International Journal of Mathematics and Statistics Studies
Vol.10, No.3, pp.39-48, 2022
Print ISSN: 2053-2229 (Print),
Online ISSN: 2053-2210 (Online)

Important Comparison About the Use of Multiple Linear Regression and Logistic Regression with Applications

AbuElgasim Abbas Abow Mohammed

College of Business and Economic Qassim University Kingdom of Saudi Arabia

Citation: AbuElgasim Abbas Abow Mohammed (2022) Important Comparison About the Use of Multiple Linear Regression and Logistic Regression with Applications, *International Journal of Mathematics and Statistics Studies*, Vol.10, No.3, pp.39-48

ABSTRACT: *This is paper presented a multiple linear regression model and logistic regression model, according to assumptions of both models. The paper depended on logistic regression model because the dependent variable is nominal. It also used preparatory year data collected from the College of Business and Economics, Qassim University to know the effect of student grades and gender, as independent variables, on the student status as dependent variable. The paper found that the grades have not a significant effect on the student status while the gender has a significant effect on the student status. The comparison between the multiple regression model and the logistic regression model showed that the logistic regression model is the most appropriate for determining the relationship between the student's status as a dependent variable and the students' grades, and gender as independent variables. The paper recommends the use of logistic regression, especially if the dependent variable is nominal.*

KEY WORD: Logistic regression, Multiple regression, significant effect, student's status، students' grades

INTRODUCTION

Regression analysis is one of statistical tools that utilize the relationship between two or more variables El-Habil, Abdallah (2012) [1]. Multiple Linear Regression (MLR) is an extension of linear regression there is a two or more independent variables affect the value of dependent variable, Multiple regression is highly useful tool in a wide range of applications from business, marketing, and sales analytics …ext. The MLR helps professionals evaluate diverse data that supports goals, processes and outcomes in many industries. In the multiple regression model, we assume that a linear relationship exists between some variable Y, which we call the dependent variable, and k independent variables, $X_1, X_2, \ldots . X_k$ .The independent variables are sometimes referred to as explanatory variables, because of their use in explaining the variation in Y. They also called predictor variables, because of their use in predicting Y.

The logistic regression is used widely to examine and describe the relationship between a binary or multiple response as dependent variable and independent variable. In general, we shall be interested in finding out how changes in the predictor variables affect the values of the response variables. In both models we will discuss the possible inferential procedures to provide a useful comparison.

Problem Statement

The aim of this study is to compare between multiple linear regression and logistic regression model, as the paper tried through this comparison to find out which is the best model for data representation, in side this try to know the effect of each of these variables: grades student and sex as explanatory variables, on the student status as predict variable, and what is the statistical significance of any of these explanatory variables?

The importance of this study stems from the use of both multiple linear regression and logistic regression

LITERATURE REVIEW

Abow, A.A (2020) [2] presented a comparison between a simple linear regression model and binary logistic regression model, used preparatory year student's data of Business and Economic College to know the effect of students grades and term level on student status, found that the grades have a significant effect on student. Rudd &Priestly (2017))[3] presented a comparison of decision tree with logistic regression model for prediction of worst – financial payment status in commercial credit, used both model to predict worst non-financial payment status among businesses, and evaluate decision tree model performance against traditional logistic regression model for task , paper found that decision tree performed as well as the logistic regression model.

METHODOLOGY

The study depended on theoretical approach that dealt with multiple linear regression and logistic regression model supported the practical side that based on preparatory year data of Business and Economics College in Qassim University Kingdom of Saudi Arabia, the study used SPSS for analyzing data

Assumption of multiple linear regression model

The assumption underlying multiple regression analysis are as follows

0. The X_i are nonrandom(fixed) variables, this condition indicate that any inferences that are drawn from sample data apply only to the set of X values observed and not to some larger collection of X's.
1. For each set of X_i value there is a subpopulation of Y values. To construct certain confidence intervals and test hypotheses it must be known, or the researcher must be willing to assume, that these subpopulations of Y values are normally distributed.
2. The variances of the subpopulations of Y are all equal.

