VIETNAM NATIONAL UNIVERSITY

UNIVERSITY OF INFORMATION TECHNOLOGY

INFORMATION SYSTEMS FACULTY

REPORT LAB 3

SUBJECT: DATA ANALYSIS IN BUSINESS

Lecturer: Assoc. Prof. Nguyen Dinh Thuan

Instructor: TA. Nguyen Minh Nhut

Class: IS403.O22.HTCL

Group 3:

21521049 – Ho Quang Lam

21521586 – Le Thi Le Truc

21521938 – Nguyen Thanh Dat

Ho Chi Minh City, April 2024

ACKNOWLEDGEMENT

First of all, we would like to express our deepest gratitude and appreciation to our lecturers, Mr. Nguyen Dinh Thuan and Mr. Nguyen Minh Nhut, for their teaching and sharing of extensive knowledge as well as practical examples during the lectures. They have guided us in completing our Lab 01 report by providing valuable feedback, suggestions, and assistance with exercises and revisions.

The Data Analysis in Business course is an interesting and highly practical subject. However, due to our limited expertise and initial unfamiliarity with real-world applications, we acknowledge that our Lab 01 report may contain some shortcomings and inaccuracies despite our best efforts. We sincerely hope to receive further guidance and feedback from Mr. Nguyen Dinh Thuan and Mr. Nguyen Minh Nhut to improve our knowledge and equip ourselves for future projects as well as for our academic and professional endeavors.

Once again, we would like to extend our heartfelt and sincere gratitude to our lecturers and peers.

Ho Chi Minh City, April 2024

Group of student performers

Ho Quang Lam

Le Thi Le Truc

Nguyen Thanh Dat

LECTURER'S COMMENTS

TABLE OF CONTENTS

WORK DISTRIBUTION

Works \ Members	Le Thi Le Truc (Leader)	Ho Quang Lam	Nguyen Thanh Dat
Problem statement	✓	✓	✓
Build the report template	✓		
Do all task 1a		✓	
Do all task 1b	✓		
Do all task 1c			✓
Do all exercise with Excel in task 2	✓		
Do all exercise with R in task 2			✓
Do all exercise with Python in task 2		✓	
Summarize and edit reports	✓	✓	✓
Completion	**100%**	**100%**	**100%**

Chapter 1. EXPLANATION AND EXAMPLE

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

a) Multivariable Linear Regression.

b) Multivariable Nonlinear Regression

c) Logistic Regression

1.1. Multivariable Linear Regression

1.1.1. What is Multivariable Linear Regression?

Regression simply means finding the relationship of a variable, or a vector that depends (Y) on the independent set X = {X1, X2, ..., Xn} by a particular means

Linear regression model is an analysis tool in statistics and machine learning, are used to predict the value of dependent variable based on independent variables. It is based on the assumption that the relationship between the variables is linear.

1.1.2. How does Multivariable Linear Regression work?

Multivariate linear regression is a method that describes the relationship between dependent variables independent variables X1, X2,.....Xp and error term ε

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon}$$

In which:

- **Y:** dependent variable
- **X1, X2, Xp:** independent variables
- **β0:** regression constant (also known as intercept coefficient). This is an index that shows the value of. What would Y be if all X were equal to 0 (no X) when performing on map represents Oxygen, β0 is the point on the Oy axis that the regression line crosses.

$$b_0 = \overline{Y} - b_1 \overline{X}$$

- **β1, β2, βp:** regression coefficient, also known as slope coefficient. This indicator gives us know about the change in Y caused by the corresponding X.

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}$$

- **ε:** error. The larger this index is, the worse the regression's predictive ability becomes should be less accurate or more misleading than reality.

1.1.3. Why use Multivariable Linear Regression?

Multivariable linear regression is a statistical technique that allows us to analyze the relationship between multiple independent variables and a dependent variable. It offers several advantages in data analysis and modeling.

First, it enables us to examine the individual contributions of each independent variable to the dependent variable while controlling for other variables. By considering multiple factors simultaneously, we can better understand the unique impact of each variable on the outcome of interest.

Second, multivariable linear regression can be used for prediction. By fitting a regression model to historical data, we can make informed estimates or forecasts of the dependent variable's values based on the values of the independent variables. This predictive capability is valuable in fields where future projections are essential for decision-making.

Third, multivariable linear regression helps us assess the strength and direction of the relationships between independent variables and the dependent variable. By examining the estimated coefficients, we can determine which variables have significant associations with the outcome. This information aids in identifying key factors or variables that influence the dependent variable.

1.1.4. Example

A data collection software company a sample of 20 programming members. People recommend using regression analysis to determine whether salary is associated with years of experience and capacity about the city organization installer?

Number of years of experience, competency test score and qualified annual salary ($1000) of 20 installations users are presented in the following table:

	A	B	C	D	E
1	Programmer	b0	Experience	Score	Salary
2	1	1	4	7.8	24
3	2	1	7	10	43
4	3	1	1	8.6	23.7
5	4	1	5	8.2	34.3
6	5	1	8	8.6	35.8
7	6	1	10	8.4	38
8	7	1	0	7.5	22.2
9	8	1	1	8	23.1
10	9	1	6	8.3	30
11	10	1	6	9.1	33
12	11	1	9	8.8	38
13	12	1	2	7.3	26.6
14	13	1	10	7.5	36.2
15	14	1	5	8.1	31.6
16	15	1	6	7.4	29
17	16	1	8	8.7	34
18	17	1	4	7.9	30.1
19	18	1	6	9.4	33.9
20	19	1	3	7	28.2
21	20	1	3	8.9	30

Suppose we believe that the annual salary (y) is related to number of years of experience (x1) and performance test scores aptitude (x2) according to the following regression model:

$$y = b0 + b1x1 + b2x2 + e$$

With

y: Annual salary

x1: Number of years of experience

x2: Aptitude test score

Regression results by tool:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.913334059
R Square	0.834179103
Adjusted R Square	0.814670762
Standard Error	2.418762076
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	500.3285303	250.1643	42.76013	2.32774E-07
Residual	17	99.45696969	5.85041		
Total	19	599.7855			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	**3.17393627**	6.156066829	0.515579	0.612789	-9.814229424	16.16210196	-9.814229424	16.16210196
Experience	**1.403902485**	0.198566912	7.070173	1.88E-06	0.984962921	1.822842049	0.984962921	1.822842049
Score	**2.508854478**	0.77354127	3.243336	0.00478	0.876825058	4.140883899	0.876825058	4.140883899

Regression results by step – by – step:

XT	1	1	1	1	1	1
	4	7	1	5	8	10
	7.8	10	8.6	8.2	8.6	8.4
XT*X	20	104	165.5			
	104	708	875			
	165.5	875	1380.53			
(XT*X)-1	6.477692832	0.037845783	-0.80054			
	0.037845783	0.006739497	-0.00881			
	-0.80054271	-0.0088086	0.102278			
(XT*X)-1*XT	0.384842846	-1.262813761	-0.36913	0.102472	-0.104208186	0.131591922
	-0.00390331	-0.003063741	-0.03117	-0.00069	0.016007796	0.031248509
	-0.03801156	0.16057344	0.070236	-0.00591	0.008576151	-0.029496576
(XT*X)-1*XT*Y	**3.17393627**					
	1.403902485					
	2.508854478					

Regression equation estimate:

SALARY = 4.09 + 0.499(EXPER) + 2.894(SCORE)

- **F test:**

Hypothetical:

H0 : b1 = b2 = 0

Ha : There is at least 1 non-zero bi parameter

Rejection rule:

With a = 5% and Degrees of Freedom are 2 and 17

Look up table F.05 = 3.59

F = MSR/MSE = 250.16/5.85 = 42.76

Reject H0 if p-value < .05 or F > 3.59

Conclusion: p-value < .05, so H0 can be rejected (also, F = 42.76 > 3.59)

⇨ Concluding that at least one of the independent variables has a significant significance effect on the dependent variable in the regression model.

