Regression theory in linear and logistic model
With an illustrated example in programming language R

MEDINI Atmane [1*], AOUADI Mostafa [2],

[1] Département de sciences de gestion Faculty des SECG/Jijel (Algérie)
Financial, accounting, collection and insurance, University of Souk Ahras.
(a.medini@univ-jijel.dz)
[2] Département de sciences de gestion El oued (Algérie) ,Labortory of economic
Business and management, (pr.aouadi@gmail.com)

Received: 29/04/2021; **Revised:** 17/05/2021; **Accepted:** 10/07/2021

Summary: The validity of the model in linear theory regression, was determined through the optimal estimation of the line variation through the analysis of sediments on OLS, but if study was concerned on logistic model, a number of different statistics will have to be resorted to, such as Homs -limcho 'statistics to examining the quality of conformity ,such classification tables where there is on other statistics that's give a partial indication. and significance of the variables Interpretation. Where the study aimed to meet the regression theory between the two models while supporting the idea with an applied example in the programming language R.

Keywords: Least Squares ; Linear Regression ; Logistic Model ; Homes statistics; Coefficient.
Jel Classification Codes : C18 ; C99; C16 ; C19 ; C46.

[*] Corresponding author.

I- Introduction :

The validity of the statistical model for estimation requires knowledge of the extent to which it provides the appropriate conditions for the appropriate tools to be adopted, as residual analysis and differences between the original model and the estimated model are among the most important parts in determining the path of analysis and the accuracy of the results in the regression theory. The process of analysis is summarized in comparing the observed data of the original model with the expected data, that is, before and after the introduction of the interpreted variables, so that the researcher can acknowledg+e the ability of the estimated model to explain the relationship in the long term. However, this is linked to overlapping criteria related to the nature of data distribution and type. This will entail identifying different paths in the analysis according to their appropriate tests.

I.1. Problématique of study:

Determining the statistical method supports the theory of measurement, as well as avoids falling into the error of identification, so the question lies in identifying the difference between the linear model and the logistic model in analyzing sediments and differences?

I.2. hypotheses:

h0: There is no fundamental difference between the linear model and the logistic model in the residual analysis method and the differences in the regression theory.
h1: There is a fundamental difference between the linear model and the logistic model in the method of sediment and variance analysis to identify the suitability of the statistical model.

1.3. Objectives of the study:

• Identify the mathematical foundations in the theory of sediment analysis and the differences in regression theory .
• Identify the more significant differences in sediment analysis and the differences between the linear model and the logistic model .
•Drop those differences on the example attached to the study.

1.4.. Previous studies:

 study of Muhammad Amin Ayyash,)The Logistic Regression Model, Its Concept, Characteristics and Applications), 2017, Jalat Al-Sarraj in Education and Community Issues, Issue 1, where the study aimed not to highlight the characteristics of logistic regression and its importance in predicting nominal and categorical variables by highlighting in detail the tests and statistics related to the model, which It differs from the method for estimating suicide streak variance in linear models.
 study by Adel bin Ahmed bin Hussein Babtain, (Logistic suicide and how to use it in building prediction models for data), 2017, PhD thesis, Umm Al-Qura University, the aim of the study was the theoretical rooting of the models of the binary logistic regression model, and it reached a difference in the various statistics related to the logistic model .

I.1. Theoretical rooting of regression models

I.1.1.Analysis of residual and variances in linear model:

The purpose of the residual and differences analysis process is to reach the possible maximum for the proposed model)the quality of model matching), in order to adopt the interpreted variables to answer the problems of the study. However, these tests are supposed to study the model on its total level first before exposure to its detailed parameters.

I.1.2. Theoretical rooting of linear model

The model described as linear when the data explaining the phenomenon to it with the predictor variable takes a linear form, where the researcher can observe the distribution of the data in the spread plate within the graph in harmony that prompts the belief that there is a linear point continuum between the variables. On the mathematical level, it is expressed in the following formulas

$$y = \sum_{i=1}^{n} \beta x_i = \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n$$

As it will have define the model parameters according to an aggregate method, so that the parameter β indicates the change in the value of the dependent variable with the change of one unit of the interpreted variable. The significance of the model will be related to the partial significance of its parameters by the same grouping.