3. The Y values are independent. That is, the values of Y selected for one set of X values do not depend on the values of Y selected at another set of X values. There are rules for assumption that are as follows:

a. interprets the impact of each independent variable relative to the other variable in the model, because model respecification can have a profound effect on the remaining variables:
b. use beta weights when comparing relative importance among independent variables
c. regression coefficients describe changes in the dependent variable, but can be difficult in comparing across independent variables if the response formats vary
d. multicollinearity may be considered "good" when it reveals a suppressor effect, but generally it is viewed as harmful because increases in multicollinearity
e. reduce the overall R^2 that can be achieved
f. confound estimation of the regression coefficients
g. negatively affect the statistical significance tests of coefficients
h. generally accepted level of multicollinearity (tolerance values up to.10, corresponding to a VIF of 10) almost always indicate problems with multicollinearity, but these problems may also be seen at much lower levels of collinearity and multicollinearity:
i. bivariate correlations of .70 or higher may result in problems, and even lower correlation may be problematic if they are higher than the correlations between the independent and dependent variables
j. values much lower than the suggested thresholds (VIF values of even 3 to 5) may result in interpretation or estimation problems, particularly when the relationship wish the dependent measure are weaker.

Multiple linear regression model

the multiple regression model written as follow

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_k x_{kj} + e_j \quad (1)$$

Where y_j is a of typical value from one of the subpopulation Y values, β_j are called the regression coefficients, $x_{1j}, x_{2j}, \ldots, x_{kj}$ are, respectively, particular values of the independent variables $X_{1j}, X_{2j}, \ldots, X_{kj}$ and e_j is a random variable with mean zero and variance σ^2, the common variance of the subpopulation of Y values. To construct confidence intervals for and test hypotheses about the regression coefficients, we assume that the e_j are normally and independently distributed. The statement regarding e_j are a consequence of the assumptions regarding the distribution of Y values. We will refer to Equation (1) as the multiple linear regression model. Wayne, Daniel [4]

The regression model is an extension of the model for a single X. For each particular set of values of $X_1, X_2, \ldots, X_k$ there exists a population of Y's. Each population of Y's is normally distributed; the mean of the various population lies on a plane, and the variances of the population are equal. The plane on which the population mean lie is called the population regression plane. It can be written

$$E((Y|X_1, X_2, \ldots, X_k) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (2)$$

Where α is the height of the regression plane the point $X_1 = 0, \ldots, X_k = 0$. The parameters $\beta_1, \ldots . \beta_k$ are the partial regression coefficients. If $\beta_1 = 2$, for example, the height of the regression plane increases by two units if $X_1$is increased by one unit and $X_1, \ldots, X_k$ are kept unchanged Dunn Clark [5]

Assumption of logistic regression

The assumption underlying multinomial logistic regression analysis are as follows

1. Multinomial logistic regression assumes that the observations are independent
2. Multinomial logistic regression also assumes that the natural log of the odds ratio and the measurement variables have a linear relationship. It can be hard to see whether this assumption is violated, but if you have biological or statistical reasons to expect a non-linear relationship between one of the measurement variables and the log of the odds ratio you may want to try data transformation
3. Multinomial logistic regression does not assume that the measurement variables are normally distributed
4. Multinomial logistic regression assumes no multicollinearity that mean, when two or more of independent variable are substantially correlated among each other Benotti et al [6].

In addition to some rules, it can be mentioned as follows:

1. model significance tests are made with a chi-square test on the differences in the log likelihood values (-2LL) between two models
2. sample size considerations for logistic regression are primarily focused on the size of each group, which should have 10 times the number of estimated model coefficients
3. sample size requirements should be met in both the analysis and holdout samples
4. coefficients are expressed in two forms: original and exponentiated to assist in interpretation
5. interpretation of the coefficients for direction and magnitude is as follows:

a. direction can be directly assessed in the original coefficients (positive or negative sign) or indirectly the exponentiated coefficients (less than 1 are negative, greater than 1 are positive)
b. magnitude is best assessed by the exponentiated coefficient, with the percentage change in the dependent variable shown by:
percentage change = (Exponentiated coefficient -1)×100

Logistic regression model

Use logistic regression when you have one a nominal variable and two Or more measurement variables, and you want to know how the measurement variable affect the nominal variable you can use it to predict, probabilities of dependent nominal variable or if you are carful, you can use it for suggestion about which independent variables have a major effect on the dependent variable.

The logistic regression model and other General Linear Models (GLM), like ordinary regression models for normal data, generalize to allow for several explanatory variables. The predictors can be quantitative, or of both types.