- **t Test:**

Hypothetical:

H0 : bi = 0

Ha : bi # 0

Conclusion:

p-value < 0.05 => Reject both H0: b1 = 0 and H0: b2 = 0

⇨ Both independent variables (Experience and Score) are significant

1.2. Multivariable Nonlinear Regression

1.2.1. What is Multivariable Nonlinear Regression?

Multivariate nonlinear regression is a method of determining the nonlinear relationship between a dependent variable and multiple independent variables

1.2.2. How does Multivariable Nonlinear Regression work?

The multivariate nonlinear regression model can be represented by a nonlinear function:

$$Y = f(X1, X2, ..., Xn) + \varepsilon$$

In there:

Y: dependent variable

X1, X2,..., Xn: independent variables

f: a nonlinear function

ε: random error

Non-linear regression functions include exponential functions, logarithmic functions, trigonometric functions, and exponential functions factorial, Gauss function, and Lorenz curve

1.2.3. Why use Multivariable Nonlinear Regression?

When there are multiple independent variables affecting the dependent variable, multiple nonlinear regression variables can be used to evaluate the impact of each independent variable on the dependent variable.

In some cases, the relationship between variables is not a relationship simple linear. Multivariate nonlinear regression can be used to model termites this complex relationship.

Minimize error: When there are many independent variables, multivariate nonlinear regression can help minimizes error compared to a univariate linear regression model

1.2.4. Example

Problem request: The influence of house area and number of bedrooms on the price of the house

area	rooms	price_usd
38	1	33,333
40	1	51,316
42	1	24,123
42	1	78,947
46	1	46,009
50	1	4,386
50.17	1	118,421
54	1	76,754
40	2	48,246
48	2	70,175
50	2	60,526
50	2	62,281
50	2	65,789
50	2	65,789
50	2	65,789
50	2	71,491
50	2	72,368
50	2	74,123

Sample regression model:

$$\text{Price} = \beta 0 + \beta 1 * log10(\text{area}) + \beta 2 * log10(\text{rooms})$$

We use the tool in Excel to calculate the following results:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.6169008
R Square	0.3805667
Adjusted R Square	0.3798984
Standard Error	99736.141
Observations	1857

ANOVA

	df	SS	MS	F	ignificance F
Regression	2	1.1331E+13	5.6653E+12	569.529	2E-193
Residual	1854	1.8442E+13	9947297799		
Total	1856	2.9773E+13			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	**-995828.7**	37361.805	-26.6536558	9E-133	-1069104	-922553	-1069104	-922553
log10(area)	**643464.05**	22834.9205	28.1789484	1E-145	598679	688249	598679	688249
log10(room)	**-177970.2**	23052.0794	-7.72035361	1.9E-14	-223181	-132759	-223181	-132759

Or we calculate coefficisigmodent step by step:

XT	1	1	1	1	1
	1.579783597	1.60206	1.62324929	1.62324929	1.66276
	0	0	0	0	0
XT*X	1857	3411.3759	549.083788		
	3411.375851	6307.2515	1037.79334		
	549.0837884	1037.7933	202.027986		
(XT*X)-1	0.140330018	-0.084925	0.05485346		
	-0.084925261	0.0524196	-0.03845829		
	0.054853455	-0.038458	0.05342138		
(XT*X)-1 * XT	0.006166483	0.0042747	0.00247515	0.00247515	-0.00088
	-0.002113602	-0.000946	0.00016485	0.00016485	0.00224
	-0.005902319	-0.006759	-0.00757394	-0.00757394	-0.00909
(XT*X)-1 * XT*Y	**-995828.6909**				
	643464.0462				
	-177970.2045				

⇨ From the above results we can draw the following conclusion:

Regression equation:

$$\text{Price} = -\,995828 + 643464 * log10(\text{area}) - 177970 * log10(\text{rooms})$$

1.3. Logistic Regression

1.3.1. What is Logistic Regression?

Logistic regression analysis is a statistical technique to examine the association between variables independent (numerical or categorical variable) with the dependent variable being a binary variable

1.3.2. How does Logistic Regression work?

Logistic icon recovery method:

$$log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X + e$$

Let P be the probability of event A occurring and 1-P be the opposite event of event A

$$ODDs = \frac{P}{1-P}$$

If ODDs > 1, the probability of event A occurring is higher than its counterpart.

If ODDs < 1, the probability of event A occurring is less likely than its counterpart.

If ODDs = 1, the probability of event A occurring is equal to its opposite event

From the above equation we can calculate the probability p according to the X value

$$p(X) = \frac{e^{b_0 + b_1 X}}{1 + e^{b_0 + b_1 X}}$$

In which:

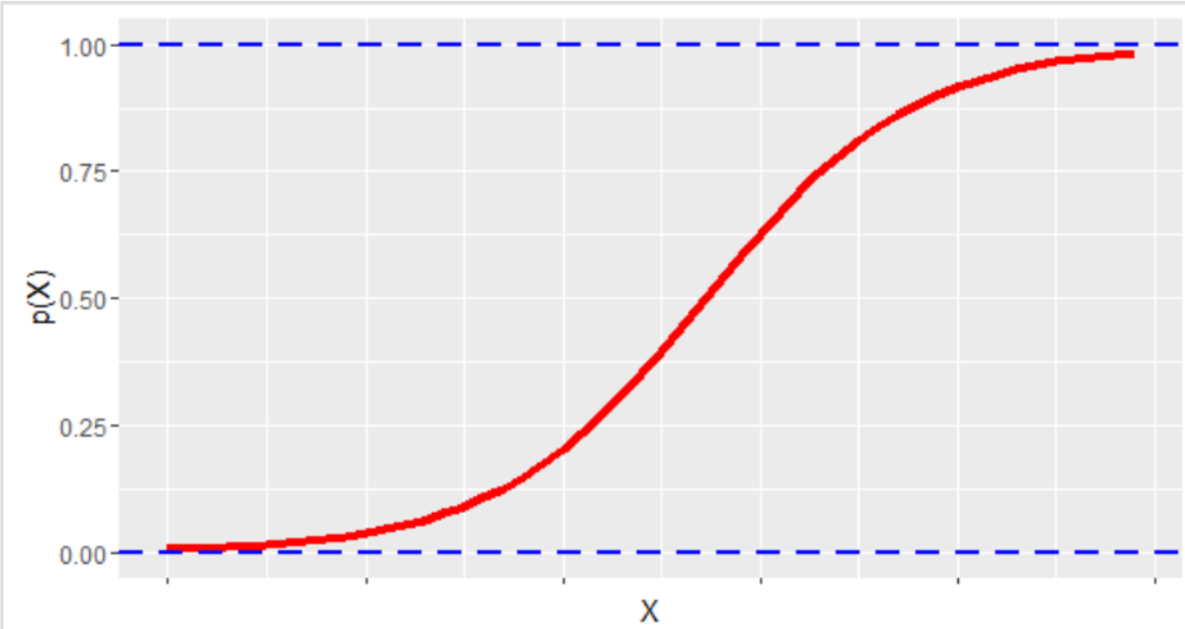
e is the mathematical constant approximately equal to 2.71828 (Euler's number).

B0 and B1 are coefficients representing the intercept and slope, respectively.

x is the predictor variable.

P(X) represents the probability of the binary outcome Y being equal to 1

And the logistic curve has the following form



This is the sigmoid function, which forms an “S” shaped curve and ensures that the output of the logistic regression lies between 0 and 1.

The sigmoid function is used to convert the output of linear regression into probabilities.