I.1.3. Matching quality within the linear model and its characteristics:

The quality of matching the statistical model within the linear model requires an analysis of the residual and the differences within the characteristics of their probability distributions, and the residual mean the analysis of the gap between the observed values and the estimated values or the reconciliation. between and yˆ, for the purpose of estimation and prediction, if we. know the value of the subject variable if the value of another variable is known (Abu Saleh and Awad, 1983, p .: 200). While the main of differences are intended to analyze the gap between the observed values and their arithmetic mean and, or between the observed values and the arithmetic mean of the estimated values yˆ and ,. In order for its results to be adopted within a comprehensive comparative structure through which the validity of the model is determined. The result of the total gap comparison is adopted based on the value of f computed with the tabular value of the Fisher distribution through a fraction that denotes the total cause of the gaps due to the regression to the gaps due to errors.

I.1.4. **Steps for analyzing residual and differences in the linear model** (Muhammad Sobhi Adnan Awad, 1983, p .: 204).

The residual and differences in the linear regression model (simple or multiple) are classified according to their source within the theory of estimating the variance of the regression line to three, where the goal is to reach estimated values that make the sum of the squares of the deviations as small as possible, that is, that makes the quantity .

$$e = Q\,(a, b\,) = \varepsilon$$

The sum of the squares deviations, denoted by SST, and expresses the sum of the squares of the deviations of the actual values (observed from their arithmetic mean)..

$$SST = \sum (y_i - \bar{y}\,)^2 \;=\; \sum y_i^2 \;\; n\bar{y}^2$$

The sum of squares of the regression expresses the sum of the squares of the deviations of the estimated values y from the arithmetic mean of the actual values y and gives an estimate of the variance between the values estimated in the model and the arithmetic mean of the actual values.

$$SSR = \Sigma(\hat{y} - \bar{y}\,).^2 \;\; = b^2[x_i^2 \text{ -n}\bar{x}^2\text{]}$$

Where the sum of squares of error, denoted as: $SSE = \Sigma(y_i - \hat{y}\ )2$, and expresses the sum of the squares of the deviations between the actual values from the associated estimated values. Where it will be based on the amount of those differences at specific degrees of freedom to arrive at the calculated value of f and then compare it with the tabular value to determine the validity of the model or not. The most appropriate and appropriate model represents that which maximizes the congruence between the original model and the model containing the variables explained by the lowest possible value of the gap.

I.1.5. Partial Parameters in a Linear Regression Model (Babtain, 2017**)**

The determination coefficient is used to interpret the significance of a whole in the linear model, expressed in the form of a mathematical fraction and indicating the relative value between the sum of the squares of the deviation due to the regression line (SSR) to the sum of the squares of the total deviation (SST). R2 = SSR / SST. To indicate the explanatory power of the independent variables in the model, and to express the partial significance of the variables to the significance of the B parameters within a direct reading of them.

I.I.I. . Analyzing sediments and differences in the logistic model

I.I.I..1 Concept of logistic model

Logistic model is used when studying the relationship of an interpreted variable to another dependent within two limits, it was a study of the effect of a particular drug on the infection of a disease, where the dependent variable will be distributed on its effect according to two values (1 or 0), to indicate the positive value of the disease, while the value zero denies the incidence of it. . This will make the distribution of the values of the dependent change to be distributed in an extreme manner around Diagonal, which requires the necessary mathematical transformations in the structure of the estimation equation, which will allow dealing with it later according to a special method. The significance of the dependent variable Y in the linear model is replaced by the Odds ratio. The parameters will be read differently from the linear model.

