Middle-East Journal of Scientific Research 13 (Mathematical Applications in Engineering): 97-102, 2013
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2013.13.mae.99999

Unavoidable Conceptual Assumptions for Regression Analysis

Jamal I. Daoud

Department of Science in Engineering, International Islamic University Malaysia

Abstract: Statistical Analysis has become an indispensable tool for most of researches. Regression analysis is one of most widely known and used statistical tools for analysing multifactor data. Nowadays, it is hard to find statistical analysis without involvement of regression analysis. It is powerful tool and at the same time it is easy and clearly showing and describing the relationship between different variables associated in a certain relationship. One of the important issues when using the regression analysis, are the assumptions made on the model based on the sample drawn from a population. To manufacture a good device, the components of the device should be manufactured in such a perfect way so that it will yield the maximum satisfactory properties, as well as working in harmony with other components which maximizes the reliability of the device. Then the device can be produced in mass production. The issue with assumptions on regression model is similar to the device reliability. The sample based model has to be verified and then used to make inferences on the population from which the sample has been obtained. In this paper the general concepts will be demonstrated without going in details on computations and calculations.

Key words: Regression % Multicollinearity % Homogeneity % Heteroscedasticity % Normality % Independence

INTRODUCTION

In different fields of science, large number of researchers is used to use statistical analysis and many of them are involved in using regression analysis and models. For statisticians should be (or at least should be) obvious to check the model suitability and accuracy by checking the validity of standard assumption in the model they are dealing with as part of built-in their scientific background.

Mostly they will make sure that the model is meeting certain standard criteria (assumptions). Sometimes researchers with different background (no statistical) may overlook or may not pay the deserved attention to these assumptions, thinking, that this is not necessary formality that can be skipped. As a result they may adapt or rely on a model including one or more violation of these standard assumptions. This will lead to wrong conclusions and understanding about the regression analysis of the population.

The ultimate objective of the regression model based on sample observation is in fact to try to project the findings on the population so that the researcher can describe or come close to the actual regression model for the population which is not always possible due to time, cost and manpower limitations. The model obtained from the sample will describe the relationship between the variables in the sample exactly, but this does mean that the same description is applicable for the population, simply because the model based on part of the population and not all population data. For this reason, it is crucial for the researcher to make sure the his model is meeting the conceptual assumption which will contribute in many ways to accuracy of the estimation of population parameters, consequently, the decision that might be made will touch the real point of the problem under study. In fact, most of researches are trying to find solutions for problems, then good information will lead to good decisions and poor information will lead to marginal and lousy decisions. That is why, it is vital to make sure that obtained model is enough solid (reliable) to extend the findings to the population with high degree of confidence.

To assure, that the model solid and reliable, researcher have to test and check the validity of these assumptions in the model.

Corresponding Author: Jamal I. Daoud, Department of Science in Engineering, International Islamic University Malaysia.

In mathematics and may be in life, there are two types or relationships deterministic and stochastic (probabilistic). The following relation (equation) is an example on deterministic one, let *r* be the radius of a circle, then the area *A* of that circle can be found by the formula;

$$A = \pi r^2$$

This relation is fixed or constant and every time when applying the formula we get the same area, regardless to the time and place where we are finding this area, that is because the area is under the influence of the radius only and no other factor(s) can affect (predict) it. On the other hand consider the relationship given in the following equation;

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

This is an example on simple linear regression (stochastic) models. In this model we can find two major components that are deciding (predicting) the destiny of the response, namely controllable component ($\beta_0 + \beta_i x_i$) and non-controllable component (ε_i) which is called the random error term and therefore, each time when we try to find the value of the response variable (Y_i), we are not sure on the value of the response value due to the impact of the random error term which the out of control. This error term in fact, is mostly, the source of all problems (violations) meanwhile; it is the beauty of regression analysis. Hopefully we have more impact for the controlled component that the one out of control. To measure the impact of the controlled part we use an indicator called coefficient of determination (R^2), which is well known to people using regression analysis.