1.3.3. Why use Logistic Regression?

Logistic regression is a popular statistical method used in situations where the dependent variable is a binary variable. Here are some reasons why logistic regression should be used:

- Logistic regression models are suitable for binary dependent variables: Logistic regression is ideal for modeling data with a dependent variable that takes only two values, such as yes or no, success or failure. lose. It allows predicting the probability of occurrence of the dependent variable based on the independent variables.

- Effective in identifying influencing factors: Logistic regression helps identify factors that influence the dependent variable. Logistic regression coefficients indicate the influence of each independent variable on the probability of occurrence of the dependent variable, based on analysis of the data sample.
- Determining thresholds and classification: Logistic regression allows determining thresholds to classify observations into different groups based on probability. This could be useful in classifying and predicting outcomes in fields such as medicine, finance and marketing.

1.3.4. Example

A survey of diabetes in area K shows the relationship between age and the likelihood of developing this disease includes 2 columns: column y (diabetes) and column x (age):

Tuổi	Tiểu đường
20	0
23	0
24	0
25	0
25	0
26	0
26	0
28	0
28	0
29	0
30	0
30	0
30	0
30	0
30	0
30	0
32	0
32	0
33	0
33	0
34	0
34	0
34	0
34	0

First we use Excel's tool to calculate:

SUMMARY OUTPUT

Regression Statistics	
Chi Square	56.18577181
Residual Dev.	80.4772109
# of iterations	7
Observations	100

	Coefficients	tandard Err	P-value	Odd Ratio	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Intercept	**-8.876712999**	1.669235	1.05E-07	0.00014	5.3E-06	0.003679	5.3E-06	0.003679
Tuổi	**0.187864621**	0.035347	1.07E-07	1.20667	1.125904	1.29323	1.125904	1.29323

Following all calculations above, Logistic regression of this data's formula:

$$\log\left(\frac{p}{1-p}\right) = -8.877 + 0,188 * (\text{age})$$

Set:

$$\mathbf{Odd(0)} = \left(\frac{p}{1-p}\right) = e^{-8.877} = 0.0001 \text{ when age} = 0$$

$$\mathbf{Odd(1)} = \left(\frac{p}{1-p}\right) = e^{-8.877+0.188} = 0.002 \text{ when age} = 1$$

Then we have the ratio between Odd(1)/ Odd(0) = 2

⇨ At this point we can conclude that for every 1 year increase in age for exercise, the likelihood of diabetes will increase 2 times.

Chapter 2. USE MS EXCEL, R, PYTHON TO PERFORM REGRESSION

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

2.1. Multivariable Linear Regression

Some Colleges and Universities try to predict Student Graduation Rates using a variety of characteristics, such as: Median SAT, Acceptance rate, Expenditures/student, Top 10% of student class.

$$y_n = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k + e$$

In which:

Y is Graduation (%)

X1, …, Xk are the independent (explanatory) variables

b0 is the intercept term

b1, … bk are the regression coefficients for the independent variables

e is the error term

⇨ So, apply the formula: Graduation %= b0 + b1 * Median SAT + b2 *Acceptance rate +b3 * Expenditures/student+ b4 * Top 10% HS

Problem statement: "With confidence 95%, consider if we can find any relationship between the Student Graduation Rate and Median SAT, Acceptance rate, Expenditures/student, Top 10% of HS class or not?"

Dependent variable: Graduation%

Independent variables: Median SAT, Acceptance rate, Expenditures/student, Top 10% of HS class

2.1.1. MS EXCEL

2.1.1.1. Using tool:

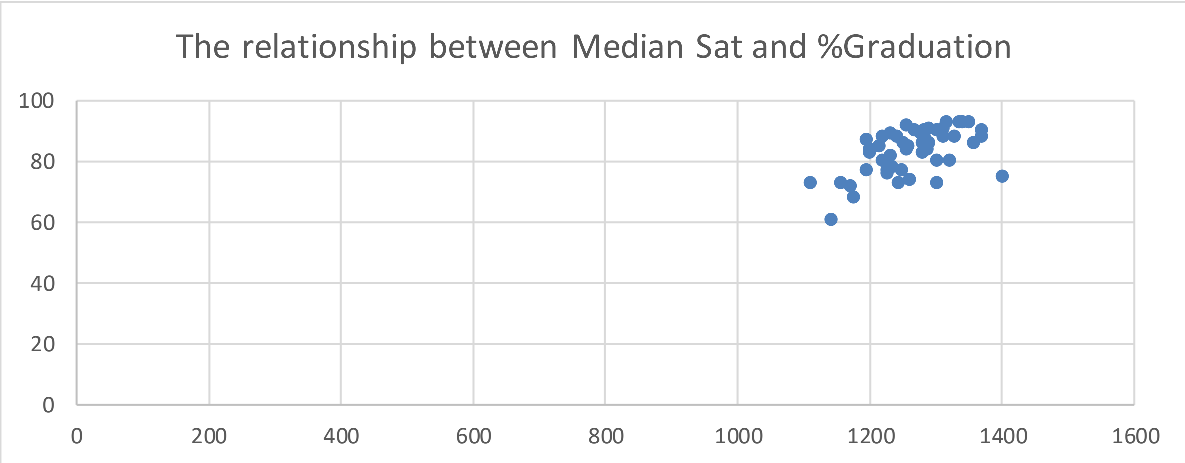
SUMMARY OUTPUT

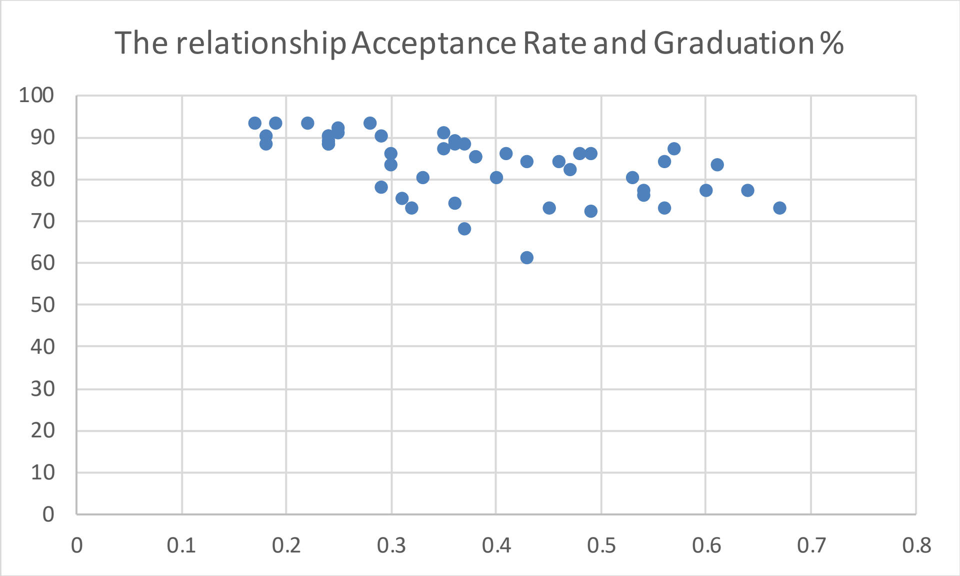
Regression Statistics	
Multiple R	0.731044486
R Square	0.534426041
Adjusted R S	0.492101135
Standard Err	5.30833812
Observations	49