I.I.I..2. Probability and the coefficient weight (Ozbon, 2018, pp: 30–34)

We defintheOdde ratio as, (odds = (y) / (1-y)) which in its quantity corresponds to ((p) / (1-p)) and denotes the ratio of the event to the complementary event. It differs from the concept of (P) in that the values of the weight coefficient may take infinitely small or large values in contrast to the probability, as a value is confined to the domain (0,1), so the zero (0)is the probability of the occurrence of the event necessarily gives a zero value also for the weight coefficient, while whenever An increase in the probability value will lead to an increase in the weighting coefficient by values that exceed the correct one (see Table 01). As the formula for weighting coefficient overcomes the problem of the upper limits of probability, while the problem of the lower limits is still present, which requires the introduction of the Ln function to move to the talk about the logarithm of the weighting function Log Likelihood Function - -2LL or LRX2

LRX2 = -2 (log likelihood without variables - log likelihood with variables).

Table (01): shows the difference between the probability and the weight coefficient

p_i	0.99999	0.9999	0.999	0.99	0
O_i	99999	9999	999	99	0

Source: Adel Bin Ahmed Bin Hassan Babtain, 2017, Logistic Regression and How to Use It in Building Prediction Models for Data with Two-Value Dependent Variables P. 47.

I.I.I..3. Logarithm of the weighting function and mathematical transformations (Babtain, p .: 77)

The introduction of the logarithm function to the explained quantity requires that it be followed by some mathematical transformations that allow a different interpretation of the significance of the coefficients. If the formula for the logistic regression equation is written as follows:

$$y = p = (e^{a+bx} /_{1-e^{a+bx}})$$

If the odds ratio expresses the value of the probability to the complementary probability, then we will obtain the following relationship:

$$O = \frac{(e^{ax+bixi} / (1+e^{ax+bixi})}{\frac{1}{e^{ax+bixi}}} = (e^{a+bixi})$$

It is the final form of the weight factor before entering the function ln. Then the parameters will be read in a linear summative manner, taking into account the relation of the parameters to the value exp.

I.I.I. 4. Measuring the validity of the logistic model:

Judging the validity of the logistic model is related to two levels, the suitability of the overall model, then the partial validity related to its interpreted parameters, where the judgment on the validity of the logistic model is based on the results of several important tests summarized in, sediment and difference analysis, and the analysis of the quality of the model's matching of its data according to the Hosmer Limsho test, Classification table, m partial estimate of the parameters (wald statistic).

❖ **Analysis of residual and differences in the logistic model** (Babtain, p. 95)

The theoretical basis of residual and variance analysis in the logistic model does not differ from that in the linear model, except as required by the consideration of the binary distribution nature of the model. It is an example of a comparative analysis between the observed values and the expected values within a different relationship, and then judged using its statistic x^2 ,Where the goal is to find the maximum reduction of the difference between the values of the variance matrix or the correlations in the data .

❖ Measuring the quality of data matching

The confidence model test is among the important tests that indicate the validity of the model, which the researcher should stand on. Where a model that is of good conformity to the estimated regression line is the one that guarantees a strong relationship, clear parameters, and high predictability (Abu Shukan and Ali, 2014, p. 04). However, in many statistical literature the concepts associated with this type of test are confused. Therefore, we shall endeavor to clarify it in proportion to the nature of the probability variations.

□ Measure the quality of confidently in the linear model

Knowing the confidently in the linear model, is not related to one hypothesis, but rather to several hypotheses, the goal was to find out the extent to which the current distribution of the data corresponds to a specific reference distribution, and it can also give an indication about the extent to which the estimated values match the actual values. It is based in the linear model on the sum of squares of deviations The estimated values refer to the sum of the squares of the actual values, and the statistical decision is based on the comparative value, to reject or accept the statistical hypothesis.

❖ Classification of the conformance quality test according to its probability distributions

If it is related to linear regression (simple or multiple), the researcher suffices to read the results of the table for estimating the variance of the regression line, while other determinants will have to be resorted to if it comes to studying logistic regression models, such as tables of classification Tables, and a similar coefficient of determination; Pseudo R2 and the partial significance of statistic variables Wald and ROC curve analysis..