It is good and desirable to find a model that is accurate and meeting all theoretical assumptions and conditions, but it seems we more likely are happy when dealing with a model that has problems. Trying to study the reasons of such problems and overcoming it (if possible) will lead to some self-satisfaction and confidence as human being nature. That is, in fact what is the research all about?

Model Criticism and Selection: The validity of a statistical method, such as regression analysis, depends on certain assumptions. Assumptions are usually made about the data and model [1].

The accuracy of the analysis and conclusion derived from the analysis depends crucially on the validity of these assumptions.

First we need to determine whether the specified assumptions hold. Then many questions to be addressed as following [2]:

- What are the required assumptions?
- For each of these assumptions, how do we determine whether or not the assumption is valid
- What can be done in cases where one or more of the assumptions does not hold?

Assumptions to Take Care of: There is a huge number of references discussing in details and demonstrating with examples these assumptions. Some are listing them in four or five assumptions, other are grouping them in many groups according to the interrelationship among factors in the same group.

One of the very important and useful such classifications are as following:

Assumptions about the Form of the Model: The model that relates the response to the predictors is assumed to be linear in the regression parameters.
Which implies that the i^{th} observation of the response can be written as?

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + ... + b_p X_{pi} + e_i \qquad i = 1,2,...n$$

This is known to be the linearity assumption.

Checking the linearity of simple regression model can be done by the scatter plot of y versus x. In multiple regressions, checking the linearity is more difficult due to the high dimensionality of the data. When the linearity assumption does not hold, transformation of the data can sometimes lead to linearity.

Assumption about the Errors: The random errors of the model ε_i, $i = 1,2,...,n$ are assumed to be independent and identically normally distribution with mean zero and an unknown variance σ^2. This implies that, The error terms are normally distributed; this is the assumption of normality. The violation of normality cannot be detected easily, but the residual plots can give an idea about the validity or violation of this assumption. The errors have mean zero and unknown variance σ^2. Lastly, the errors are assumed to be independent of each other, which mean that their pairwise covariance are zero. This is the errors-independent assumption. When this assumption is violated, the problem called autocorrelation problem.

Assumptions about the Predictor (s): There are three assumptions concerning the predictor variables:

- The predictor (independent) variables are non-random, that is, the values are assumed fixed or selected in advance. This assumption is satisfied only when the experimenter can set the values of the predictor variables at predetermined levels. When the predictors are random variables, all inferences are conditional, conditioned on the observed data. In fact this assumption is almost unknown or forgotten by majority of researchers.
- The values of independent variables are measured without error. This assumption is difficult (if not impossible) to meet. The errors in measurement will affect the residual variance, the multiple correlation coefficients, of the and the individual estimates of the regression coefficients. The exact amount of the effect will depend on several factors, the most important of which are the standard deviation of the error of measurement and the correlation structure among the errors. The effect of the measurement errors will be to increase the residual variance and reduce the magnitude of the observed multiple correlation coefficient. The effects of the measurement errors on individual regression coefficient are more difficult to assess. The estimate of the regression coefficient for a variable is affected not only by its own measurement errors, but also by the measurement errors of other variables included in the model.
- The predictor variables are assumed to be linearly independent of each other. If this assumption is violated, the problem is referred to as the collinearity problem.

Assumptions about the Observations: This is a general assumption, which states that, observations are equally important and reliable in determining the regression results. Which means that model should contain variables that have causal relationship; the absence of causal relationship will make the model unrealistic in interpreting results and make the necessary inference about the population.

Why the Assumptions Are Vital?: Many of researchers are not paying the deserved attention to these assumptions mentioned above. The ultimate objective of any regression model based on observations taken from a random sample is in fact to project the obtained results, or extend the results to the population from which this sample was taken. Now, if the regression model is good enough (meeting all assumptions) then generalizing it on the population will yield a good description of the real population model. One of the important applications of the regression model based on sample's observations is to estimate population parameters and finding confidence interval of parameters and all related issues to the inference.