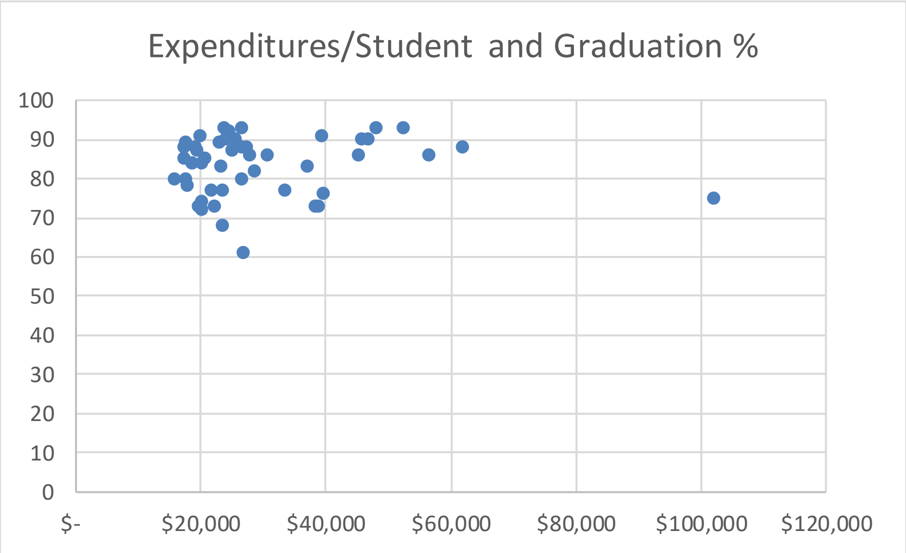
ANOVA

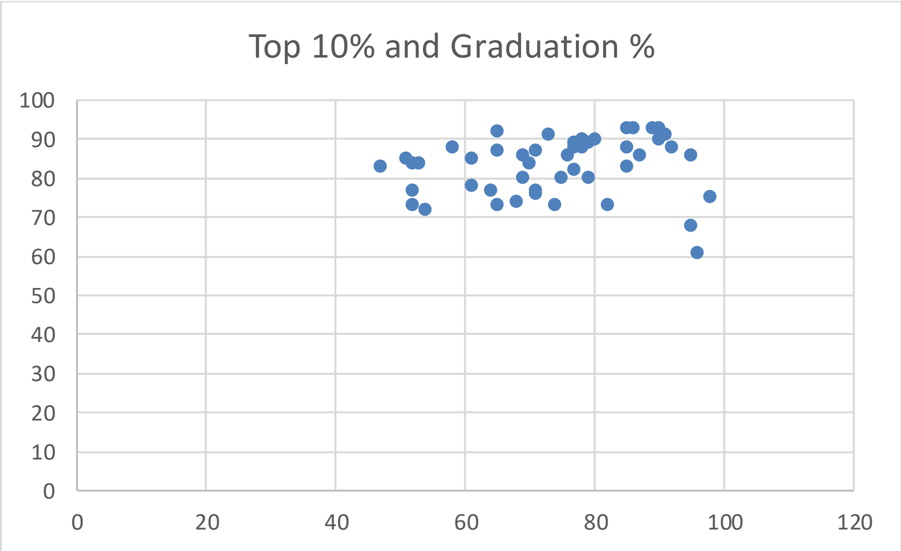
	df	*SS*	*MS*	*F*	*ignificance F*
Regression	4	1423.209	355.8023	12.62675	6.33E-07
Residual	44	1239.852	28.17845		
Total	48	2663.061			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*lpper 95.0%*
Intercept	**17.92095587**	24.55722	0.729763	0.469402	-31.5709	67.41279	-31.5709	67.41279
Median SAT	**0.072006285**	0.017984	4.003927	0.000236	0.035762	0.10825	0.035762	0.10825
Acceptance I	**-24.8592318**	8.315185	-2.98962	0.00456	-41.6174	-8.10108	-41.6174	-8.10108
Expenditures	**-0.00013565**	6.59E-05	-2.05744	0.0456	-0.00027	-2.8E-06	-0.00027	-2.8E-06
Top 10% HS	**-0.162764489**	0.079345	-2.05136	0.046214	-0.32267	-0.00286	-0.32267	-0.00286




The relationship Acceptance Rate and Graduation %
100
90
80
70
60
50
40
30
20
10
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8


Expenditures/Student and Graduation %
100
90
80
70
60
50
40
30
20
10
0
$-
$20,000
$40,000
$60,000
$80,000
$100,000
$120,000


Top 10% and Graduation %
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
120

