A black background with blue and black border

Description automatically generated

**Ho Chi Minh City, April 2023**

**Lecturers:**  **Assoc. Prof. Nguyen Dinh Thuan**

**TA. Nguyen Minh Nhut**

**Students:**

**Bui Dinh Trieu - 21521576**

**Nguyen Chi Kha - 21522179**

**Nguyễn Minh Duy - 21522005**

**LAB03 REPORT**

**IS403: DATA ANALYSIS IN BUSINESS**

**VIETNAM NATIONAL UNIVERSITY**

**UNIVERSITY OF INFORMATION TECHNOLOGY**

# TABLE OF CONTENTS

[TABLE OF CONTENTS 2](#_Toc163538805)

[TEACHER’S COMMENTS 3](#_Toc163538806)

[WORK DISTRIBUTION 4](#_Toc163538807)

[A. Task 1 5](#_Toc163538808)

[1. Multivariable Linear Regression. 5](#_Toc163538809)

[2. Multivariable Nonlinear Regression 7](#_Toc163538810)

[3. Logistic Regression 9](#_Toc163538811)

[B. TASK B: 11](#_Toc163538812)

[1. Using MS Excel, R language and Python language to perform Multivariable Linear Regression with data file: Colleges and Universities 11](#_Toc163538813)

[1.1. Analyzing by using Excel: 11](#_Toc163538814)

[1.2. Analyzing by using R: 15](#_Toc163538815)

[1.3. Analyzing by using Python: 16](#_Toc163538816)

[1.4 Conclusion: 18](#_Toc163538817)

[2. Using MS Excel, R language and Python language to perform Multivariable Nonlinear Regression with optional real data about/of Vietnam. 19](#_Toc163538818)

[2.1. Analyzing by using Excel: 19](#_Toc163538820)

[2.2. Analyzing by using R: 21](#_Toc163538821)

[2.3. Analyzing by using Python: 23](#_Toc163538822)

[2.4. Conclusion: 25](#_Toc163538823)

[3. Using MS Excel, R language and Python language perform Logistic Regression with optional real data about/of Vietnam. 27](#_Toc163538824)

[3.1. Analyzing by using Excel: 27](#_Toc163538826)

[3.2. Analyzing by using R: 35](#_Toc163538827)

[3.3. Analyzing by using Python: 36](#_Toc163538828)

[3.4. Conclusion: 39](#_Toc163538829)

[C. REFERENCES: 40](#_Toc163538830)

# TEACHER’S COMMENTS

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

# WORK DISTRIBUTION

|  |  |  |  |
| --- | --- | --- | --- |
| **Members**  **Works** | **Nguyen Minh Duy (Leader)** | **Bui Dinh Trieu** | **Nguyen Chi Kha** |
| **Problem statement** |  |  |  |
| **Build the report template** |  |  |  |
| **Do all exercise Task 2a** |  |  |  |
| **Do all exercise Task 2b** |  |  |  |
| **Do all exercise Task 2c** |  |  |  |
| **Summarize and edit reports** |  |  |  |
| **Completion (%)** | 100% | 100% | 100% |

## Task 1

**Explanation (What, How and Why) and example of:**

**a)**  [**Multivariable Linear Regression.**](https://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm)

**b) Multivariable Nonlinear Regression**

**c) Logistic Regression**

## Multivariable Linear Regression.

Multivariable Linear Regression is a statistical technique used to model the relationship between a dependent variable and two or more independent variables. It is an extension of simple linear regression, which involves only one independent variable.

**What is Multivariable Linear Regression ?**

Multivariable Linear Regression is a method that helps us understand how the value of a dependent variable changes when any one of the independent variables is varied, while the other independent variables are held constant. It is widely used in various fields such as economics, finance, engineering, and social sciences to make predictions and analyze the impact of multiple factors on a particular outcome.

**How does Multivariable Linear Regression work ?**

In Multivariable Linear Regression, the relationship between the dependent variable (y) and the independent variables (x1, x2, ..., xn) is expressed as a linear equation:



Where:

* y is the dependent variable
* x1, x2, ..., xn are the independent variables
* β0 is the y-intercept (the value of y when all independent variables are 0)
* β1, β2, ..., βn are the coefficients or slope parameters associated with each independent variable
* ε is the error term, which represents the unexplained variation in the model

The coefficients (β1, β2, ..., βn) are estimated using techniques like Ordinary Least Squares (OLS) or Maximum Likelihood Estimation (MLE). These coefficients represent the change in the dependent variable associated with a one-unit change in the corresponding independent variable, holding all other independent variables constant.

**Why is Multivariable Linear Regression useful ?**

**Multivariable Linear Regression is useful for several reasons:**

* It allows us to analyze the relationship between a dependent variable and multiple independent variables simultaneously.
* It provides insights into the relative importance of each independent variable in predicting the dependent variable.
* It can be used for prediction and forecasting by plugging in new values of the independent variables.
* It helps in understanding and quantifying the impact of different factors on a particular outcome.

**Example:**

Suppose we want to analyze the factors that influence a person's annual income (y). We consider the following independent variables: years of education (x1), years of work experience (x2), and the number of hours worked per week (x3). The multivariable linear regression model can be expressed as:

y = β0 + β1x1 + β2x2 + β3x3 + ε

Using data on annual income, years of education, work experience, and hours worked per week for a sample of individuals, we can estimate the coefficients β0, β1, β2, and β3 using techniques like OLS or MLE.

For instance, if β1 = 5000, it means that, on average, each additional year of education is associated with an increase of $5,000 in annual income, holding work experience and hours worked per week constant.

## Multivariable Nonlinear Regression

Multivariable Nonlinear Regression is a statistical technique used to model the relationship between a dependent variable and two or more independent variables, where the relationship is non-linear.

**What is Multivariable Nonlinear Regression?**

Multivariable Nonlinear Regression is a method that helps us understand how the value of a dependent variable changes as a non-linear function of two or more independent variables. Unlike linear regression, where the relationship between the variables is assumed to be linear, nonlinear regression models allow for more complex, non-linear relationships.