If we consider the regression model as a device consisting of many components, to have a good device of high quality, the components of this device must be manufactured with good quality as a component and must work in best harmony with other components so that the device will perform in the best way. This is the analogy of the regression model as the device and the assumptions as the components. That is why these assumptions are very important to check and validate in case one or more of them is not valid. Once the model is perfect in meeting all required assumptions, the estimation of the actual population model parameters will be implemented with high accuracy.

This will take us to the relation between the population and the sample. Although, the sample is part of the population, but it seems the sample is playing against the population. Using the sample will save the time, cost and manpower that might be requires when studying the population, on the other hand, the results obtained from the sample is less accurate compared to the one obtained from the population, but then it takes longer time, higher cost and more people to study the whole population.

Therefore, results obtained from the sample will lead to accurate conclusions and generalization them to the population will yield good description of the regression model of actual population or very close to the actual real life situation.

Graphical Methods: Graphical methods play an important role in data analysis. It is of particular importance in fitting linear model of data. There is no single statistical tool that is powerful as a well-chosen graph [3]. Graphical methods can be regarded as exploratory tools. They are also an integral part of confirmatory analysis or statistical inference.

Huber [4] says, " eye-balling can give diagnostic insights no formal diagnostic will ever provide" [4].

Graphical methods can be useful in many ways [5]:

- Detect errors in the data (e.g., an outlying point may be a result of typographical error)
- Recognize patterns in the data (e.g., clusters, outliers, gaps, etc.)
- Explore relationships among variables
- Discover new phenomena
- Confirm or negate assumptions
- Assess the adequacy of a fitted model.
- Suggest remedial actions(e.g., transform the data, redesign the experiment, collect more data, etc.)
- Enhance numerical analysis in general.

There are two types of graphs in regression analysis: graphs before building the model and graphs after having the model built.

First type of graphs is useful to have a rough idea about the future model form. Based on scatter plot of the raw data, the researcher can adapt a certain model which is fitting his data. Second type of graphs is important to check the accuracy of the model by testing different assumptions.

All assumptions can be tested graphically; main dominants of these tests are the residuals and their plot. The residuals for the regression model are like the temperature for human body, if the temperature of the human body deviates from the standard norm, then we suspect something abnormal is going on with a part of the body and by investigation we can know this part and take the necessary measurements for remedy. This is exactly what is happening with the regression models, if the plot of the residuals indicating some deviation from what supposed to be healthy (according to the assumptions) for the model, then we suspect that one or more assumptions en violated or not have been met. Then the action should be taken for remedy.

The graphs after fitting a model to the data help in checking assumptions and in assessing adequacy of the fit of a given model, namely:

- Graphs for checking the linearity and normality assumptions.
- Graphs for the detection of outliers and influential observations
- Diagnostic plots for the effect of variables

Remedies from Violation: What can be done when the assumptions behind regression analysis are violated?

Linearity Violation Remedy: The model of form $Y = \beta_0 + \beta_1 x_1 + ... + \beta_q x_q + \varepsilon$ is said to be linear in the parameters.

This is because the model contains terms, $q + 1$ each of which is a parameter β_i multiplied by a value determined from the data. Some models are not following this rule as in the following regression model:

$$y = b_0 b_1^x e$$

The error terms in this model are multiplicative and not additive, which results in departing from usual linear model.. To apply the regression analysis to such a model is to transform it to a linear model:

$$\log y = \log \beta_0 + x \log \beta_1 + \log \varepsilon$$

Which can written as:

$$y' = b_o' + b_1' x + e'$$

Now after linearization the model the other assumptions can be tested. There are a numerous number of models that are violating this assumption, luckily, transformation methods can deal successfully with most of them.

No Constant Error Variance Remedy: The good point estimator must meet certain requirements, namely: unbiasedness, efficiency (having minimum variance), sufficiency (depends on maximum information from the sample) and consistency (the value of the estimator approaches the value of the parameter as the sample size increases). The first two are most important. If we do not correct the non-constant error variance problem, then the least-squares estimators will still be unbiased, but they will no longer have the minimum variance property. That is the regression coefficients will have larger standard errors than necessary.

