Common Errors in Regression Analysis

Dr. Khalifa H. Jaber

Al Ain University of Science and Technology

Introduction

Regression analysis is an applied statistical method used to analyze data and make statement about variables after having controlled for values of known explanatory variables. Many applications of regression analysis were noticed in several studies like business, education, finance and others.

Therefore the challenge here is to select a statistical model that is congruence with the data and the underlying statistical theory.

In this article, we tried to identify some errors committed in the regression analysis. These errors are rarely noticed by researchers who use this technique especially those who have little knowledge in statistics although they might obtain completely different results from the reality.

Some of these mistakes need to be explained and to be avoided by the people who use the regression analysis in order to overcome conceptual and theoretical misunderstanding behind it.

We will explain some of these mistakes related to the following aspects:

The stepwise least regression, the regression coefficients, the coefficient of determination, the regression analysis of binary data and the model selection.

Finally some questions will be raised about the regression analysis and the answers for them will be given.

1) **The stepwise least Square**:

This method is also called the regression on residual (ROR) which is different from the well-known method stepwise regression, which we will discus later in this article.

Suppose we have the multiple regression model:

$$Y= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e \quad \ldots\ldots\ldots\ldots (1)$$

Where e is the error term with mean = 0 and variance = σ^2

$\beta_0, \beta_1, \beta_2$ are unknown parameters .

The least square method can be used to estimate the above parameters to obtain the regression equation :

$$Y= b_0 + b_1 X_1 + b_2 X_2 + e \quad \ldots\ldots\ldots\ldots (2)$$

This equation, which is based on sample estimates, is used to make inference to the population parameters in equation (1).

Now in some cases the method of stepwise least square is used to estimate one parameter like β_1 from the regression equation :

$$Y= b_0 + b^*_1 X_1 + e_1 \quad \ldots\ldots\ldots..(3)$$

Where b^*_1 is the first ROR estimator

Then use another equation ,which can be obtained by regress the error term in equation (3) on the second explanatory variable

$e_1 = b^*_2 X_2 + e_2$ ……(4)
Where b^*_2 is the second ROR estimator .
Now the problem is that b^*_1 from equation (3) is not the same as b1 in equation (2) since equation (3) does not control X_2.Therefore it is a biased estimator for β_1 unless either $\beta_2=0$ or X_1 and X_2 are independent and hence b^*_2 is also biased estimator for β_2 because the residual in equation (3) is calculated from the biased estimator b^*_1.
In order to estimate the regression coefficients correctly ,both variables X_1 and X_2 should be put in the regression model simultaneously.

2) **The comparison of the regression coefficients**:

Suppose that a researcher wants to explain the dependent variable y on some explanatory variables such as X_1 and X_2 and wants to make a statement about the comparison of the effect of X_1 and X_2 on Y.
This depends on the researcher stated question to be explained clearly in order to make the comparison meaningful.
If there is a common unit of measurement for X_1 and X_2 ,then the comparisons make sense like a number of working hours or a piece of cake.
Now if the units of measurements for both variables are different, then their comparison will be meaningless although both regressors explain the same dependent variable.
In addition to that ,there is another problem ,that makes the comparison difficult if not logically impossible. The X_1 coefficient represents the effect of X_1 with holding the effect of the other variable (X_2) as constant. Also the estimated coefficient for X_2 has a different set of control variable since it includes X_1 not X_2 .
To overcome the above problem, some propose the standardization of the regression coefficients .Now because the comparison of the above variables is not meaningful, then it is useless to compare standardized variables. The standardization does not add any information because the original data were meaningless.
To replace the unmeasurable by the unmeaningful is not a progress (Achen 1977)
So the standardization makes the regression coefficients more difficult to explain and doe not make sense in the comparison if one said that the increase of one standard deviation in X_1 leads to increase of b_1 standard deviation in the dependent variable Y
The other important point is that ,the standardized variables measure the relationship between the explanatory variables and dependent variables and measure the variance of the independent variables while the researches usually interest only in the relationship.
Therefore ,we can say that the standardized coefficients are more difficult to explain and do not add any information that might be helpful to compare the influence of the explanatory variables on the dependent variable and finally it may give misleading information.
Finally if one must include a variable that is difficult to interpret as a control ,then perhaps standardizing just this variable would capitalize on the standardized coefficient's simpler descriptive properties (Blalock, 1967a).So this standardization is better than standardizing all variables.