**How does Multivariable Nonlinear Regression work?**

In Multivariable Nonlinear Regression, the relationship between the dependent variable (y) and the independent variables (x1, x2, ..., xn) is expressed as a non-linear function:

y = f(x1, x2, ..., xn; β1, β2, ..., βk) + ε

Where:

* y is the dependent variable
* x1, x2, ..., xn are the independent variables
* f(...) is a non-linear function of the independent variables and parameters β1, β2, ..., βk
* β1, β2, ..., βk are the parameters or coefficients to be estimated
* ε is the error term, which represents the unexplained variation in the model

The specific form of the non-linear function f(...) is determined based on the subject matter knowledge and the nature of the relationship between the variables. Common non-linear functions include polynomials, exponential, logarithmic, and trigonometric functions.

The parameters (β1, β2, ..., βk) are estimated using iterative optimization techniques, such as the Gauss-Newton method or the Levenberg-Marquardt algorithm, which minimize the sum of squared residuals between the observed values and the predicted values from the non-linear model.

**Why is Multivariable Nonlinear Regression useful?**

Multivariable Nonlinear Regression is useful when the relationship between the dependent variable and the independent variables is non-linear, and a linear model would be inadequate or misleading. It allows for more accurate modeling and prediction in situations where the underlying relationships are complex or curved.

**Example:**

Suppose we want to model the growth of a bacterial population (y) over time (x1) in a culture, considering the initial concentration of nutrients (x2) and temperature (x3). The relationship between these variables may follow a non-linear pattern, such as the logistic growth curve:

y = α / (1 + β \* exp(-γ \* (x1 - δ))) + ε

Where:

* y is the bacterial population
* x1 is the time
* x2 is the initial nutrient concentration
* x3 is the temperature
* α, β, γ, and δ are parameters to be estimated

Using data on bacterial population growth, time, initial nutrient concentration, and temperature, we can estimate the parameters α, β, γ, and δ using non-linear regression techniques.

## Logistic Regression

Logistic Regression is a statistical technique used to model the relationship between a binary dependent variable and one or more independent variables. It is widely used in various fields, such as machine learning, biostatistics, and social sciences, for classification problems and probability estimation.

**What is Logistic Regression?**

Logistic Regression is a type of regression analysis that models the probability of an event occurring based on one or more independent variables. The dependent variable in Logistic Regression is binary, meaning it can take only two values, typically coded as 0 and 1, representing the absence or presence of a particular characteristic or outcome.

**How does Logistic Regression work?**

In Logistic Regression, the relationship between the dependent variable (y) and the independent variables (x1, x2, ..., xn) is modeled using the logistic function:

p(y = 1 | x1, x2, ..., xn) = 1 / (1 + exp(-(β0 + β1x1 + β2x2 + ... + βnxn)))

Where:

* p(y = 1 | x1, x2, ..., xn) is the probability of the event occurring (y = 1) given the values of the independent variables
* x1, x2, ..., xn are the independent variables
* β0 is the intercept term
* β1, β2, ..., βn are the coefficients or slope parameters associated with each independent variable

The coefficients (β0, β1, β2, ..., βn) are estimated using techniques like Maximum Likelihood Estimation (MLE) or iterative methods, such as Newton-Raphson or Iteratively Reweighted Least Squares (IRLS).

The estimated coefficients can be used to calculate the odds ratio, which quantifies the change in the odds of the event occurring for a one-unit change in the corresponding independent variable, holding all other variables constant.

**Why is Logistic Regression useful?**

Logistic Regression is useful for several reasons:

It models the probability of a binary outcome, which is useful in various applications, such as credit risk assessment, disease diagnosis, and customer churn prediction.

* It provides interpretable coefficients that indicate the direction and strength of the relationship between the independent variables and the probability of the event occurring.
* It can handle both continuous and categorical independent variables.
* It does not assume a linear relationship between the dependent and independent variables, making it suitable for modeling non-linear relationships.

**Example:**

Suppose we want to predict whether a customer will churn (leave the company) or not, based on their age (x1), account balance (x2), and number of customer service calls (x3). The dependent variable is "churn," coded as 1 for customers who churned and 0 for those who did not.

The logistic regression model can be expressed as:

p(churn = 1 | x1, x2, x3) = 1 / (1 + exp(-(β0 + β1x1 + β2x2 + β3x3)))

Using data on customer churn, age, account balance, and number of customer service calls, we can estimate the coefficients β0, β1, β2, and β3 using MLE or iterative methods.

If β1 = 0.05, it means that for a one-year increase in age, the odds of a customer churning increase by a factor of exp(0.05) = 1.051, holding account balance and number of customer service calls constant.

## B. TASK B:

**a) Using MS Excel, R language and Python language to perform Multivariable Linear Regression with data file: Colleges and Universities**

**b) Using MS Excel, R language and Python language to perform Multivariable Nonlinear Regression with optional real data about/of Vietnam.**

**c) Using MS Excel, R language and Python language perform Logistic Regression with optional real data about/of Vietnam.**

### Using MS Excel, R language and Python language to perform Multivariable Linear Regression with data file: Colleges and Universities

### Analyzing by using Excel:

About dataset:

A screenshot of a spreadsheet

Description automatically generated

**Problem description:**

Interpreting Regression Results for the Colleges and Universities Data

Colleges try to predict Student Graduation Rates using a variety of characteristics: Median SAT, Acceptance Rate, Expenditures/Student, Top 10% of HS class.

With the confidence level is 95% => The significance level α = 0.05.

A math equation with black text

Description automatically generated with medium confidence

* There is no relationship between the dependent variable and the set of independent variables. In this case, all of the regression coefficients  in the population model are zero. This is the claim for the null hypothesis in the overall model test.
* There is a relationship between the dependent variable and the set of independent variables. In this case, at least one of the regression coefficients  in the population model is not zero. This is the claim for the alternative hypothesis in the overall model test.

Using Data Analysis Tool:

A screenshot of a computer

Description automatically generated

The result:

A screenshot of a spreadsheet

Description automatically generated

Recheck by Calculating:

A grid with text on it

Description automatically generated

**Conclusion:**

In ANOVA table , Significance F < 0.05 => reject H0 => there is a relationship between the dependent variable and the set of independent variables

All of the slope coefficient p-values are < 0.05 => accept all of them

The multiple linear regression equation is:

**Graduation**% = 17.92 + 0.072 **SAT** - 24.859 **ACCEPTANCE**

- 0.000136 **EXPENDITURES**

- 0.163 **TOP10% HS**

### Analyzing by using R:

Text

Description automatically generated

The above analysis results show: p-value = 6.332e-07 < 0.05. So we reject H0 and accept H1, which means there exists a meaningful Multivariable Linear Regression model.

