel

**Step 1:** Open file “Lab4VietNam.xlsx” to calculate.

Graphical user interface, application, table, Excel

Description automatically generated

**Step 2:** Calculate the Autoregression

* Make the first Lag in excel file.

Table

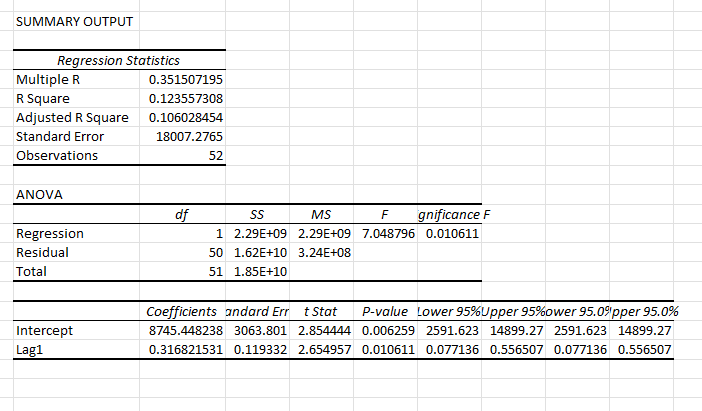
Description automatically generated

* Go to Data 🡪 Data Analysis 🡪 Choose Regression.
* For **Input Y Range**, select property columns **Weekly Exports CMY**. For **Input X LAG1.** The rest of the settings are seen in the image below.

**Graphical user interface, application

Description automatically generated**

* The result



- According to the above result, we can see that **Lag1 + Intercept > 1** and R Square is approximately 10%.

🡪 So we don't take Lag 1🡪 p =0

**Step 3: Moving Average**

* We calculate the Moving Average – is the process of finding the relationship between data current data and q part of the previous past error.
* Go to Data 🡪 Data Analysis 🡪 Choose **Moving Average**.
* For **Input Range**, select property columns **Weekly Exports CMY**. For the interval choose 2

Graphical user interface

Description automatically generated

* The result like below:

Table

Description automatically generated

* Then we calculation the Regression.

Graphical user interface, application, table, Excel

Description automatically generated

* Here because R Square is approximately 0.64, we accept Moving Average
* So that, we choose **q = 1**
* For the value of d, we choose d=1
* The best model is ARIMA(0,1,1)
  1. R

**Step 1:**

* Convert **Lab4VietNam.xlsx** file to **Lab4VietNam**.**csv**. Select \*CSV format like the picture below

Graphical user interface, text, application, email

Description automatically generated

**Step 2:**

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

vn<-read.csv(choose.files())



**Step 3:**

* Install and use these library to calculate the ARIMA









**Step 4:**

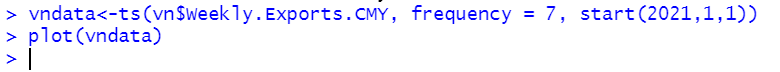
* Define a index column

Text

Description automatically generated

**Step 5:**

* Plotting time series graphs



* Plot graph like the picture below

A picture containing text, boat

Description automatically generated

**Step 6:**

* ADF testing for the time series

Text

Description automatically generated

* We can see that **p-value > 0.01**, so this time series **is not stationary**. So we have to find the correct model for them.

**Step 7:**

* Find the best model of the data

A picture containing text, receipt, screenshot

Description automatically generated

* The above formula has suggested that the most suitable model for this Data is A(0,1,1) with p=0, q=1 and d=1.

**Step 8:** Fit the ARIMA.



**Step 9:** Forecast test

Text

Description automatically generated

**Step 10:** Plot graph for the forecast



Chart

Description automatically generated

**Conclusion:** So the most suitable model for this data is ARIMA(0,1,1). With the above predictions, we can see that, in the following years, the United States will export a lot of soybeans to Vietnam.

* 1. Python

**Step 1:**

* Import the library

A screenshot of a computer

Description automatically generated with medium confidence

**Step 2:**

* Choose the file and open it

Graphical user interface, text, application

Description automatically generated

**Step 3:**

* Make the index

Graphical user interface, text

Description automatically generated with medium confidence

**Step 4:**

* Show the data.

A screenshot of a computer

Description automatically generated with medium confidence

**Step 5:**

* Plot graph of the data

Chart, line chart

Description automatically generated

**Step 6:**

* ADF Test

Text

Description automatically generated

Text

Description automatically generated

**Step 7:**

* Find the best model of the Data

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

* The best model of the data is ARIMA(0,1,1)

**Step 8:**

* Calculate root mean squared and Calculate mean squared

Text

Description automatically generated

* 1. Evaluate model accuracy

To evaluate the accuracy of ARIMA, we use AIC values. We can see that in the following picture:

Graphical user interface, text

Description automatically generated

The lower AIC, the better model. So we choose ARIMA(0,1,1)(0,0,0)[0] because this model has the smallest AIC (=1173.883) and we conclude that is the most useful model.

In another way, we can use the **Box-Jenkins method** on R language

Graphical user interface, text, application, email

Description automatically generated

We can see that with lag values of 5.15 and 20 respectively, the p-values are all greater than 0.05. It follows that model A(0,1,1) is completely suitable.

REFERENCES

1. <https://www.datatechnotes.com/2019/02/regression-model-accuracy-mae-mse-rmse.html>
2. <https://timeseriesreasoning.com/contents/f-test-for-regression-analysis/>
3. <https://rpubs.com/lengockhanhi/445130>
4. <https://mosl.vn/danh-gia-muc-do-phu-hop-cua-mo-hinh-hoi-quy-tren-r/>
5. <https://luanvanhay.org/huong-dan/r-cach-tinh-loglikelihood-aic-bic/>
6. <https://www.statology.org/sst-ssr-sse/>
7. <https://www.investopedia.com/terms/r/residual-sum-of-squares.asp#toc-is-rss-the-same-as-the-sum-of-squared-estimate-of-errors-sse>
8. <https://www.youtube.com/watch?v=eVp0oyKxtrI>
9. <https://www.youtube.com/watch?v=q5M5N66QFBk&list=PLbRKZL7ww3qjGLYt9DyN2zh6o21Phn7IH&index=5&t=692s>
10. <https://sites.google.com/site/nttrungmtwiki/home/it/data-mining/time-series-analysis/arima-ar-ma/a-guide-to-time-series-forecasting-with-arima-in-python-3>