Denote a set of k predictors for a binary response Y by $X_{1j}, X_{2j}, \ldots, X_{kj}$ model (3) for the logit of the probability π that Y=1 generalizes to

$$logit(\pi) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (3)$$

The parameter β_i refers to the effect of X_i on the log odds that Y=1, controlling the other Xs. For instance, exp (X_i) is the multiplicative effect on the odds of 1 –unit increase in X_i , at fixed levels of the other Xs. Alan Agresti [7]

The goal of a logistic regression is finding an equation that the best predicts the probability of a value of the Y variable as a function of the X Variable Log Jian Liu [8]

A comparison between multiple linear regression and logistic regression models

In the following table, we show the correspondence between concepts used in multiple regression and their counterparts in logistic regression

Table1.Correspondence both use Multiple regression and logistic regression

Correspondence of primary elements of model fit	
Multiple regression	**logistic regression**
Total sum of squares	-2LL of base model
Error sum of squares	-2LL of proposed model
Regression sum of squares	Difference of -2LL for base and proposed models
F-test of model fit	Chi-square test of -2LL difference
Coefficient of determination (R^2)	Pseudo R^2 measures

As we can see, the concept between multiple regression and logistic regression are similar. The basic approaches to testing overall fit are comparable, with the differences arising from the estimation methods used in the two techniques Hair, Black, Babin, Anderson (2010) [9]. There are two primary reasons for choosing the logistic distribution. First a mathematical point of view, it is an extremely flexible and easily used function. Second it lends itself to a clinically meaningful interpretation. A detailed discussion of the interpresent the mean of y given x when the logistic regression model we use is

$$\pi(x) = \frac{e^{\beta_0+\beta_i x}}{1+e^{\beta_0+\beta_i x}} \quad (4)$$

A transformation of $\pi(x)$ that is central to our study of logistic regression is the logit transformation. This transformation is defined, in terms of $\pi(x)$ as

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] \quad (5)$$

The importance of this transformation is that $g(x)$ has many of the desirable properties of a linear regression model. the logit, $g(x)$, is a linear in its parameters, may be continuous and may range from $-\infty \ to + \infty$ depending on the range of x. The second importance difference between the linear and logistic regression models concerns the conditional distribution of the outcome variable.in the linear regression model we assume that an observation of the outcome variable may be expresses as$y = E(y/x_i) + \epsilon$ The quantity ϵ is called the error and expresses as observation deviation from the conditional mean. The most common assumption is that ϵ follows normal distribution with mean zero and some variance that is constant across levels of independent variable David, Hosmer (2013) [10].

Fitting the logistic regression model

The Conditional distribution of outcome variable given x will be normal with E(y/x_i), and a variance that is constant. This is not the case with a dichotomous outcome variable given x as$y = \pi(x) + \varepsilon$. Here the quantity ε may assume one of two possible values. If $y = 1$ then $\varepsilon = 1 - \pi(x)$ with probability $\pi(x)$, and if $y = 0$ then $\varepsilon = -\pi(x)$ with probability $1 - \pi(x)$. Thus ε has a distribution

with mean zero and variance equal to $\pi(x)[1-\pi(x)]$. That is, the conditional distribution of the outcome variable follows a binomial distribution with probability given by the conditional mean,$\pi(x)$. In summary, we have seen that in a regression analysis when the outcome variable is dichotomous:

(1) The conditional mean of regression equation must be formulation to be bounded between zero and 1 we have stated that the logistic model, $\pi(x)$ given in equation (4) satisfies this constraint.
(2) The binomial, not the normal distribution describes the distribution of the errors and will be the statistical distribution up on which the analysis is based.
(3) The principles that guide an analysis using linear regression will also guide us in logistic regression.

Fitting the logistic regression model

Suppose we have a sample of independent observation of fair (x_i,y_i) , $i = 1,2 \cdots ,n$ where $y_i$denotes the value of a dichotomous outcome variable and x_i is the values of the of the independent variable for the i^{th} subject, furthermore, assume that the outcome variable has been coded as 0 or 1, representing the success or the failure of the characteristic, respectively, this coding for a dichotomous outcome is used through-out the text. To fit the logistic model in equation (4) to a set of data requires that we estimate the values of β_0 and β_i the unknown parameters.