❖ Test the quality of data in the logistical model (Babtain, p .: 104)

The quality of confidently in the logistic model knows as Hosmer-Lemshow Goodness-Of-Fit Test where it is based on the idea of finding a measure that expresses the extent of the difference between the numbers of the observed values and the numbers of the expected values, where the sample cases are grouped based on the expected cases according to the two values (Y = 0 and y = 1) and then classifying the cases that have a probability less than (0.1) in the first group, while the cases that have a probability greater than (0.9) are placed in the tenth group, and so on for the rest of the groups. Also, the difference if it comes to the logistic regression will include the statistic The approved Hosmer-Lemshow statistic is used, which takes the following formula:

$$\hat{C} = \sum_{K=1}^{n} (O_k - n'_k \overline{P}_k))^2 \;/\; n'_k \overline{P}_k \;(1 - \overline{P}_k)$$

n'_k: represents the total number of cases in Cluster K.

$O_k = \sum_{i=1}^{n'_k} y_i$: Represents the number of responses y = 1

$\overline{P}_k = \sum_{i=1}^{n'_k} p_{i\,/n'_k}$: represents the average expected probability of group K

Where the statistic $\hat{C}$ follows the distribution $x^2$with degrees of freedom (n-2)

To billed the hypotheses as follows:

h0: There is a match for the data if it is:

$$\hat{C} = \sum_{K=1}^{n}(O_k - n'_k \overline{P}_k))^2 / n'_k \overline{P}_k(1 - \overline{P}_k) > \; x^2\,(n-2\,)$$

h1: The data does not match if it:

$$\hat{C} = \sum_{K=1}^{n}(O_k - n'_k \overline{P}_k))^2 / n'_k \overline{P}_k(1 - \overline{P}_k) \; < \; x^2\,(n-2)$$

❖ Classification Tables (Tom F, p: 862, 2005)

The method of analysis according to classification tables based on classifying the data according to the probability of their occurrence, through a comparison between sensitivity, which expresses the probability that the classification will be positive when it is expected to be a positive act, and the accuracy that expresses the probability that the expected classification will be negative when it is actually negative, and whenever a percentage Correct classification indicates that the model has good predictability.

❖ **Partial estimation of the logistic regression parameters** (Babtain, pg: 99)

If his statistics are unimportant and the classification tables are concerned with the overall significance of the validity of the logistic model, then there are detailed tests whose mission is to give the researcher a partial indication of the importance of the variables included in his statistical model .

❖ **wald statistic for partial significance:**

His statistic denotes the partial significance of the variables, each separately in their ability to interpret the dependent variable in isolation from other variables according to the following hypothesis structure:

$$(h_0: b = 0 \ / \ h1: b \neq 0)$$

It expresses the value of the standard error of the logistic regression coefficient of the listed variable and traces the distribution 〖ch〗 ^2,

If the value of the statistic indicates its importance, then this means rejecting the null hypothesis that the value of the regression coefficient is equal to zero and accepting the alternative hypothesis that there is an effect of the independent variable on the dependent variable.

❖ **pssodo R2:**

In logistic regression, the analogous coefficient of determination is used instead of the coefficient of determination, which approaches the fractional value of the comparison between the sum of the squares of the return deviation and the sum of the squares of the total deviation. It is estimated by the phrase :

$$R_L^2 = {}^{GM}/_{DO},$$

among the most important of its measures are Cox & Snell and Nagelkerke R. It should also be noted that there are other statistics that depend on its significance, among them, the statistic of the compatibility coefficient , $R_c^2 = \frac{G_M}{(G_M + N)}$

and wald statistic $R_w^2 = \frac{W}{(W+N)}$

N is the number of cases and W is the parent statistic of Wald Statistic.

❖ **Log coefficient β:**

The log coefficient is a measure that corresponds to the parameter B in a normal linear regression and indicates the partial effect of the independent variable. However, since the dependent variable is related to the variables interpreted in binary form, its interpretation will be by resorting to the function exp, where the value of the consequent effect of the explained variable will be expressed by the value of the log parameter raised to the function exp.