3- **The Coefficient of Determination**:

The coefficient of determination is the square of the correlation coefficient, usually denoted by (R^2) . It is a measure for the specification of the regression model in the regression analysis. Others consider it as a measure of the influence of the independent variable(X) on the dependent variable (Y) or a measure of the fit between the true model and the suggested model, where a large value of R^2 is considered as a good indication for the chosen model.
All the above interpretations are not correct because R^2 is designed to measure something else. it is a measure of the spread of the points around the regression line. Where R^2 is the ratio between the sum of squares due to the regression line to the total sum of squares.
R^2 Can also be used to compare between two regression equations for the same dependent variable and different independent variables. So it is like a measure for goodness of fit.
Therefore R^2 is a measure for the proportion of variance explained because it explains the proportion of the reduction error in the current model compared with the theoretical model. So large R^2 accompanied by small variances, large regression coefficients and short confidence intervals. So R^2 is useful but we should knows, when we can use it .
The researchers can get higher R^2 but, it does not mean that, the chosen model is good and it does not add anything to the analysis. The following actions can be taken to increase the value of R^2 :
1) Choose independent variables that are similar to the dependent variable.
2) Increase the number of independent variables. As the number of variables approaches the number of observations, R^2 becomes close to 1.
3) Add or delete observations to increase the variance of independent variables.

In addition to all the above, there is no statistical theory behind the coefficient of determination because there is no R^2 for the population.

4- **The regression Analysis for Binary Variables**

In some cases, there are binary independent or dependent variables. There is big confusion in the analysis of such variables.
For the binary dependent variable ,the expected value of that variable is a probability and that supposed to be between zero and one. The fitted values are supposed to be desirable approximations of expected values. However, it is possible that the fitted values may be less than zero or greater than one, resulting in unreasonable approximations of the expected values.
Another drawback of the least square regression model is that patterns in the residuals are always evident when the response is binary.
That is the ith residual is :

$$\hat{e}_i = y_i - \hat{y}_i = \begin{cases} -\hat{y}_i & \text{if } y_i=0 \\ 1-\hat{y}_i & \text{if } y_i=1 \end{cases}$$

So the scatter plot of the standardized residuals versus the fitted values will be two parallel lines correspond to the cases of Y=0 and Y=1 respectively.
Therefore the plot is much less informative than the usual case when Y is continuous, so it does not help us to formulate an improved linear regression model.
Finally the binary nature of the response induces heteroscedasticity. The transformation is usually used to improve the estimation procedure like the logarithmic transformation, but the transformation is not appropriate for the case of a binary dependent variable.

The alternative approach for estimation the parameters in this case is the weighted least square.
It is worth mentioning that there is misunderstanding for the correlation coefficient in case of dichotomous independent variable.
The sample correlation coefficient (r) is considered as an estimator for the population correlation coefficient (ρ) ,which is one of five parameters for bivariate normal probability distribution . In this case the variables X and Y should be drawn from the above distribution and the marginal distribution of one of these variables is normal.
Now the above assumptions are not possible for the binary independent variable.
So one should not relate the causal assumptions to regression coefficients due to the correlation coefficient because these assumptions are not required in the regression analysis.

5- **The model selection** :

The well known method for bringing explanatory variables into the regression model is the stepwise regression.
This procedure employs a series of t-test to check the significance of explanatory variables entered into, or deleted from the model.
This method has several drawbacks, some of them are as follows :

1) It is an automatic procedure which depends mainly on computer algorithms to choose the variables and does not take into account an investigator's prior knowledge about these variables.
2) The algorithm does not consider models that are based on nonlinear combinations of explanatory variables.
3) By considering each variable separately, the method does not take into account the joint effect of independent variables and might neglect some correlated variables. This will lead to bad effect on the estimation of the parameters.
4) Because in this procedure there is a sequence of significance tests, the significance level that determines the t-value is meaningless.