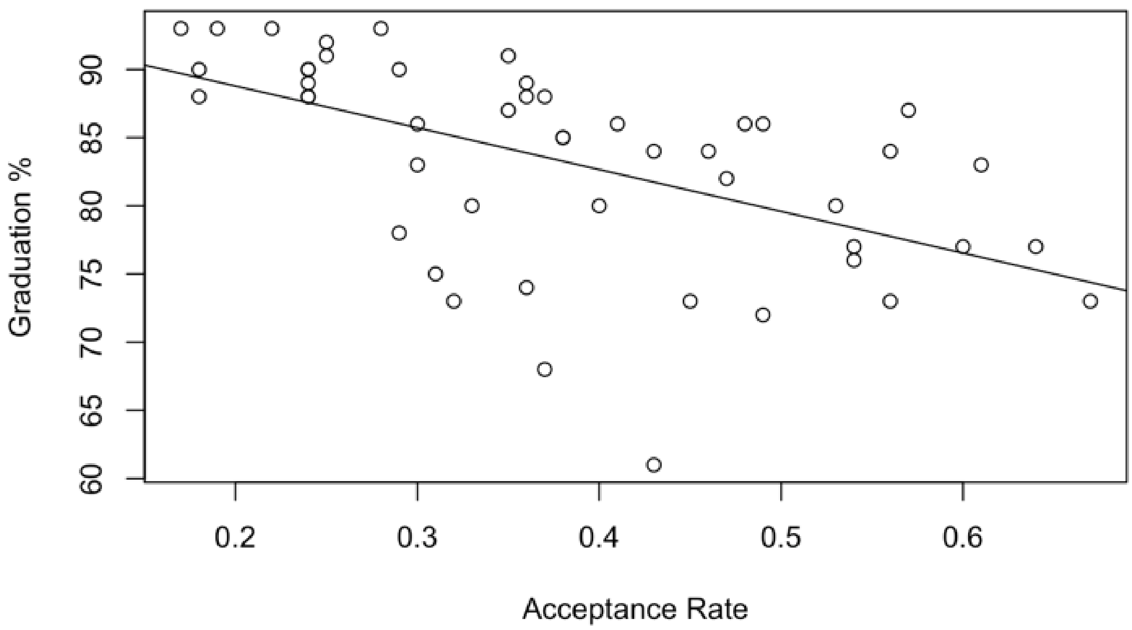
2.1.1.2. Calculating Step-by-step

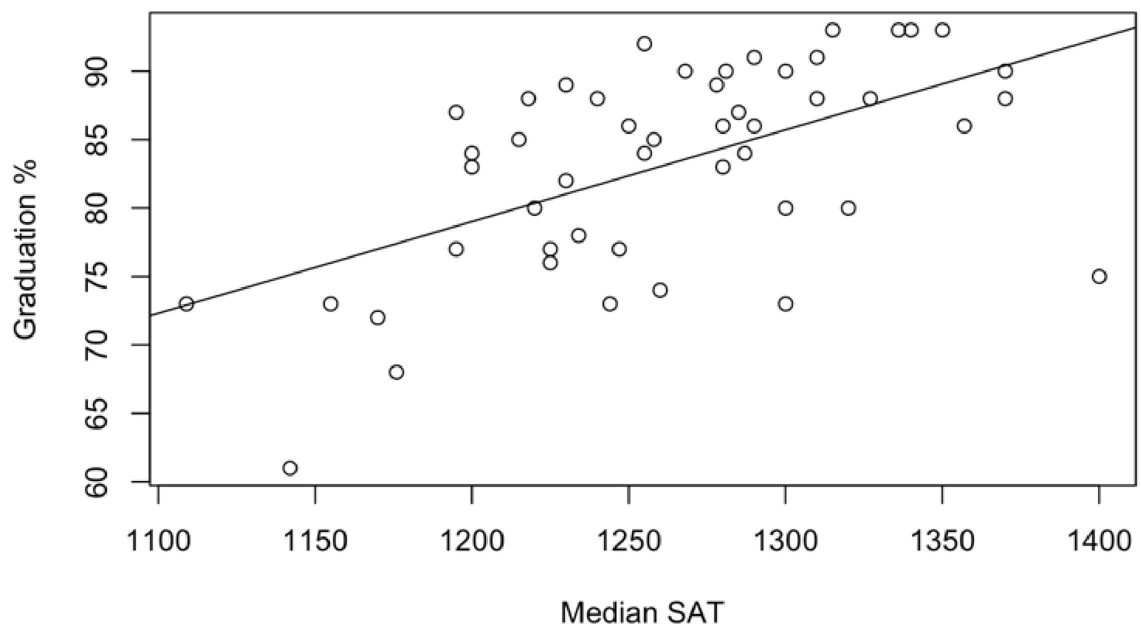
XT	1	1	1	1	1	1	1	1
	1315	1220	1240	1176	1300	1281	1255	1400
	0.22	0.53	0.36	0.37	0.24	0.24	0.56	0.31
	26636	17653	17554	23665	25703	24201	18847	102262
	85	69	58	95	78	80	70	98
XT*X	49	61892	18.67	1472956	3636			
	61892	78364472	23339.98	1.89E+09	4613164			
	18.67	23339.98	7.9719	533014.1	1332.36			
	1472956	1.89E+09	533014.1	5.58E+10	1.14E+08			
	3636	4613164	1332.36	1.14E+08	278620			
(XT*X)-1	21.40136017	-0.01494	-4.95058	2.88E-05	-0.02001			
	-0.014942977	1.15E-05	0.002617	-2.1E-08	9.19E-07			
	-4.950578724	0.002617	2.453729	-5.2E-06	0.011683			
	2.87721E-05	-2.1E-08	-5.2E-06	1.54E-10	-7.2E-08			
	-0.020013406	9.19E-07	0.011683	-7.2E-08	0.000223			
(XT*X)-1*XT	-0.272545944	-0.32589	0.43415	0.776325	-0.03416	0.166511	-0.98307	-0.0725
	0.000254781	0.000146	-7.7E-05	-0.00088	0.000148	-3.8E-05	0.000603	-8.1E-05
	-0.115266332	0.256748	-0.23604	0.021352	-0.18236	-0.20087	0.427408	0.084959
	-1.53952E-06	-1.4E-06	-1.7E-07	-6.4E-07	-9.7E-07	-9.6E-07	-2.2E-06	6.96E-06
	0.000826606	0.001437	-0.00298	0.004901	-0.00045	8.83E-05	0.001957	-0.00062
(XT*X)-1*XT*Y	**17.92095587**							
	0.072006285							
	-24.8592318							
	-0.00013565							
	-0.162764489							

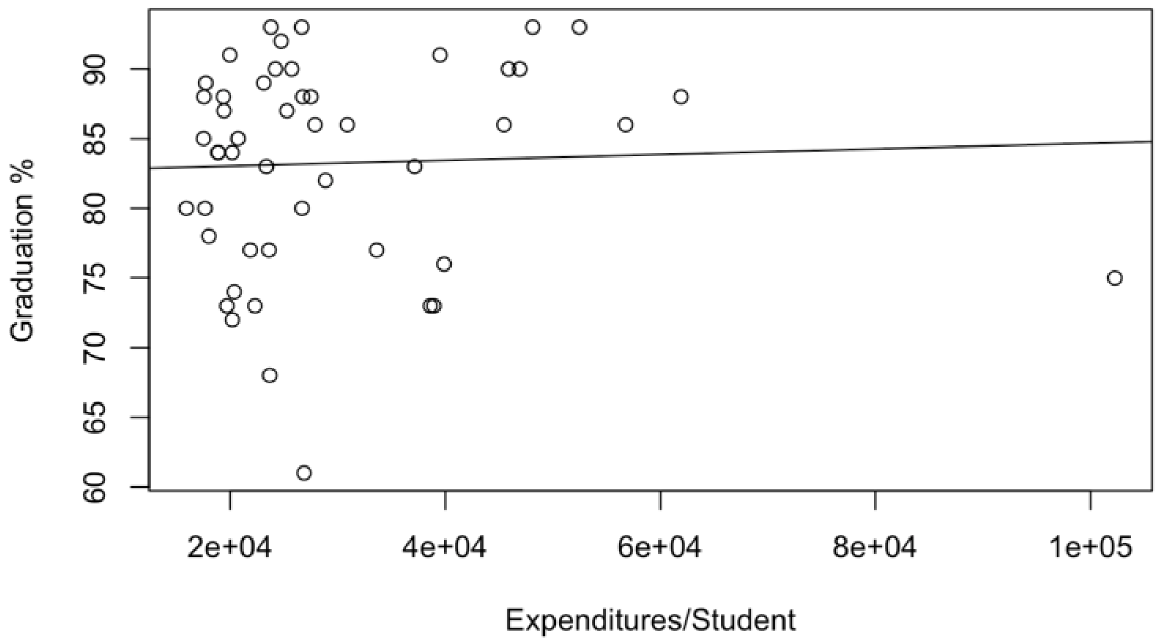
⇨ Conclusion: The result of manual calculation again is the same as the result excel

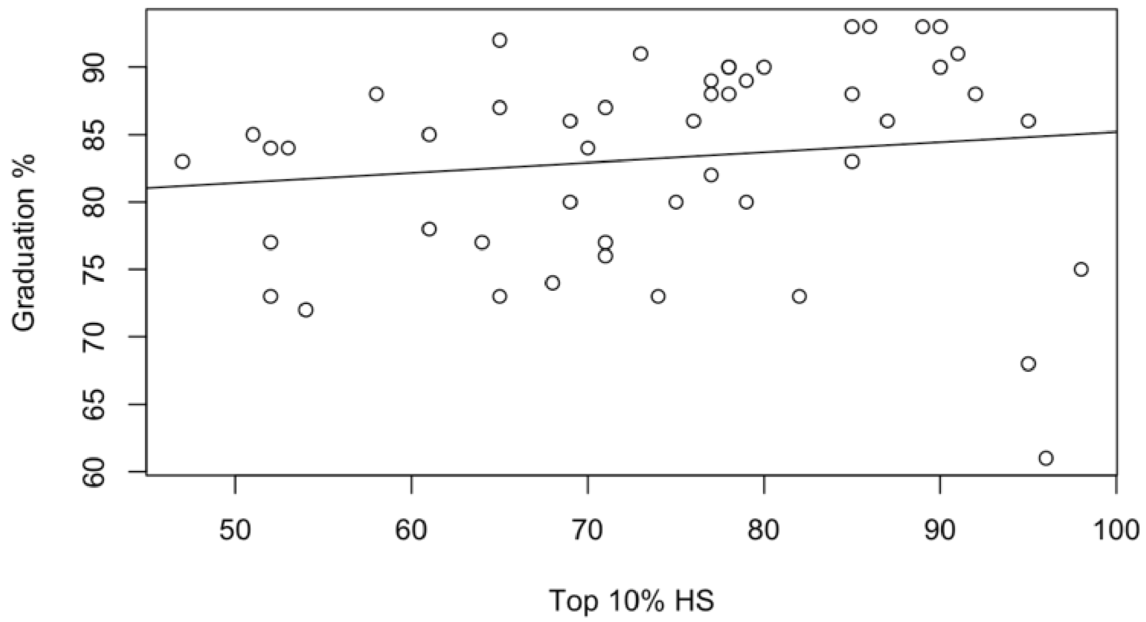
2.1.2. R LANGUAGE

```
> reg1=lm(`Graduation %` ~ `Median SAT` + `Acceptance Rate` + `Expenditures/Student` + `Top 10% HS`)
>
> summary(reg1)

Call:
lm(formula = `Graduation %` ~ `Median SAT` + `Acceptance Rate` +
    `Expenditures/Student` + `Top 10% HS`)

Residuals:
     Min       1Q   Median       3Q      Max
-11.8674  -2.0462   0.6193   3.6417  11.2090

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.792e+01  2.456e+01   0.730 0.469402
`Median SAT`            7.201e-02  1.798e-02   4.004 0.000236 ***
`Acceptance Rate`      -2.486e+01  8.315e+00  -2.990 0.004560 **
`Expenditures/Student` -1.356e-04  6.593e-05  -2.057 0.045600 *
`Top 10% HS`           -1.628e-01  7.934e-02  -2.051 0.046214 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.308 on 44 degrees of freedom
Multiple R-squared:  0.5344,    Adjusted R-squared:  0.4921
F-statistic: 12.63 on 4 and 44 DF,  p-value: 6.332e-07
```