❖ **ROC curve** (586Tom F, p: 862 -, 2005)

The analysis according to the ROC curve is used for observation, regulation and taxonomic testing, which depends on the calculation of sensitivity and accuracy. This classification is drawn in a graph of probabilities, starting with coordinates (0 0,), which is the point that all states are considered negative, versus the point (11), which is considered positive among all cases, and the line connecting the two points is called the chance diagonal. Below the curve it ranges from zero to one. The analysis by ROC is a measure of the model's ability to distinguish between cases that have the feature under examination and those that are considered a measure of the model's ability to distinguish and classify between cases that have the feature under examination and those that do not.

- ROC = 0.5, so the model does not have the ability to discriminate that differs from chance.
- ROC≤ 0.8≥ 0.7, the model has an acceptable discriminant ability,
- ROC ≤ 0.9≥ 0.8, the model has excellent discriminant ability,
- ROC≥ 0.9, the model has a very excellent discriminant ability.

IV– practical of the concepts:

After discussing the conceptual frameworks for both theories, we will work to project them into two explanatory examples that present the different methods of analysis. Where the first example represents data processing within the theory of simple linear regression, while the second example represents data processing within the logistic regression approach, and we will also rely on the programming language R in both cases.

IV.1. . Regression Analysis in a Linear Model

The data for the first example constitute company sales attached to advertising costs, where the aim of the study is to identify the relationship of sales volume with spending on advertising.

Table (02): The sales volume to the company's advertising expenses

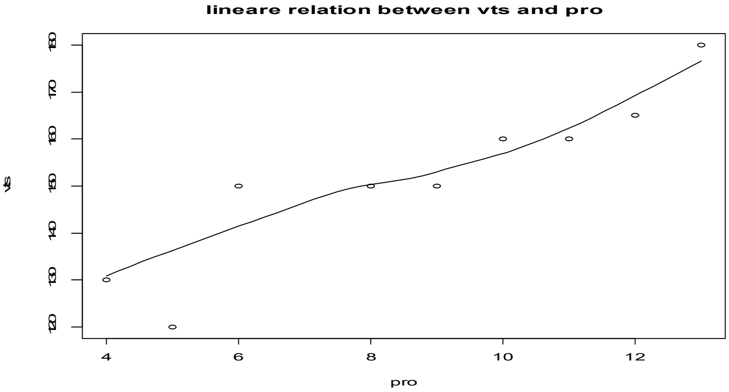
Advertising costs	8	10	6	12	13	5	11	9
Sales	150	160	130	165	180	120	160	150

Source: Muhammad Sobhi, Saleh and Adnan Muhammad Awad, Introduction to Statistics, John Wabel and Sons Publishing House, First Edition, p: 220

o **Spread plate data**

```
#curve plot of logestique regression between vts  and promotion
plot(vts~pro,xlab="pro",
           ylab="vts",
            main="relation vts and Promotion")
```

Figure (01) Distribution of publicity and advertising data to sales volume



Source: Prepared by the student based on the outputs of R

o Interpretation of the results of the linear model

Studying the correlation between the two variables to determine the form and direction of the relationship.

❖ **Multiple regression model between the depositors ’share of profits with the net result of equity**

In this part of the study, we will learn about the relationship of the depositors ’share of profits as a dependent variable to each of the net result and shareholder equity within a linear regression model, allowing the estimation of its parameters. But we will first have to ascertain the validity of its application. Where (Bin Hussein,2019,p:27-29) said that the Disrespect of these conditions would estimate false parameters of the model.

❖**Check the conditions for applying normal linear régression mode**l (Iade ،2021)

r(correlations)			Conditions for applying linear regression) Sel~Adve(
SEl		ADver	
	netr	Dac	*Absence of autocorrelation between independent variables*
netr	1	0.978**	
dac	0.978**	1	
Variance			*The values of the variable x are*

Sel		Adv	*variable and not fixed,*
370.892		7.929	
Mahal. Distance			*Normal distribution of residual residues and no outliers*
2.278			
Shapiro -teste			*Normal distribution of variables*
Sel			
0.719			
Collinearity Statistics			*multiple linear correlation*
tolerance		vif	
Adv	1.000	1.000	

Source: Prepared by the two researchers, based on R

□The condition of linearity of the relationship between the independent variables and the dependent variable: He revealed the existence of an average correlation between the equity of shareholders and the depositors ’share of profits, which explains the linear relationship with the dependent variable, but the correlation is characterized by weakness with the net result.