The multiple linear regression equation is:

**Graduation**% = 17.92 + 0.072 **SAT** - 24.859 **ACCEPTANCE**

- 0.000136 **EXPENDITURES**

- 0.163 **TOP10% HS**

### Analyzing by using Python:

Import libraries, read and check data:

**Table

Description automatically generated with medium confidence**

Create an array X consisting of many variables and Y is the Graduation %

****

Use LinearRegression function and fit a&b

**Graphical user interface, text, application, email

Description automatically generated**

R Square Score = 0.534

The multiple linear regression equation is:

**Graduation**% = 17.92 + 0.072 **SAT** - 24.859 **ACCEPTANCE**

- 0.000136 **EXPENDITURES**

- 0.163 **TOP10% HS**

### Conclusion:

* SAT score has a positive correlation with graduation rate
* **The remaining variables** has a negative correlation with graduation rate
* R-squared: R^2 = 0.5344 indicates that 53.44% of the variation in graduation rate is explained by the independent variables (SAT score, acceptance rate, expenditure per student, top 10% HS ratio).
* Assessing the Model's Reliability:
* P-values: The p-values for all independent variables are less than 0.05, indicating that the relationships between these variables and graduation rate are statistically significant.
* F-statistic: The F-statistic value is 6.33E-07, which is less than 0.05, indicating that the regression model is statistically significant.
* General Comments:
* The data shows that SAT score is positively correlated with graduation rate, while acceptance rate, expenditure per student, and top 10% HS ratio are negatively correlated with graduation rate.
* The regression model has a high reliability with R^2 = 53.44% and p-values less than 0.05.
* Limitations:
* This model is based on data from only 49 schools. Therefore, the results may not be representative of the entire population of schools.
* This model cannot explain 46.56% of the variation in graduation rate.
* Recommendations:
* More data needs to be collected from more schools to improve the accuracy of the model.
* Further research is needed to identify other factors that may affect graduation rate.
* Applications:
* This model can be used to predict the graduation rates of schools based on SAT score, acceptance rate, expenditure per student, and top 10% HS ratio.
* This model can be used to identify the factors that affect graduation rate and develop appropriate policies to improve graduation rate.

### Using MS Excel, R language and Python language to perform Multivariable Nonlinear Regression with optional real data about/of Vietnam.

First, we extracted 34 rows of data from the [Automobile dataset (kaggle.com)](https://www.kaggle.com/datasets/qucvinhdng/automobile-dataset). We use multivariate nonlinear regression to find an equation that fits the data set. The problem is to apply Vietnam's data set, specifically Pickup Car Prices, to answer the question of whether car prices are affected by measured factors or not. In this case, we choose the measured factors as the number of kilometers traveled (Km) and the vehicle's cylinder capacity (cylinder capacity).



### Analyzing by using Excel:

Dataset:

A screenshot of a table

Description automatically generated

Using Analys Tool:

A screenshot of a computer

Description automatically generated

The result:

A screenshot of a graph

Description automatically generated

### Analyzing by using R:

Read and check data:

A screenshot of a computer

Description automatically generated

**Result:**

**A close up of a number

Description automatically generated**

**A diagram of a graph

Description automatically generated**

**A diagram of a graph

Description automatically generated**

### Analyzing by using Python:

Read Data :

A screenshot of a computer

Description automatically generated

Train data:  
A screenshot of a computer program

Description automatically generated

Result:

*y = 2.9672 \* x^2 - 377.0203 \* x + 1368.4431*

A graph of data on a grid

Description automatically generated

### Conclusion:

The fact that Python, R, and Excel all give similar results shows the consistency and reliability of this linear regression model. This shows some of the following points:

* Objectivity and Trust: When different software/tools, developed by different teams, use different algorithms, but still produce similar results, this increases the reliability of the results. It shows that the results are not affected by the error of a particular software/algorithm.
* Unity of theory and practice: The fact that different software all apply the same principle/theory of linear regression and obtain similar results proves that there is unity between theory and practice in the field of regression analysis.
* Reproducibility: The same results on different platforms indicate the reproducibility of regression analysis. This is important in scientific research and data analysis.
* Cross-checking: The comparison of results between tools helps to cross-check each other, detect errors if any, enhance reliability on the final result.

In summary, the similarity results between Python, R, and Excel for this regression model is a good indication of the accuracy, objectivity, and reliability of the analysis results. This strengthens confidence in the conclusions drawn from this regression model.

Based on the provided documents, I can draw the following conclusions about the given model:

* This appears to be a multiple linear regression model with two predictor variables.
* The model is trying to predict some quantity (likely "price" based on the data) using two predictor variables: "Km" (kilometers) and "Cylinder Capacity".
* The coefficients of the model are:
* Coefficient a (intercept) = 1368.4431
* Coefficient b (for Km) = -377.0203
* Coefficient c (for Cylinder Capacity) = 2.9672
* The model has a multiple R value of 0.639363237, which indicates a moderate positive correlation between the predictor variables and the response variable.
* The R-squared value is 0.408785349, meaning that about 40.9% of the variation in the response variable can be explained by the two predictor variables (Km and Cylinder Capacity).
* The adjusted R-squared value, which accounts for the number of predictor variables, is 0.369371039, slightly lower than the regular R-squared.
* The F-statistic (10.37149574) and its associated p-value (0.000376858) suggest that the model as a whole is statistically significant, meaning at least one of the predictor variables has a significant impact on the response variable.
* The regression analysis partitions the total sum of squares into a portion explained by the regression (210599.25) and a residual portion (304583.7197).
* The data includes both real and predicted values, plotted in a 3D scatter plot, allowing for visual inspection of the model's fit.

Overall, this appears to be a moderately effective regression model for predicting the response variable based on Km and Displacement, but there is still a significant amount of variation because of other factors at play. like brand, color, number of seats, ...