In linear regression the method used most often for estimating un known parameters is least square. In that method we choose those values of β_0 and $\beta_i$which minimize the sum of squared deviations of the observed values of y from the predicted values based up on the model under the usual assumption for linear regression the method of the least square yields estimator with a number of desirable statistical proper Dunn, Oliver, Clar, Virginia (1987)[11]

Deviance and likelihood ratio tests

In linear regression analysis, one is concerned with partitioning variance via the sum of squares calculations –variance in the criterion is essentially divided in to variance accounted for the predictors and residual variance. In logistic regression analysis, deviance is used in lieu of a sum of squares calculation Cohen, Aiken, Leona S (2002) [12]. Deviance is analogous to the sum of square calculation in linear regression and is a measure of the lack of fit to the data in a logistic model. When a (saturated) model is available (a model will a theoretically perfect), deviance is calculated by comparing a given model with the saturated model. This computation gives the likelihood- ratio test

$$D = -2ln\frac{likelihood\ of\ the\ fitted\ model}{likelihood\ of\ the\ saturated\ model} \qquad (6)$$

If the model deviance is significantly smaller than the null deviance the one can conclude that the predictor or set of predictors significantly improved model fit. This is analogous to the F-test used in linear regression analysis to assess the significance of prediction Hosmer, Stanley (2000) [13].

Wald Statistic

Alternative, when assessing the contribution of individual predictors in a given model, one way examines the significance of the Wald statistic. The Wald statistic, analogous to the t-test in linear regression, is used to assess the significance of coefficients the Wald statistic is the ratio of the square

of the regression coefficient to the square of the standard error of the coefficient and is asymptotically distributed as Chi-square distribution

$$W_j = \frac{{\beta_j}^2}{SE^2{}_{\beta_j}} \quad (7)$$

The Wald statistic has limitations when the regression coefficient is large, the standard error of the regression coefficient also tends to be larger increasing the probability of type-II error the Wald statistic also tend to be biased when data are square Menard, Scott W (2002) [14].

RESULTS AND DISCUSSION

In this section, the paper collected data of a sample size of 633 students grades of the preparatory year from the College of Business and Economics at Qassim University, as shown in the table1bellow in Table 2.

Table2. Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	633	100.0
	Missing Cases	0	.0
	Total	633	100.0
Unselected Cases		0	.0
Total		633	100.0

From Table3 and Table4 the paper found that R- square value is 0.636 while the coefficient determination of multiple linear regression model in a manner similar to Pseudo R square values Cox and Snell is 0.597 Nagelkerke 0.848 of logistic regression model. However, the indicators differ between the two models as the coefficient of determination in the multiple linear regression model indicates the percentage of variation in the dependent variable, while in the logistic regression model it is indicates that the model is fit. Also, this indicates the convergence of measures in the two models.

Table3. Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.797[a]	.636	.635	.282

Table4.Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	195.221[a]	.597	.848

From Table 5 of ANOVA the multiple regression employs the method of least square, which minimizes the sum of the square differences between the actual and predicted values of the dependent variable while the maximum likelihood estimation procedure fits are the likelihood value, similar to sums of squares value of -2 times the log of the likelihood value referred to as -2LL or -2 log likelihood from table5, multinomial logistic regression maximizes the likelihood that an event will occur. The likelihood value instead of the sum of square is then used when calculating a measure of overall model fit. From below tables, we find information for both model fitting F.test=549.922 and sig=0.00 From

table5 of multiple linear regression corresponding the Chi-Square=214.753 and sig=0.00 from table6 of Hosmer and Lemeshow test that we assess model fit in different ways.

Table5. Test Hosmer and Lemeshow

Step	Chi-square	df	Sig.
	214.753	8	.000

Table6. ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	87.303	2	43.652	549.922	.000
	Residual	50.008	630	.079		
	Total	137.311	632			

Logistic regression test hypotheses about individual coefficients have been calculated based on multiple linear regression in multiple whether the multiple regression coefficients are different from 0. The paper found that $\beta_1 = 0.100$ and sig= 0.00 inspect to degrees of student variable. Therefore, the $\beta_2 = 0.022$ and sig = 0.00 inspect to gender variable. this is referring to coefficient β_1 has an impact on the dependent variable, and coefficient β_2 has an impact on the dependent variable too. Thus, grade of student and gender have a signified affect to the student status(β_i are parameters the regression coefficient). However, in logistic regression coefficient values for some group indicate that no effect of the independent variables on predicting group membership logistic regression uses a different statistic, the Wald statistic. The paper provides the statical significance for each estimated coefficient. So that hypothesis testing can occur just as it does in multiple regression if the logistic coefficient is statistically significant can interpret it in term of how it impacts the estimated probability and thus the prediction of group membership (according to Table8 we find that values of Wald statistic is tend to be biased, because the regression coefficient is large, the standard error of the regression coefficient also tends to be larger increasing the probability of type-II error and its sig =0.00 refer to effect sex to the students status).