□ The condition of the absence of Auto-correlation between the independent variables: the correlation matrix between the independent variables revealed the existence of autonomy between the variables, that is, they are not related to each other by strong coefficients.

□ the condition that the values of the variable x are variable and not fixed: the matrix of covariance between the independent variables explains their non-fixed nature.

□ Condition of normal distribution of residual residues and absence of outliers: the Mahal value. Distance at a degree of freedom equal to the number of variables (2) equals 2.278, which is less than the tabular value.Therefore, we judge the moderation of the residual distribution and the absence of extreme values.

```
>cor.test(vts,pro)
Pearson's product-moment correlation
data:  vts and pro
t = 6.2511, df = 7, p-value = 0.0004237
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6611411 0.9835168
sample estimates:
cor
0.9209109
```

Source: Prepared by the researcher based on R

The correlation analysis gives a preliminary idea about the relationship of the explained variable to the dependent, as it is indicated in our example that advertising expenses are related within a strong positive relationship to sales.

□ **Validity of the written form**

The validity of the model is determined by the value of F-statistic: 39.08, then by its probability 0.00042 = p, which indicates the significance of the quantity. Comparison between the sum of the squares of the deviations due to the regression line and the sum of the squares of the total deviations. Where we can adopt the result to analyze the relationship between the two changes within a simple regression relationship with the R program as follows:

```
mydata<-read.table(file.choose(data),header = T,sep= "  "
Call:
lm(formula = vts ~ pro(
Residuals:
   Min      1Q  Median      3Q     Max
 12.333  2.833   1.333   3.917-  12.417-
Coefficients:
  Estimate Std. Error t value Pr(>|t    (|
)Intercept) 106.1667     7.6974   13.793 2.49e-06***
pro            5.2500       0.8399     6.251 0.000424***
                                                                                              ---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 7.512 on 7 degrees of freedom
                                                 Multiple R-squared:  0.8481, Adjusted R-squared:  0.8264
F-statistic: 39.08 on 1 and 7 DF,  p-value: 0.0004237
```

Source: Prepared by the researcher based on R

The determination factor interprets the total impact of advertising expenses on the volume of sales by 0.8264. Also, both the constant and the variable explained (publicity expense) are important in the model.

Table (03): Cancer incidence to smoking habit

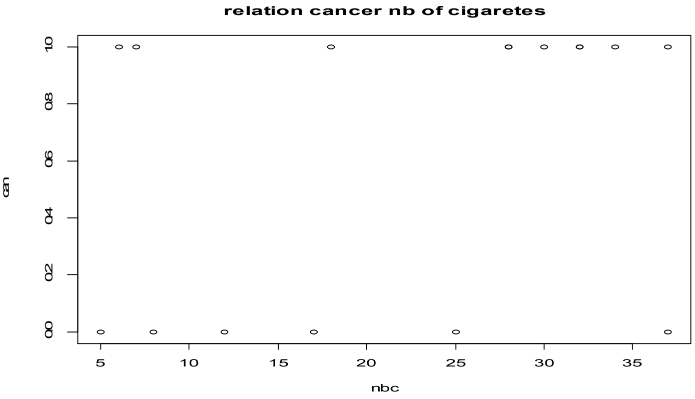
Cancer	0	0	1	1	1	0	0	0	1	1	1	1	1	0	1	1
Nb of cigarettes	5	25	6	18	7	8	12	17	32	37	28	34	32	37	30	28

Source: From the proposal of the research'

- Dashboard of data

```
#curve plot of logestique regression between cancer and nb of cigarefes
plot(Can~nbC,xlab="nbc",
ylab="can",
main="relation cancer nb of cigaretes")
```

Figure (02) shows the spread of the number of cigarettes to cancer



Source: prepared by the researcher based on the outputs of R

The prevalence panel of the variable A (cancer incidence) shows that it is of a dichotomous nature.