### Using MS Excel, R language and Python language perform Logistic Regression with optional real data about/of Vietnam.

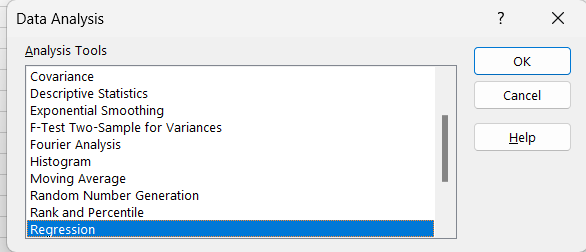


### Analyzing by using Excel:

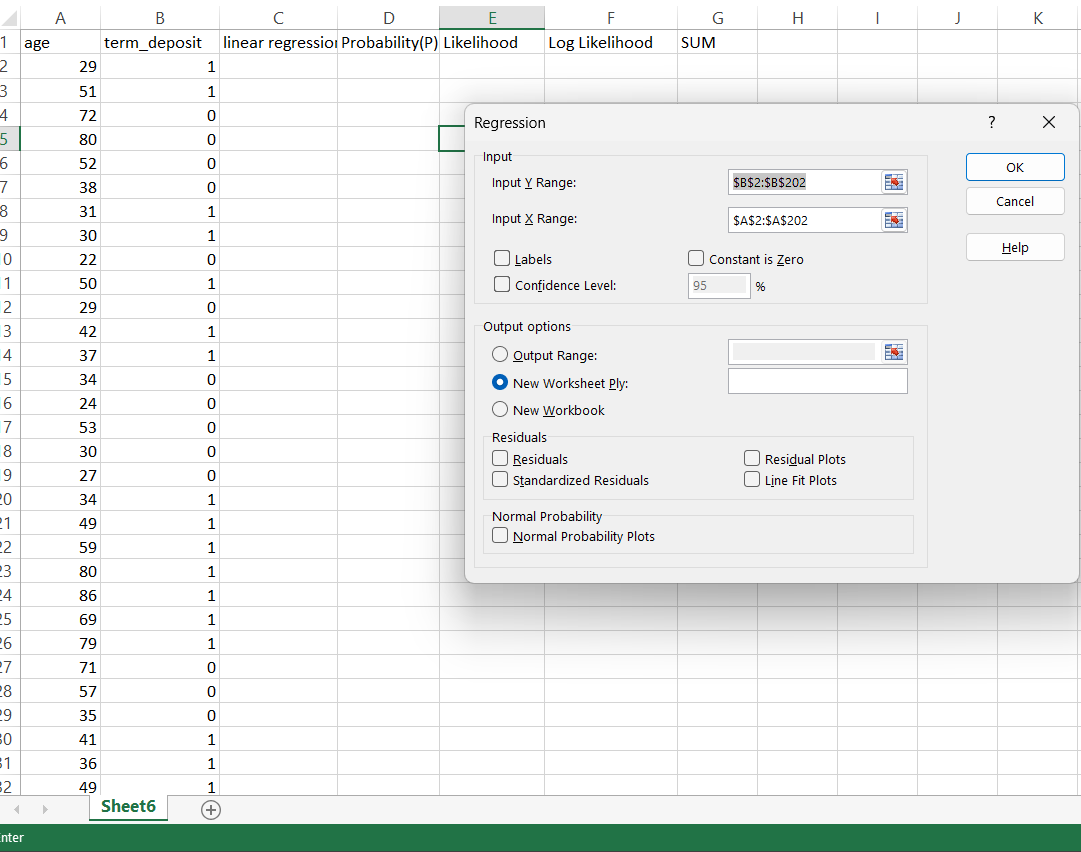
Get the first 100rows of BankCustomerData.csv and perform Logistic Regression on the dataset to predict the likelihood of the client will subscribe a term deposit based on age in Viet Nam.

Link dataset: [Bank Customer Data in VietNam | Kaggle.](https://www.kaggle.com/datasets/tomculihiddleston/bank-customer-data-in-vietnam)

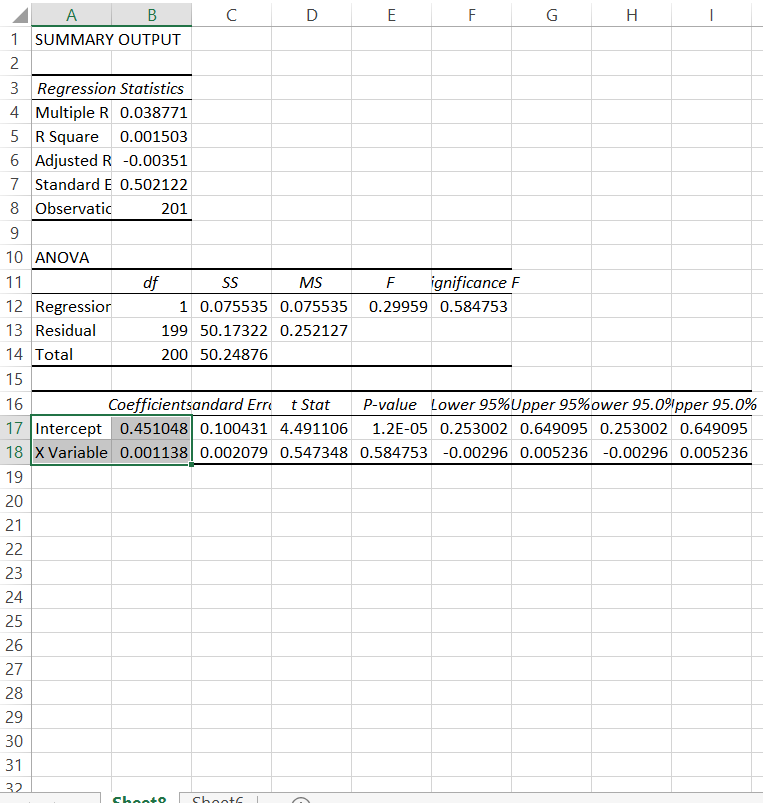
Step 1: Choose Data Anays from Data tab to find Linear Regression:



Step 2: Choose Regression Tool, select Input range for X(independent variable) and Y(out come) and press OK:

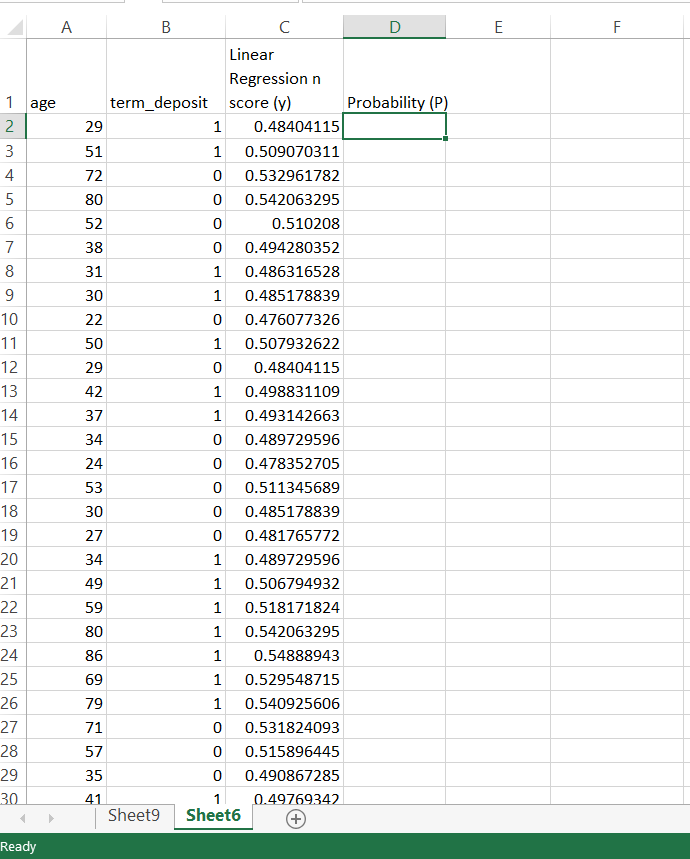


Step 3: Get the Intercept(β0) and X Variable 1(β1) to form the logit:

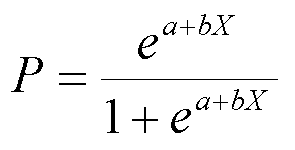


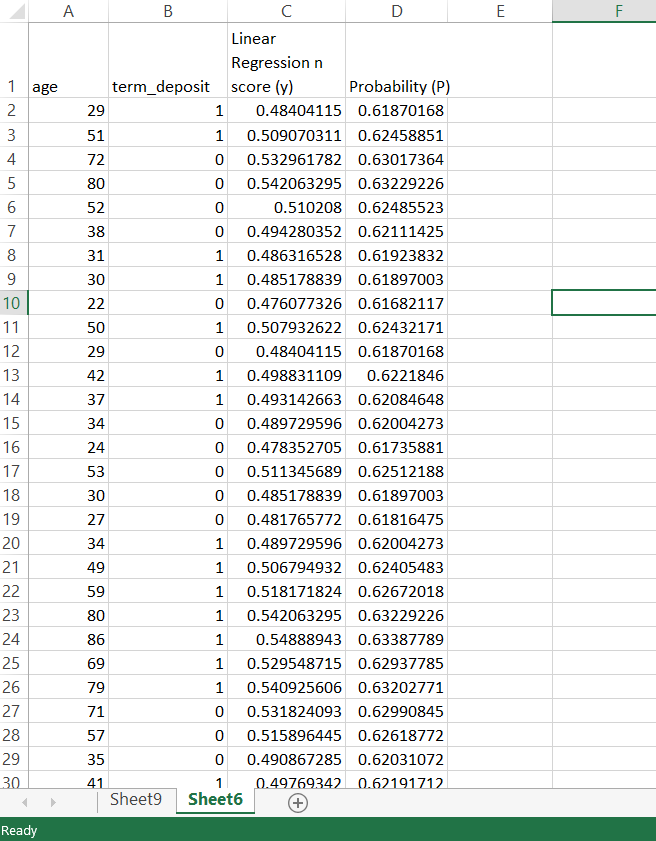
Step 4: Calculate Linear Regression n score (y):

The formula: Y = β0 + β1\*X



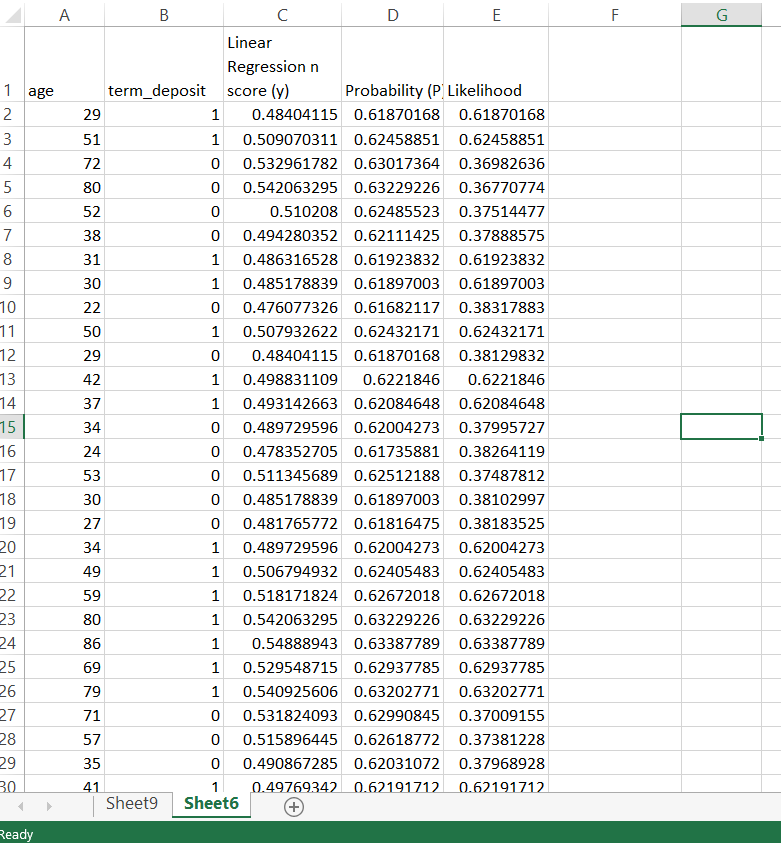
Step 5: Calculate the Probability:



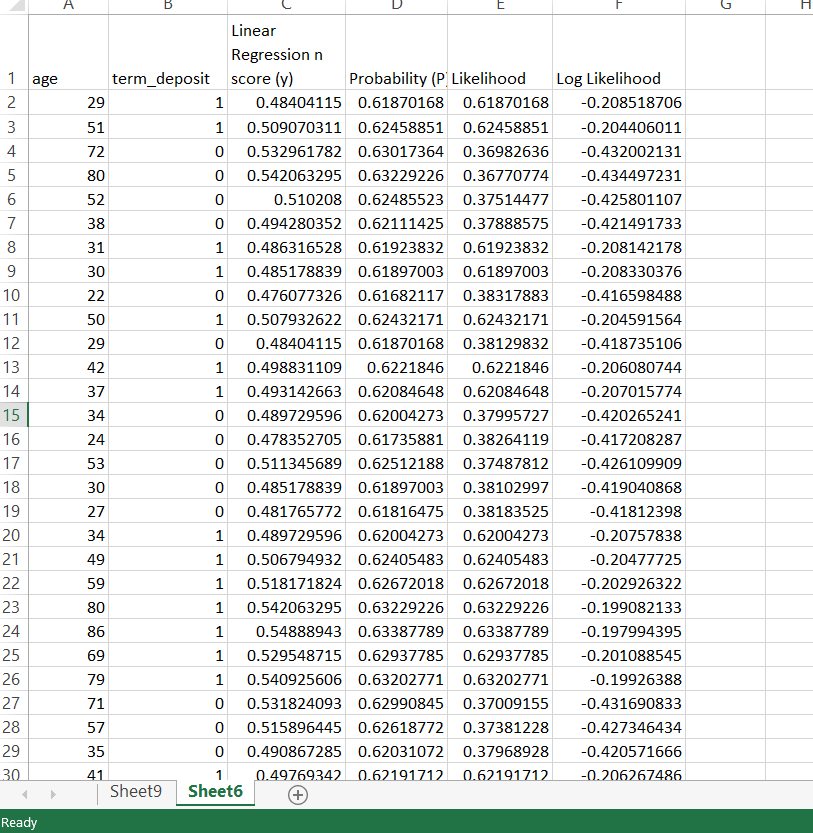


Step 6: Calculate the Likelihood:

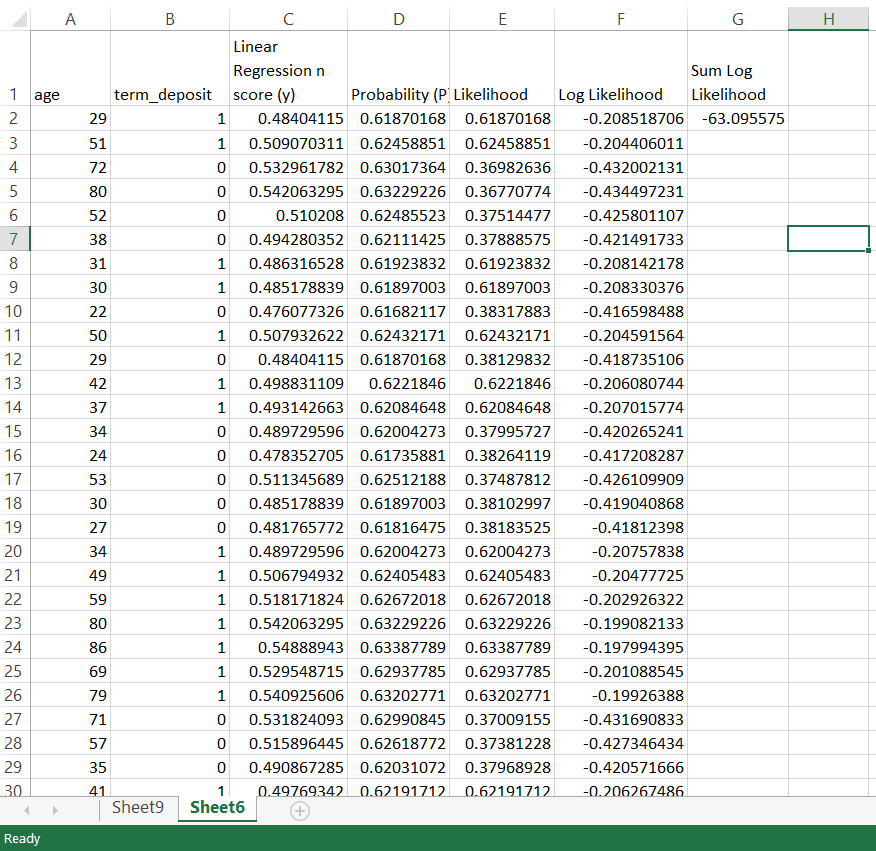
We use IF(), if the housing loan is true, the Likelihood = P, or else Likelihood = 1 – P.



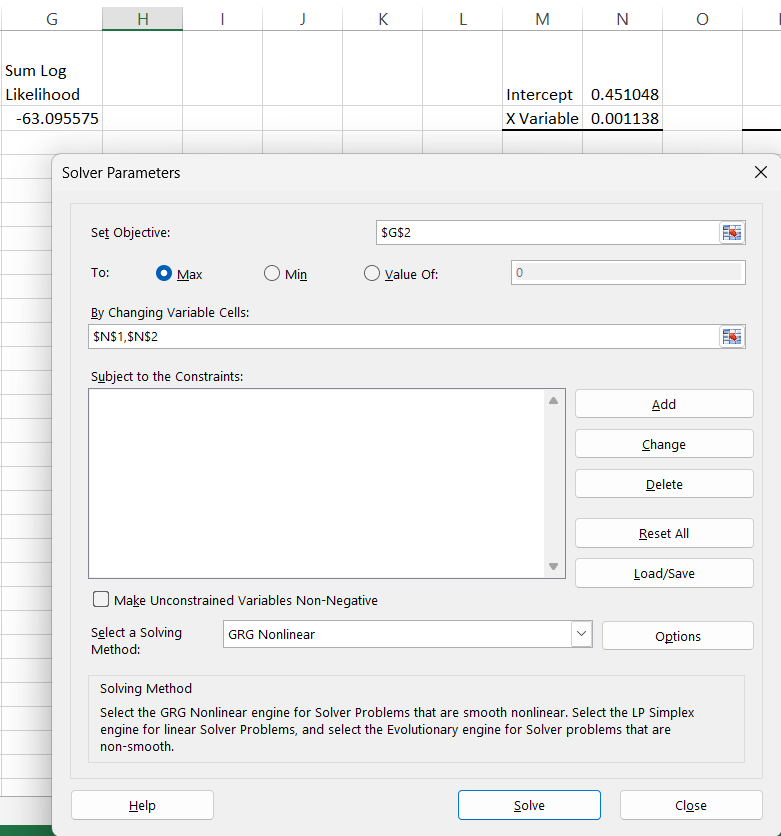
Step 7: Calculate Log(Likelihood):