There are many ways to get rid of the violation of constant error variance assumption, among which:

- Variance-Equalizing Transformations
- Weighted Least Squares

Which are most well-known techniques to overcome this type of violation? They can be found in most of textbooks dealing with the assumptions of regression analysis.

Multicollinearity Remedy: Multicollinearity, or near-linear dependence, is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated. If there is no linear relationship between the predictor variables, they are said to be orthogonal. In most applications of regression, the predictor variables are usually not orthogonal. Multicollinearity can be observed in one of the following:

- Large changes in the estimated coefficients when a variable is added or deleted.
- Large changes in the coefficients when a data point is altered or dropped

  Once the residual plots indicate that the model has been satisfactorily specified, multicollinearity may be present if:

  1- the algebraic signs of the estimated coefficients do not conform to the prior expectations; or

  coefficients of variables that are expected to be important have large standard errors (small t-values)

  One of the test approaches for multicollinearity presence is to use a quantity called the variance inflation factor (VIF), which can be found using;

$$VIF_j = \frac{1}{1-R_j^2}, j = 1,2,..., p$$

where R_j^2 is the square of the multiple correlation coefficient that results when the repressor X_j against all the other predictor variables. In absence of any linear relationship between the predictor variables would be zero and VIF_j would be one. The deviation of VIF_j value from 1 indicates departure from orthogonality and tendency toward collinearity. In polynomial regression models, it is suggested to use the centered data for the predictor variables. Centering the data often reduces the multicollinearity.

To control and reduce the impact of multicollinearity, one or more independent variable might be dropped from the model, this will improve the estimation precision of the remaining regression coefficients.

If none of the predictor variables can be dropped, we should consider alternative (biased) methods of estimation like Principal Component Regression (PCR) as well as many other well-known methods.

Non-Normality Remedy: Mild departure from the normality assumption are not serious.

If examination of the residuals indicates a pronounced departure from the normality assumption, there are remedies that can employed. These remedies involve transformation of the data.

Incorrect functional forms, omitted variables and violation of the constant variance assumption can cause the error terms to look non-normal.

Fortunately, remedies for these problems (for example, transformations to achieve equal variances) often correct the non-normality problem [1].

Autoregressive Error Remedy: When the error terms for a regression model are autocorrelated, we can remedy the problem by modeling the autocorrelation.

If the error terms are autocorrelated and we ignore this fact, the least squares procedure tends to produce values of the standard error of the estimate b_j) that are too small.

Consequently, values of t statistic are too large, as a result we tend to get declarations of significance when variables are really not important.

CONCLUSIONS

This paper is discussing some important steps to be taken by researchers using regression analysis. The conclusions can be summarized in the following:

- Highlight the importance of these assumptions and their impact on the results accuracy.
- Remind the researchers who are aware about these assumption and enhance and raise their awareness.
- Advice and convince researchers without background on regression analysis about the importance of these assumptions and their role in the accuracy of parameters estimation. This will contribute positively in drawing and describing the actual population regression model with high accuracy and reliability.

REFRENCES

1. Samprit Chatterjee and Ali S. Hadi, 2006. Regression Analysis by Example.4th ed., Wiley Interscience.
2. Jamal I. Daoud and Samir Salim Fadhil, 2004. Analysis of Correlation and Simple Regression Models. University of 7th of April.

3. Chambers, J.M., W.S. Cleveland, B. Kleiner and P.A. Tukey, 1983. Graphical Methods for Data Analysis. Belmont, CA: Wadsworth.
4. Joel Huber and Noreen M. Klein, 1991.Adapting cut-offs to the choice environment:The effects of Attribute Correlation and Reliability.Journal of Consumer Research, 8(3): 346-357.
5. Bowerman, O'Connell, 1990. Linear Statistical Models (an applied approach). 2nd ed. PWS-KENT Publishing Company.