In order to overcome the above drawbacks, some action can be taken to do so.
For example, the first drawback can be addressed by using some statistical software have options for forcing variables into the model. So if one or more variables should be included in the model, then one can force the inclusion of these variables.
For drawback three ,Bendel and Afifi (1977) suggest using a cut-off smaller than you ordinarily might .For example ,instead of using a t-value =1.96 corresponding to a 5% significance level , consider using a t-value=1.645 corresponding to a 10% significance level . In this way, there is less chance of screening out variables that might be important.
Finally ,the combination of variables may be deleted by using the backwards selection algorithm because this procedure starts with all variables and will detect and retain variables that are jointly important.

In addition to all the above mistakes, there are some other questions that might be raised about the regression analysis :
-Is the chosen model applicable for whole range of each variable or for a part of it ?
In fact, the regression equation may be used for interpolation within the range of measured values but not to make predictions for some conditions that not previously investigated.
-Should we depend only on the statistical test to determine the kind of relationship between the variables?

The answer is that we should also depend on the mechanisms that give rise to the data, which might determine the relationship between the variables. So the best way to make sure about the relationship is to plot the data before take the decision about the model.

-Is there any effect for an additional variable other than the explanatory variable X on the dependent variable Y?

If there is, then the additional effect should be taken into account by using multivariate regression equation. We should not depend only on the correlation to determine if there is a relationship between the two variables X and y because both might depend on a third variable.

-Are the assumptions underlying the statistical analysis met?

We should make sure that our assumptions are met before we make inference about regression coefficients.

The estimates of the coefficients depend upon the chosen model (linear, nonlinear, logistic,..) and also on the assumptions related to the error terms in the model .

-Is the data appropriate for regression calculation?

Sometimes if the data has an outlier value, then this value should be excluded and the recalculate coefficients with its confidence intervals.

-Is the data need transformation in order to be in a linear form ?

If a log transformation is used, then even the error in the original equation is normal but unfortunately, the new error term after transformation is not and the resulting confidence intervals need to be adjusted. (Zhou, Gao 1997).

References

1- Achen, Christopher H. (1977) Measuring Representation; Perils of the correlation coefficient.
American journal of political science, 21 Nov. 805-15

2- Agostinelli, C(2002) Robust stepwise regression, J. of applied statistics vol. 29, No.6, p. 825-840

3- Antoniadis, A. and Leblanc, F. (2000), Nonparametric Wavelet Regression for Binary Response.
J. of theoretical and applied statistics, vol. 34, Issue 3, p. 183-213

4- Blalock, Hubort M. (1967a) Casual Inference, closed populations and measures of association
American political science review 61 (March) 130-36

5- ChutAhu, Sin-Ho, Seung-Hokang (2002), Modified Regression Coefficient Analysis for Repeated Binary Measurements
J. of applied statistics vol. 129, issue 5, p. 703-710

6- Draper, N.R. and Smith H. (1998), Applied Regression Analysis, J. Wiley, New York

7- Good, I. J. (1989), Interpretation of a Large Coefficient of Determination, J. of statistical computation and simulation, vol. 31, issue 1, p. 63-64

8- Goldberger, Arthur S. (1961) Stepwise Least Squares: Residual Analysis and Specification Error
J. of the American statistical association 56, (Dec.) 998-1000

9- Hargens, Lowell L. (1976), A note on Standardized Coefficients, Socialogical Methods and Research, 5 (Nov.) 247-56

10- Leon, J. Glesser, (1981): Estimation in a Multivariate "Errors in Variables' Regression Model: Large Sample Results", the annuals of statistics vol. 9, no. 1,p. 24-44

11- Qaqish, B.F. (2003), A Family of Multivariate Binary Distributions for Simulating Correlated Binary Variables with Specified Marginal Means and Correlations, Biometrika 90, p. 455-463

12- Qinggang, W. Ebal (2008), Determination of the Selection Statistics and Best Significance Level in Backward Stepwise Logistic Regression
Communications in statistics simulation and computation, vol. 37, issue 1, p. 62-72

13- Ronchetti, E (1997) Robustness Aspects of Model Choice, statistica sinica 7, p. 327-338

14- Routledge, R. D. (1990), When Stepwise Regression Fails: Correlated Variables Some of Which are Redundant
Int. J. of mathematical education in science and technology, vol. 21, issue 3, p. 403-10

15-Zhou, X-H, Gae, S. (1997), Confidence Intervals for the Log Normal Mean . Stat. Med., vol. 17, p. 2251-2264