Table7. Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
	B	Std. Error	Beta	t	Sig.	Tolerance	VIF
(Constant)	-.541-	.047		-11.439-	.000		
	-.100-	.024		-4.220-	.000		
Students grades	.022	.001	-.107-	32.504	.000	.897	1.115
Sex			.825			.897	1.115

Table8.Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Students' grades	-.610-	.397	2.362	1	.124	.543	.249	1.183
	-.342-	.035		1	.000	.710	.663	.761
Sex	19.993	2.191	95.601	1	.000	48164546		
Constant			83.276			6.711		

CONCLUSION AND RECOMMENDATION

From the results of the multiple regression and logistic regression analysis, we found that the sex variable effect of the probability of student status and indicate that the sex has significant effect on student status, on the other hand the students grade not significant effect of the student status at 5% level of significance.

By comparing between multiple regression model and logistic regression model paper focus on study the effect of variables grades' student and sex on the student status using binary logistic regression, according to results found that the students' grades are not effects on student status. This result is logical because the grades greater than 60 all are pass considered as the same is one status. The comparison of multiple regression and logistic regression showed that the logistic regression is more stable to determine the relationship between the student status as dependent variable and students' grades, sex as independent variables The paper recommends the use of logistic regression, especially if the dependent variable is nominal.

References

[1] Abdalla El-Habill. (2012). An Application of Multinomial Logistic Regression Model, pak.j.stat.paper.res, pp271

[2] Abow, A.A (2020), Inference About the Use of linear regression and logistic regression, International Journal of Recent Scientific Research, India, Issue 8, vol 11p 2.

[3] Rudd. Jessica M.MPH,GStat & Priestly .Pennifer L (2017, A Comparison of Decision Tree with logistic Regression Model for Prediction of Worst Non– Financial Payment Status in Commercial Credit,Grey Literature from PhD Candidaters.5.http://digitalcommons.kennesaw.edu/dataphdgreylit/5

[4] Wayne W. Daniel, (2005), Biostatistics a Foundation for Analysis in the Health Sciences, Eighth Edition, John Wiley and Sons, United State of America, P 488-489.

[5] Dunn, Olive Jean Clark, Virginia, (1987)," Applied Statistics: Analysis of Variance and Regression", 2nd Edition, John Wiley and Sons, New York

[6] Benotti, (2014), " Risk factor associated with mortality after Roux-en-Y gastric bypass surgery. A nals of surgery259: 123-130

A Global Perspective"(7nd) pearson Education, United State of American .421

[7] Alan Agresti, (1996), An Introduction to Categorical Data Analysis, John Wiley and Sons, New York, p 122

[8] Log Jian Liu, 2018, MD, PHD, MSC (LSHTM), FAHA, in Heart Failure: Epidemiology and Research Method

[9] Hair, Joseph. Black, wellian. C, Babin, Barry J, Anderson, Ralph. E, (2010), " Multivariate Data Analysis

[10] David, W. Hosmer, (2013), Applied Logistic Regression, Third Edition John Wiley and Sons, New York

[11] Dunn, Oliver Jean and Clark, Virginia. A, (1987), Applied Statistics: Analysis of Variance and Regression, 2nd Edition, John Wiley and Sons, New York

[12] Cohen, Jacob, Cohen, Patricia West, Steven G; Aiken, Leona S (2002) Applied Multiple Regression/ Correlation Analysis for the Behavioral Sciences (3rd ed). Routledge.

[13] Hosmer, David W; Lemeshow, Stanley (2000). Applied Logistic Regression (2nd ed) Wiley.

[14] Menard, Scott W. (2002). Applied Logistic Regression (2nd ed). SAGE, International ISBN Agency P 99-100