□ **Results of logistic regression related to the example**

```
mydata<-read.table(file.choose(data),header = T,sep = "")
head(mydata)
attach(mydata)
names(mydata)
logesmydata<-glm(Can~nbC)
summary(logesmydata)
exp(cbind("OR"=coef(logesmydata),confint(logesmydata)))
>logesmydata<-glm(Can~nbC)
summary(logesmydata)
Call:
glm(formula = Can ~ nbC)
Deviance Residuals:
Min       1Q       Median       3Q      Max
-0.8378  -0.4339   0.2199   0.2920   0.6094
Coefficients:
```

```
Estimate  Std. Error t   value Pr(>|t|)
(Intercept )     0.30403          0.26894      1.131        0.277
nbC            0.01443      0.01078            1.339           0.202
(Dispersion parameter for gaussian family taken to be 0.2374607)
Null deviance: 3.7500  on 15  degrees of freedom
Residual deviance: 3.3244  on 14  degrees of freedom
AIC: 26.265
Number of Fisher Scoring iterations: 2
```

Source: prepared by the researcher based on the outputs of R

Interpretation of the logistic model results

- **the overall significance of the model**

The validity of the total logistic regression model can be known by comparing the two values of the Null deviance and the Residual deviance, where the first reveals the differences before including the interpreted variable, while the second after it. As the difference increases, it indicates the validity of the model and its ability to predict the dependent variable. As for the AIC coefficient that came with 26,265, it has no significance as long as we are not in the case of multiple models.

- **likelihood function (-2LL) for the logistic relationship**

Table (04) shows the maximum potential function of the independent and constant variable,

Iritation	) -2 Log likelihood(	constant
1	19.353	0.784-
2	19.323	0.857-
3	19.323	0.960-
4	19.323	0.860-

Source: prepared by the researcher based on the outputs of Spss, 25

Maximizing the possibility function by repeating the calculation process, with the aim of reaching the maximum possible reduction of the difference between the original model and the model containing the explained variable, so that the absolute maximization is the one through which we reach to find a theoretical total match between the two models, allowing a high prediction process Where in our example the results indicated that the value of the greatest potential reached (19.323) after the (04) stage of iteration until the incident change became less than (0.001), and it stabilized at (19.323) after the second stage of iteration until the incident change became less than (0.001), which explains why there are no differences between the conceptual model and its data.

- **of hosmers hypothesis of the relationship of the number of cigarettes to cancer**

Table (05) statistic for Homes to walk to test the quality of conformity

Step

1	Chi-square	df	Sig.
	4.617	5	0.464

Source: Prepared by the student based on the outputs of Spss, 25

- **partial significance**

The significance of the dependent variable, p = 0.202, greater than 0.05. No cancer was explained as a result of a number
Smoked cigarettes.

o **Pseudo R-squared**

The analogous coefficient of determination expresses the extent of the variables' ability to explain their relationship with the dependent variable in the logistic model. The following table was reached through the programming language R.

```
>exp(cbind("OR"=coef(logesmydata),confint(logesmydata)))
                              OR                 2.5 %                97.5 %
             (Intercept) 1.355315       0.8000576          2.295935
```

```
nbC          1.014530    0.9933276        1.036185
```

Source: Prepared by the researcher based on the outputs of R

Interpretation of (ratio-odds) is by comparing its result obtained with the correct one. If it exceeds one, it indicates a steady positive correlation, while if its value is less than one this indicates a negative correlation between the two changes, as long as the index is a value Fractional comparison between an event and its complement event. Where the relationship (that is, the relationship of the variable to the dependent within its explanatory capacity) will be read, provided that smoking explained cancer within a probability of (1-1.014530) and equal to 0.0145 or 1.45%, but it should be noted that this reading will be correct only when there is a significant effect of the variable .