Graduation %
60
65
70
75
80
85
90
1100
1150
1200
1250
1300
1350
1400
Median SAT


Graduation %
60
65
70
75
80
85
90
2e+04
4e+04
6e+04
8e+04
1e+05
Expenditures/Student



2.1.3. PYTHON LANGUAGE

```
results_mul=sm.OLS(y,x1).fit()
results_mul.summary()
```

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.534
Model:	OLS	**Adj. R-squared:**	0.492
Method:	Least Squares	**F-statistic:**	12.63
Date:	Sun, 07 Apr 2024	**Prob (F-statistic):**	6.33e-07
Time:	11:10:33	**Log-Likelihood:**	-148.69
No. Observations:	49	**AIC:**	307.4
Df Residuals:	44	**BIC:**	316.8
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	17.9210	24.557	0.730	0.469	-31.571	67.413
x1	0.0720	0.018	4.004	0.000	0.036	0.108
x2	-0.2486	0.083	-2.990	0.005	-0.416	-0.081
x3	-0.0001	6.59e-05	-2.057	0.046	-0.000	-2.77e-06
x4	-0.1628	0.079	-2.051	0.046	-0.323	-0.003

Omnibus:	1.954	**Durbin-Watson:**	2.010
Prob(Omnibus):	0.376	**Jarque-Bera (JB):**	1.833
Skew:	-0.450	**Prob(JB):**	0.400
Kurtosis:	2.706	**Cond. No.**	1.09e+06

2.1.4. CONCLUSION

If Pr(>|t|) of expend > 0.05, it means there is no statistical evidence to conclude that the expend variable has a significant effect on the dependent variable. In other words, the expend variable may not be necessary for the model.

In this case, we already have a suitable model.

%graduation = 17.921 + 0.072*median_sat – 24.8592*accept – 0.0001* expend – 0.1628*top

- **F test:**

Hypothetical:

H0 : b1 = b2 = 0

Ha : There is at least 1 non-zero bi parameter

Rejection rule:

Reject H0 if p-value < .05

Conclusion: p-value < .05, so H0 can be rejected

⇨ Concluding that at least one of the independent variables has a significant significance effect on the dependent variable in the regression model.

- **t Test:**

Hypothetical:

H0 : bi = 0

Ha : bi # 0

Conclusion:

p-value < 0.05 => Reject both H0: b1 = 0 and H0: b2 = 0

Both independent variables (Median SAT, Acceptance rate, Expenditures/student, Top 10% of HS class) are significant

2.2. Multivariable Nonlinear Regression

The dataset: Vietnamese car price

Description: This is data about buying and selling used cars in the Vietnamese market at the beginning of 2023. The data includes information about the car such as price, number of kilometers traveled, and information about the seller such as phone number… , name, sales website (if exists)

Dataset overview:

ad_id	origin	condition	car_model	mileage	exterior_cc	interior_co	num_of_dc	seating_ca	engine	fuel_syster	transmissi	drive_type	fuel_consu	describe	brand	grade	year_of_ma	car_name	price. price	url
17042	Domestic a	New car	Truck	0 Km	White	gray	2-door	2-seat	Petrol1.0 L		Manual	RFD - Rear	L/100Km	Super Carr	Suzuki	Super Carr	2022	Suzuki Sup	249 Million	https://bonba
53794	Imported	New car	SUV	0 Km	Black	Black	5-door	7-seat	Petrol3.4 L		Automatic	AWD - 4-w	10L/100Kn	New Toyota	Toyota	Land Cruis	2022	Toyota Lan	4 Billion 28	https://bonba
73954	Domestic a	New car	Crossover	0 Km	Silver	Brown	5-door	8-seat	Petrol2.0 L		Automatic	RFD - Rear	L/100Km	**Registra	Toyota	Innova	2023	Toyota Innc	885 Million	https://bonba
74150	Imported	New car	SUV	0 Km	White	Black	5-door	5-seat	Petrol1.8 L		Automatic	FWD - Fror	L/100Km	2 interior c	Toyota	Corolla Crc	2023	Toyota Cor	754 Million	https://bonba
87573	Domestic a	New car	Crossover	0 Km	Silver	gray	5-door	8-seat	Petrol2.0 L		Automatic	RFD - Rear	L/100Km	Toyota Innc	Toyota	Innova	2022	Toyota Innc	850 Million	https://bonba
97011	Domestic a	New car	Van/Miniva	0 Km	White	gray	5-door	2-seat	Petrol1.0 L		Manual	RFD - Rear	7L/100Km	Suzuki Blin	Suzuki	Super Carr	2023	Suzuki Sup	299 Million	https://bonba
101726	Domestic a	New car	SUV	0 Km	White	Black	5-door	7-seat	Petrol1.5 L		Automatic	FWD - Fror	L/100Km	Honda CR-	Honda	CRV	2023	Honda CR\	984 Million	https://bonba
135739	Imported	New car	SUV	0 Km	Copper	Black	5-door	7-seat	Petrol2.7 L		Automatic	RFD - Rear	L/100Km	Toyota For	Toyota	Fortuner	2023	Toyota For	1 Billion 22	https://bonba
142495	Domestic a	New car	BaÌn taÌ‰i	0 Km	Grey	Black	4-door	5-seat	Diesel2.2 L	Single turb	Automatic	RFD - Rear	7L/100Km	FORD RAN	Ford	Ranger	2023	Ford Range	688 Million	https://bonba
143308	Domestic a	New car	BaÌn taÌ‰i	0 Km	Black	Black	4-door	5-seat	Diesel2.2 L	Single turb	Automatic	4WD - Fou	7L/100Km	FORD RAN	Ford	Ranger	2023	Ford Range	830 Million	https://bonba
174951	Domestic a	New car	Van/Miniva	0 Km	White	Cream	4-door	7-seat	Diesel2.2 L	TDCi	Manual	RFD - Rear	9L/100Km	FORD TRAI	Ford	Transit	2023	Ford Trans	940 Million	https://bonba
182003	Domestic a	New car	Crossover	0 Km	Silver	gray	5-door	8-seat	Petrol2.0 L		Manual	RFD - Rear	L/100Km	Toyota Innc	Toyota	Innova	2022	Toyota Innc	750 Million	https://bonba
182298	Imported	New car	SUV	0 Km	White	Cream	5-door	5-seat	Petrol2.7 L	VVTi	Automatic	4WD - Fou	10L/100Kn	Land Cruis	Toyota	Prado	2022	Toyota Pra	2 Billion 58	https://bonba
183963	Domestic a	New car	Sedan	0 Km	White	Yellow	4-door	5-seat	Petrol1.5 L	Multi-point	Automatic	FWD - Fror	6L/100Km	Toyota Vios	Toyota	Vios	2022	Toyota Vios	542 Million	https://bonba
203337	Imported	New car	SUV	0 Km	White	Black	5-door	7-seat	Diesel2.0 L	Single turb	Automatic	RFD - Rear	7L/100Km	FORD EVEI	Ford	Everest	2023	Ford Evere	1 Billion 99	https://bonba
211394	Imported	New car	Hatchback	0 Km	Red	Cream	5-door	5-seat	Petrol1.5 L		Automatic	FWD - Fror	L/100Km	Toyota Yari	Toyota	Yaris	2022	Toyota Yari	650 Million	https://bonba
304604	Domestic a	New car	Sedan	0 Km	Black	Cream	4-door	5-seat	Petrol1.5 L	VVTI	Automatic	FWD - Fror	6L/100Km	Toyota Vios	Toyota	Vios	2022	Toyota Vios	592 Million	https://bonba
449343	Domestic a	New car	SUV	0 Km	White	Black	5-door	7-seat	Petrol2.0 L	Multi-point	Automatic	FWD - Fror	8L/100Km	Free prem	Mitsubishi	Outlander	2022	Mitsubishi	825 Million	https://bonba