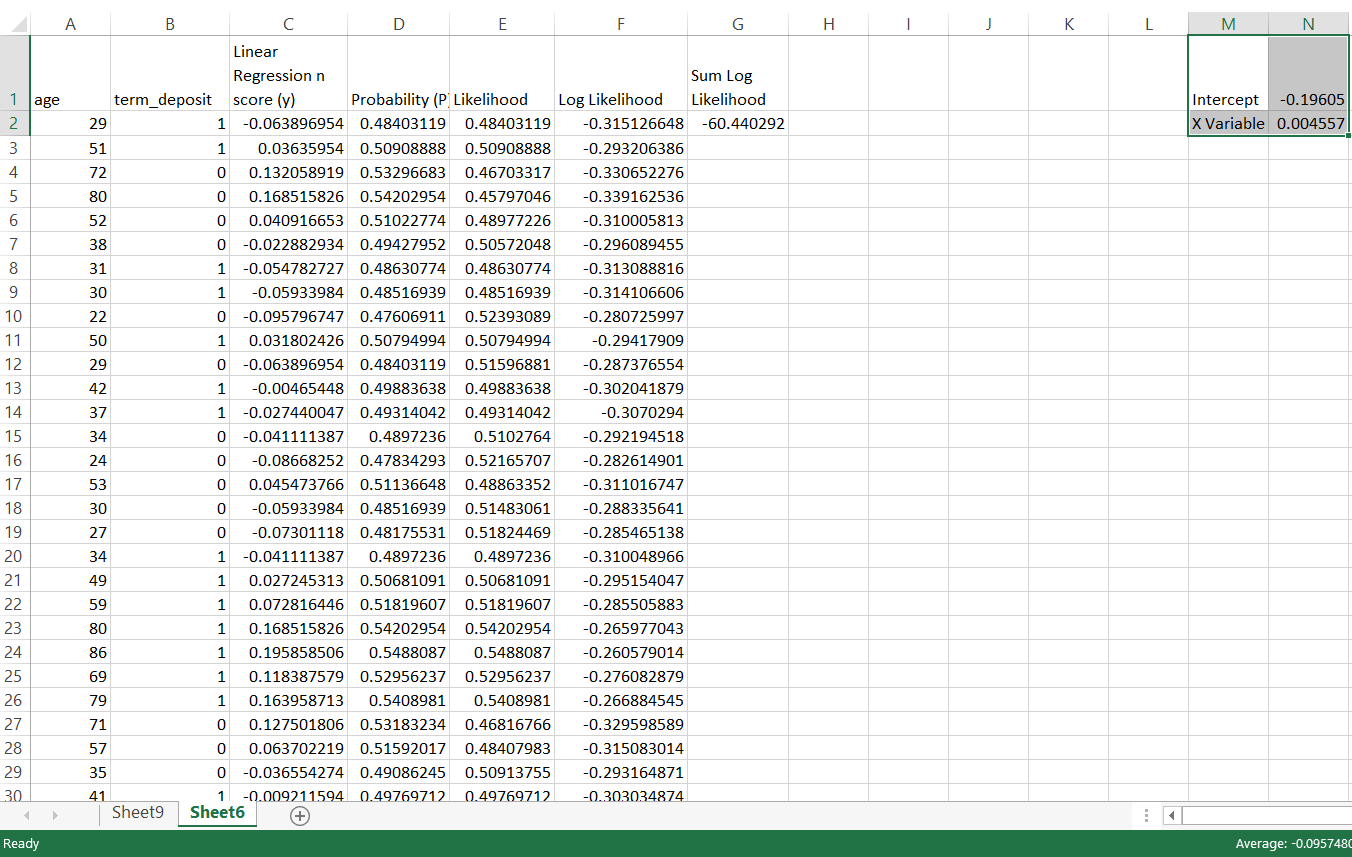
Step 8: Calculate sum of log(Likelihood):



Step 9: Using Solver in Data tab of Excel to find the β0 and β1:

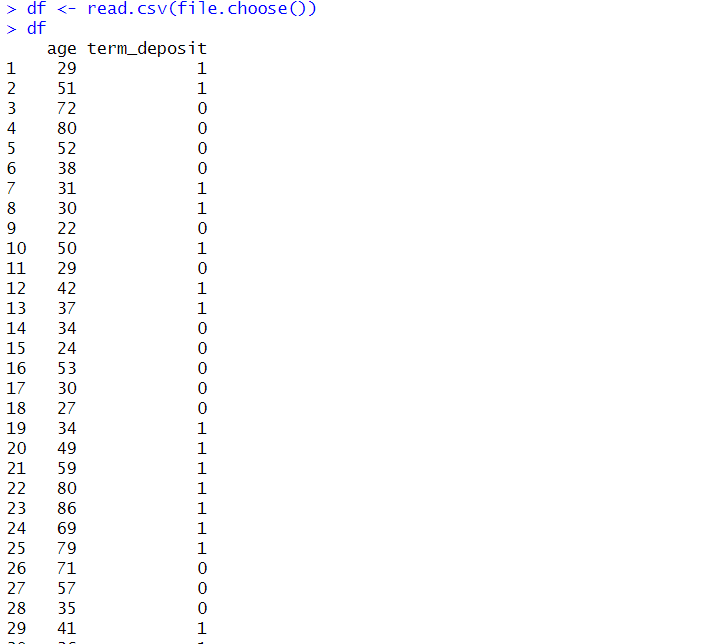


Step 10: Get the result:



### Analyzing by using R:

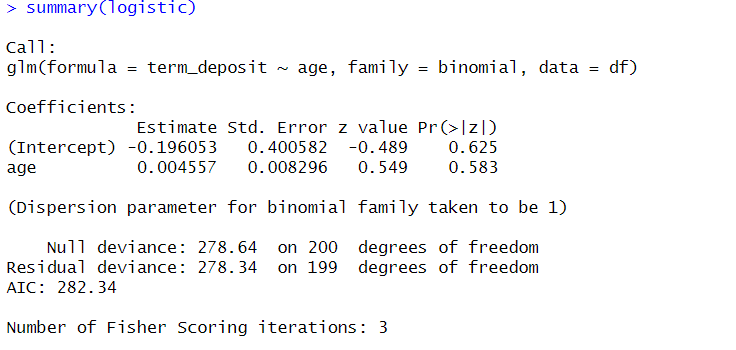
Step 1: Import the dataset:



Step 2: Fit generalized linear models (GLMs), specifies the formula for the model. It means that term\_deposit is the dependent variable, and age is the independent variable being used to predict term\_deposit.

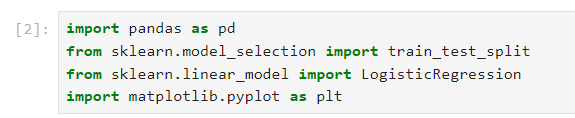


Step 3: View the result using summary():

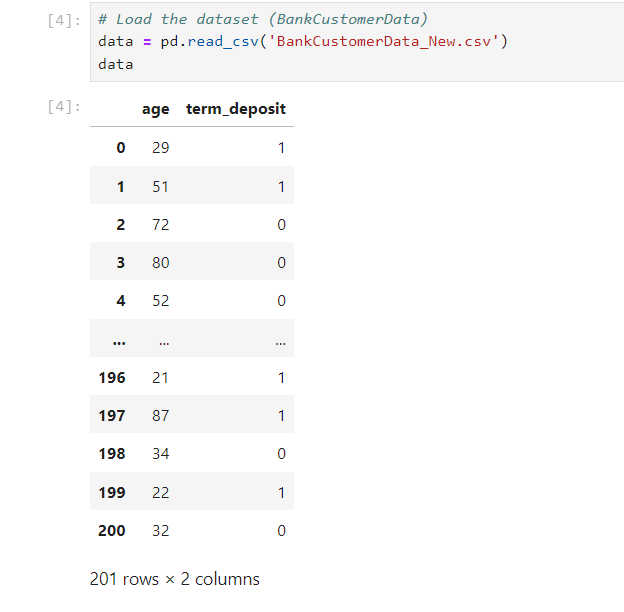


### Analyzing by using Python:

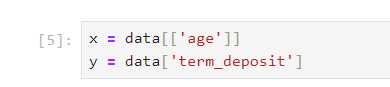
Step 1: Import the libraries:



Step 2: Load the dataset:



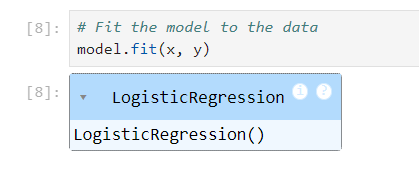
Step 3: Split the dataset into features (x) and target variable (y):



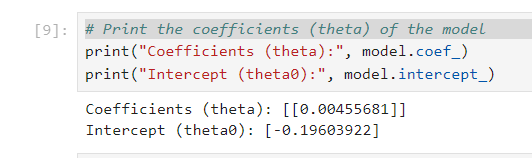
Step 4: Create a Logistic Regression model



Step 5: Fit the model to the data:

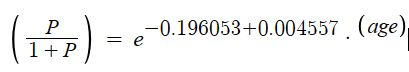


Step 7: Print the coefficients (theta) of the model:



### Conclusion:

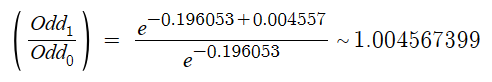
Based on the result, we can form a regression:



We set (P / 1 – P ) is Odd.







We can say: when the age increases by 1, the likelihood of subscribing to a term deposit increases ~1.004567399 time(s).

Overall, based on these results, age alone may not be a strong predictor of term deposit subscription in this logistic regression model. It's essential to consider other variables or interactions that might better explain the variation in term deposit subscriptions. Additionally, further analysis or refinement of the model may be necessary to improve its predictive performance.

## REFERENCES:

1. Wooldridge, J. M. (2015). Introductory econometrics: A modern approach (pp. 67-145). Nelson Education.
2. James R. Evans, *Business Analytics*: *Methods, Models, and Decisions*, 1st edition, Pearson Education,2019
3. [David F. Groebner](http://www.amazon.com/s/ref=ntt_athr_dp_sr_1?_encoding=UTF8&field-author=David%20F.%20Groebner&search-alias=books&sort=relevancerank), [Patrick W. Shannon](http://www.amazon.com/s/ref=ntt_athr_dp_sr_2?_encoding=UTF8&field-author=Patrick%20W.%20Shannon&search-alias=books&sort=relevancerank), [Phillip C. Fry](http://www.amazon.com/s/ref=ntt_athr_dp_sr_3?_encoding=UTF8&field-author=Phillip%20C.%20Fry&search-alias=books&sort=relevancerank), *Business Statistics: A Decision-Making Approach* (9th Edition), Prentice-Hall, Inc. 2012
4. Gerhard Bohm, Günter Zech, *Introduction to Statistics and Data Analysis*, Verlag Deutsches Elektronen, 2018
5. Seber, G. A. F., & Wild, C. J. (2003). Nonlinear regression. Hoboken, NJ: Wiley-Interscience.
6. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Hoboken, NJ: Wiley.
7. LAB03\_THONGKEHOIQUY\_SLIDES-1.pdf
8. Lecture 6 - Thống kê hồi quy ALL.pdf
9. Chapter 9: Regression Analysis (Business Analytics: Methods, Models, and Decisions, 1st edition - James R. Evans)