- ***Classification Table***

```
>predict<-predict(logesmydata,type = 'response')
>table(mydata$Can,predict>0.5)
  FALSE TRUE
 0      3       3
 1      2       8
```

Source: Prepared by the researcher based on the outputs of R

The model was able to classify eight individuals in their correct place and failed to classify two items. Whereas, three individuals with cancer were classified equally as non-infected, and in parallel, three individuals were classified as correct. Where the total ability of classification is determined by adding the diameter that represents sensitivity and accuracy, then dividing it by the degree of the overall classification as follows:

$$Sum(3+8)/(3+8+2+3)= 11/16=0.6875$$

As the overall rating rate of the model reached 68.75%, which is a good percentage that can be relied upon to predict the phenomenon.

- **Wald statestic of partial significance**

Table (06) logistic regression parameters for the effect of number of cigarettes on cancer incidence

varibales	Wald .value	P.value.Wald	B logite	SEE	p.value B
const	.558	.277	-.866	1.151	.423
nbs	1.699	.202	.064	.049	1.066

Source: Prepared by the researcher based on the outputs of R

wald statistic explained and fixed indicated its lack of significance in the interpretation of cancer (sig = 0.277). As for reading its significance in all cases, it depends on its reading within the aforementioned conceptual frameworks. That is, after raising its significance to the exp. We say that increasing the number of cigarettes by one unit increases the incidence of cancer by 〖$exp^{(0.064)}$. However, it should be noted that, in our example, and due to the lack of meaning of the explained variable, a relationship cannot be adopted as a statistically significant result.

VI. Conclusion

The aim of the study was not related to partial results about the study variables, but rather to determine the difference in the two methods of sediment analysis and the differences in the regression theory between the linear and logistic model. Where the linear model is based on examining the total significance of the statistic for the parameter F, and then for the explanation of the amount of the effect by which the variables entered the variable directly linearly. As for the analysis according to the logistic model, it requires the inclusion of more numerous and complex measures of the fact that the dependent variable is described as being a two-valued nominal variable, where the validity of the model is linked to the greatest possibility function 2 Log likelihood-Hosmer and Lemeshow's statistic, and a similar coefficient of determination, in addition to the classification tables. It is related to studying the log parameters of the variables and the significance of its wald statistic. It also enabled reliance on the software by moving from ready-made packaging methods to dealing with data within special orders.

V. II Bibliographical References

1. F. Tom F (An introduction to ROC analysis. Institute for the Study of Learning and Expertise), 2164 Staunton Court, Palo Alto, CA 94306, USA, 2005 p: 862.
2. c. Uzbon. (2018) (Best Practices in Logistic Regression), Institute of Public Administration, King Fahd National Scribes, Riyadh, p: 43.
3. p. Ahmad bin Hassan Babtain. (2017),(Logistic regression and how to use it in building forecast models for data). Research submitted to obtain a PhD in Statistics and Research. Umm Al-Qura University, Kingdom of Saudi Arabia, pp: 39-100
4. M. Subhi Abu Saleh, and Adnan Muhammad Awad. (1983). (Introduction to Statistics). Yarmouk, Jordan: John Wiley & Sons House., P .: 216
5. M. Abu Shukan, and Ibrahim Ali. (2014), (binary logistic regression model in the interpretation of the two-value dependent variables in the field of physical activities and sports). Journal of the Sciences and Practices of Sport and Artistic Activity's.

1. Adel l bin Ahmed Bin Hussein Babtain, Adel bin Ahmed .(2019) .Logistic regression and how to use it in constructing forecast models for bivariate data .Research submitted for a PhD in Statistics and Research.29-27 ‹
2. hicham Iade .(2021) .Econometrics, lectures and exercises .A pedagogical publication in econometrics .29-27 ‹Faculty of Economic Sciences, University of Tlemcen.

How to cite this article by the APA method:

MEDINI Atman ,AOUADI Mostafa (2021), Regression theory in linear and logistic model- With an illustrated example in programming language R, Journal of quantitative economics studies, Volume 07 (Number 01), Algeria: Kasdi Marbah University Ouargla, pp. 437-448.