The price of the car depends on the number of seats and the year of manufacture

- ⇨ **Independent variables:** seating_car and year_of_manufacture
- ⇨ **Dependent variable:** price.price_car

Sample regression model:

Price_car = β0 + β1 ∗ log10(seating_car) + β2 ∗ log10(year_of_manufacture)

2.2.1. MS EXCEL

2.2.1.1. Using tool:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.14198641
R Square	0.02016014
Adjusted R Square	0.01976797
Standard Error	2682.41726
Observations	5000

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	2	7.4E+08	3.7E+08	51.40647	0
Residual	4997	3.6E+10	7195362		
Total	4999	3.67E+10			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 95.0%*
Intercept	**-768239.07**	125309.3	-6.13074	9.42E-10	-1013900	-522578	-1013900	-522578
log(seating)	**-3078.0393**	367.6619	-8.37193	7.28E-17	-3798.82	-2357.26	-3798.82	-2357.26
log(year)	**233608.261**	37916.79	6.161078	7.79E-10	159274.7	307941.8	159274.7	307941.8

2.2.1.2. Step by step:

XT	1	1	1	1	1	1
	0.301029996	0.845098	0.90309	0.69897	0.90309	0.30103
	3.305781151	3.305781	3.305996	3.305996	3.305781	3.305996
XT*X	5000	3661.935	16526.12			
	3661.935089	2735.337	12103.54			
	16526.11825	12103.54	54622.52			
(XT*X)-1	2182.299258	0.329554	-660.331			
	0.329554301	0.018786	-0.10387			
	-660.3306314	-0.10387	199.807			
(XT*X)-1 * XT	-0.51009083	-0.33079	-0.45347	-0.52074	-0.31168	-0.65188
	-0.008161574	0.00206	0.003127	-0.00071	0.003149	-0.00818
	0.156197678	0.099685	0.136567	0.157769	0.093662	0.199103
(XT*X)-1 * XT*Y	**-768239.5156**					
	-3078.039406					
	233608.3947					

⇨ Comparing with the Data Analysis results calculation tool, we see that the value systems in the two tables are the same

2.2.2. R LANGUAGE

```
> attach(df)
> reg <- lm(df$price..Price ~ df$log.seating. + df$log.year.)
> summary(reg)

Call:
lm(formula = df$price..Price ~ df$log.seating. + df$log.year.)

Residuals:
   Min     1Q Median     3Q    Max 
 -2866  -1158   -784    -17  37935 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -768239.1   125309.3  -6.131 9.42e-10 ***
df$log.seating.   -3078.0      367.7  -8.372  < 2e-16 ***
df$log.year.     233608.3    37916.8   6.161 7.79e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2682 on 4997 degrees of freedom
Multiple R-squared:  0.02016,	Adjusted R-squared:  0.01977 
F-statistic: 51.41 on 2 and 4997 DF,  p-value: < 2.2e-16
```

2.2.3. PYTHON LANGUAGE

```
y = data['price. Price']

# For independent variables (x), select the desired columns
x = data[['log(seating)', 'log(year)']]

# Add a constant term to the independent variables
x = sm.add_constant(x)

reg = sm.OLS(y,x).fit()
reg.summary()
```

OLS Regression Results

Dep. Variable:	price. Price	**R-squared:**	0.020
Model:	OLS	**Adj. R-squared:**	0.020
Method:	Least Squares	**F-statistic:**	51.41
Date:	Sun, 07 Apr 2024	**Prob (F-statistic):**	7.96e-23
Time:	09:38:12	**Log-Likelihood:**	-46566.
No. Observations:	5000	**AIC:**	9.314e+04
Df Residuals:	4997	**BIC:**	9.316e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-7.682e+05	1.25e+05	-6.131	0.000	-1.01e+06	-5.23e+05
log(seating)	-3078.0393	367.662	-8.372	0.000	-3798.818	-2357.261
log(year)	2.336e+05	3.79e+04	6.161	0.000	1.59e+05	3.08e+05

Omnibus:	5614.000	**Durbin-Watson:**	1.660
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	605360.083
Skew:	5.729	**Prob(JB):**	0.00
Kurtosis:	55.673	**Cond. No.**	1.22e+04

2.2.4. CONCLUSION

From the above results we can draw the following conclusions:

Regression equation:

Price_car = -768239 - 3078*log10(seating_car) + 233608 *log10(year_of_manufacture)

The coefficients in the equation represent the contribution of each independent variable to the value of Price_car:

The coefficient -3078*log10(seating_car) represents the relationship between the number of car seats and the Price_car. The negative coefficient (-3078) indicates that as the number of seats increases, the Price_car tends to decrease.

The coefficient 233608*log10(year_of_manufacture) represents the relationship between the year of car manufacture and the Price_car. The positive coefficient (233608) indicates that as the year of manufacture increases, the Price_car tends to increase

2.3. Logistic Regression

The dataset: [Vietnamese car price]

Description: This is data about buying and selling used cars in the Vietnamese market at the beginning of 2023. The data includes information about the car such as price, number of kilometers traveled, and information about the seller such as phone number… , name, sales website (if exists)

Dataset overview:

ad_id	origin	condition	car_model	mileage	exterior_cc	interior_co	num_of_dc	seating_ca	engine	fuel_syster	transmissi	drive_type	fuel_consu	describe	brand	grade	year_of_ma	car_name	price. pric	url
17042	Domestic a	New car	Truck	0 Km	White	gray	2-door	2-seat	Petrol1.0 L		Manual	RFD - Rear	L/100Km	Super Carr	Suzuki	Super Carr	2022	Suzuki Sup	249 Million	https://bonba
53794	Imported	New car	SUV	0 Km	Black	Black	5-door	7-seat	Petrol3.4 L		Automatic	AWD - 4-w	10L/100Kn	New Toyota	Toyota	Land Cruis	2022	Toyota Lan	4 Billion 28	https://bonba
73954	Domestic a	New car	Crossover	0 Km	Silver	Brown	5-door	8-seat	Petrol2.0 L		Automatic	RFD - Rear	L/100Km	**Registra	Toyota	Innova	2023	Toyota Inn	885 Million	https://bonba
74150	Imported	New car	SUV	0 Km	White	Black	5-door	5-seat	Petrol1.8 L		Automatic	FWD - Fror	L/100Km	2 interior c	Toyota	Corolla Cr	2023	Toyota Cor	754 Million	https://bonba
87573	Domestic a	New car	Crossover	0 Km	Silver	gray	5-door	8-seat	Petrol2.0 L		Automatic	RFD - Rear	L/100Km	Toyota Inn	Toyota	Innova	2022	Toyota Inn	850 Million	https://bonba
97011	Domestic a	New car	Van/Miniva	0 Km	White	gray	5-door	2-seat	Petrol1.0 L		Manual	RFD - Rear	7L/100Km	Suzuki Blin	Suzuki	Super Carr	2023	Suzuki Sup	299 Million	https://bonba
101726	Domestic a	New car	SUV	0 Km	White	Black	5-door	7-seat	Petrol1.5 L		Automatic	FWD - Fror	L/100Km	Honda CR-	Honda	CRV	2023	Honda CR\	984 Million	https://bonba
135739	Imported	New car	SUV	0 Km	Copper	Black	5-door	7-seat	Petrol2.7 L		Automatic	RFD - Rear	L/100Km	Toyota For	Toyota	Fortuner	2023	Toyota For	1 Billion 22	https://bonba
142495	Domestic a	New car	BaÌn taÌ‰i	0 Km	Grey	Black	4-door	5-seat	Diesel2.2 L	Single turb	Automatic	RFD - Rear	7L/100Km	FORD RAN	Ford	Ranger	2023	Ford Range	688 Million	https://bonba
143308	Domestic a	New car	BaÌn taÌ‰i	0 Km	Black	Black	4-door	5-seat	Diesel2.2 L	Single turb	Automatic	4WD - Fou	7L/100Km	FORD RAN	Ford	Ranger	2023	Ford Range	830 Million	https://bonba
174951	Domestic a	New car	Van/Miniva	0 Km	White	Cream	4-door	7-seat	Diesel2.2 L	TDCi	Manual	RFD - Rear	9L/100Km	FORD TRAI	Ford	Transit	2023	Ford Trans	940 Million	https://bonba
182003	Domestic a	New car	Crossover	0 Km	Silver	gray	5-door	8-seat	Petrol2.0 L		Manual	RFD - Rear	L/100Km	Toyota Inn	Toyota	Innova	2022	Toyota Inn	750 Million	https://bonba
182298	Imported	New car	SUV	0 Km	White	Cream	5-door	5-seat	Petrol2.7 L	VVTi	Automatic	4WD - Fou	10L/100Kn	Land Cruis	Toyota	Prado	2022	Toyota Pra	2 Billion 58	https://bonba
183963	Domestic a	New car	Sedan	0 Km	White	Yellow	4-door	5-seat	Petrol1.5 L	Multi-point	Automatic	FWD - Fror	6L/100Km	Toyota Vios	Toyota	Vios	2022	Toyota Vios	542 Million	https://bonba
203337	Imported	New car	SUV	0 Km	White	Black	5-door	7-seat	Diesel2.0 L	Single turb	Automatic	RFD - Rear	7L/100Km	FORD EVEI	Ford	Everest	2023	Ford Evere	1 Billion 99	https://bonba
211394	Imported	New car	Hatchback	0 Km	Red	Cream	5-door	5-seat	Petrol1.5 L		Automatic	FWD - Fror	L/100Km	Toyota Yari	Toyota	Yaris	2022	Toyota Yari	650 Million	https://bonba
304604	Domestic a	New car	Sedan	0 Km	Black	Cream	4-door	5-seat	Petrol1.5 L	VVTI	Automatic	FWD - Fror	6L/100Km	Toyota Vios	Toyota	Vios	2022	Toyota Vios	592 Million	https://bonba
449343	Domestic a	New car	SUV	0 Km	White	Black	5-door	7-seat	Petrol2.0 L	Multi-point	Automatic	FWD - Fror	8L/100Km	Free prem	Mitsubishi	Outlander	2022	Mitsubishi	825 Million	https://bonba

⇨ **Independent variables:** mileage

⇨ **Dependent variable:** condition

Sample regression model:

$$Logit = X = b0 + b1 * (\text{mileage})$$

2.3.1. MS EXCEL

SUMMARY OUTPUT								
Regression Statistics								
Chi Square	4471.115515							
Residual Dev.	2456.593132							
# of iterations	18							
Observations	4999							
	Coefficients	Standard Error	P-value	Odd Ratio	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Intercept	1.804696891	0.052671703	2.8E-257	6.078129	5.481958	6.739135	5.4819575	6.739135

SUMMARY OUTPUT								
Regression Statistics								
Chi Square	4472.153908							
Residual Dev.	2456.897702							
# of iterations	18							
Observations	5000							
	Coefficients	Standard Error	P-value	Odd Ratio	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Intercept	**1.805088769**	0.052670219	2.1E-257	6.080511	5.484122	6.741757	5.4841222	6.741757
mileage	**-0.004142683**	0.000628764	4.44E-11	0.995866	0.994639	0.997094	0.9946394	0.997094

2.3.2. R LANGUAGE

```
> logistic <- glm(df$condition ~ df$mileage, data = df, family = binomial())
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(logistic)

Call:
glm(formula = df$condition ~ df$mileage, family = binomial(),
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9786   0.0000   0.5518   0.5518   6.1495

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.8050888  0.0526702  34.272  < 2e-16 ***
df$mileage  -0.0041427  0.0006287  -6.589 4.42e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6929.1  on 4999  degrees of freedom
Residual deviance: 2456.9  on 4998  degrees of freedom
AIC: 2460.9

Number of Fisher Scoring iterations: 16
```

2.3.3. PYTHON LANGUAGE

```
logistic = sm.formula.glm('condition ~ mileage', data=data, family=sm.families.Binomial()).fit()
print(logistic.summary())


# x = np.array(data["duration"]).reshape((-1, 1))
# y = np.array(data["term_deposit"])
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              condition   No. Observations:                 5000
Model:                            GLM   Df Residuals:                     4998
Model Family:                Binomial   Df Model:                            1
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                    nan
Date:                Sun, 07 Apr 2024   Deviance:                       2456.9
Time:                        14:49:36   Pearson chi2:                 1.63e+08
No. Iterations:                    17   Pseudo R-squ. (CS):                nan
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.8051      0.053     34.272      0.000       1.702       1.908
mileage       -0.0041      0.001     -6.589      0.000      -0.005      -0.003
==============================================================================
/usr/local/lib/python3.10/dist-packages/statsmodels/genmod/families/links.py:198: RuntimeWarning: overflow encountered in exp
  t = np.exp(-z)
/usr/local/lib/python3.10/dist-packages/statsmodels/genmod/families/family.py:1056: RuntimeWarning: divide by zero encountered in log
  special.gammaln(n - y + 1) + y * np.log(mu / (1 - mu + 1e-20)) +
/usr/local/lib/python3.10/dist-packages/statsmodels/genmod/families/family.py:1056: RuntimeWarning: invalid value encountered in multiply
  special.gammaln(n - y + 1) + y * np.log(mu / (1 - mu + 1e-20)) +
```

2.3.4. CONCLUSION

From the results we obtain the Logistic Regression equation as follows:

$$\text{Log}\left(\frac{p}{1-p}\right) = 1.805 - 0.004 * \text{mileage}$$

⇨ $\frac{p}{1-p} = e^{1.805 - 0.004 * mileage}$

Set odd = $\frac{p}{1-p}$

Odd(0) = $\left(\frac{\boldsymbol{p}}{\boldsymbol{1-p}}\right) = e^{1.805}$ =6.08 when mileage = 0

Odd(1) =$\left(\frac{\boldsymbol{p}}{\boldsymbol{1-p}}\right) = e^{1{,}805 - 0.004}$ = 6.06 when mileage = 1

Then we have the ratio between Odd(1)/ Odd(0) = 0.997

REFERENCES

[1] https://fsppm.fulbright.edu.vn/cache/FSLM-10-MultipleRegressionV-2021-03-01-16463454.pdf

[2] https://www.youtube.com/watch?v=AP_K7SaKkIE

[3] https://www.kaggle.com/datasets/nguynthanhlun/vietnamese-car-price

[4] https://xdulieu.com/da-bien/db1-tuong-quan-hoi-quy/ht7-hoi-quy-logistic.html

[5] J.Donohue. (2013). Ch9-Regression Analysis-T [PowerPoint Slides]

[6] N.M.Nhut - "Lecture-6-Thống-kê-hồi-quy-ALL"

LINK CANVA