On the Assessment of the Adequacy of the Fitted Regression Model Using the Confidence Interval

Odior, K. A.[1]; Emudiaga, R. E.[2]
[1]Department of Statistics, Delta-State Polytechnic, Otefe- Oghara, Delta State
[2]Kruskal Statistical Services, Delta-State Polytechnic, Otefe- Oghara, Delta State
E-mail: [1]odifullness @ gmail.com, [2]emudiagaeric @ gmail.com

Abstract— *In this study, an attempt was to critically evaluate the application of confidence interval approach in selecting most suitable predictor variables in a regression model. Data were sourced secondarily on some economic factors with probable effects on crime rate. The multiple regression model was fitted with the F-test statistic and confidence interval estimates as measuring criteria for assessing the model aptness. The result from the study revealed that the number of unemployed male citizens is evident as an increasing factor of crime rate in the society. This result was notable from the evaluation of parameter estimates using the confidence interval at α = 0.05, since only its predictor coefficient interval estimate excluded 0. The result also reported that the narrower the interval of an estimate the more efficient it becomes. In general, the study found that confidence interval estimation was more efficient in assessing a regression model than the F-global test (ANOVA) which ascertained that the model was significant at α = 0.05. Also, it is obvious that the confidence interval criterion gives a better understanding and insight of on the predictive power of the model.*

Keywords— *Confidence interval, global F-test, multiple regression, parameter estimates, crime rate.*

I. Introduction

Regression modeling is a sophisticated statistical tool used in data analysis for explaining the relationship between two or more variables, in situation where one variable is assumed to be a function others. A regression model have gained wide acceptance and application in both quantitative and qualitative researches including market analysis, economic analysis, agricultural survey, biometrical studies, etc. Regression is aimed at fitting a model that is capable of explaining the effect of some variables on a response variable. The fitted model is often used for decision making (drawing conclusion) on a particular event and at such, the fitted/estimated model is expected to possess significant explanatory strength of the situation. If the fitted model is inadequate, wrong and misleading decisions are likely to be reached and this is the relevance of critically evaluating a regression model.

The validity of a regression model is very vital to inference, and for this reason, there is need to ensure that all fitted models are cross examined for efficiency sake. Sarkar and Midi (2010) pointed that in order for analyses to be valid, the fitted models must in all certainty have to satisfy the assumptions of regression such as the observations are independent, the explanatory variables are not linear combinations of each other, errors are normally distributed. When these assumptions of regression analysis are not met, we may have problems, such as biased coefficient estimates or very large standard errors for the regression coefficients and these problems may lead to invalid statistical inferences. They suggested that a critical step in assessing the appropriateness of a regression model is to examine its fit, or how well the model describes the observed data. Without such an analysis, the inferences drawn from the model may be misleading or even totally incorrect.

This practically implies that in all sense, that after fitting a model, the first step for making any inference is to critically examine the fitted model. A well-fitted regression model results in predicted values using the predictor variable coefficients estimated. There is now a general need to test for the significance of the relationship existing between the variables. This may be done using the hypothesis testing and the confidence interval. Russell and James (2009) argued that hypothesis testing is easier to understand than the construction of verifying the truthfulness of a relationship between variables.

There are several useful criteria for measuring the goodness of fit of the multiple regression model. One such criteria is to determine the square of the multiple correlation coefficient R^2 (also called the coefficient of multiple determination). The R^2 value in the regression output indicates the percentage of the total variation of the Y values about their mean can be explained by the predictor variables used in the model. The adjusted R^2 value indicates the total variation of the Y values about their mean can be explained by the predictor variables used in the model (Shakil, 2010).

Draper and Smith (1998), proposed the examination of the estimates of variance statistics in evaluating the aptness of a regression model that the smaller it is the better, that is, the more precise will be the predictions. A useful way of looking at the decrease in is to consider it in relation to response.

Aside the use of the t-test statics, F-test and r-square, some researchers also engage the use of multicollinearity diagnostics to examine the goodness of a regression model. By multicollinearity, we mean that some predictor variables are correlated with other predictors. Various techniques have been developed to identify predictor variables that are highly collinear, and for possible solutions to the problem of multicollinearity, (Draper and Smith (1998)). For example, we can examine the variance inflation factors (VIF), which measure how much the variance of an estimated regression coefficient increases if the predictor variables are correlated. If the VIF is 5 - 10, the regression coefficients are poorly estimated. Since the variance inflation factors (VIF) for each of the estimated regression coefficient in our calculations are less than 5, there does not seem to be multicollinearity in our model.

Gibbs et al. (2006) developed a regression model on establishing the relationship between home environments and reading achievement in Zimbabwe. The study included seven predicting variabbles of reading achievement. The aptness of the model parameter estimates was examined using the standard error measurement which enable their study select only the best performing predictors to be included in the model estimated.

Moustris et al. (2012) employed the multiple linear regression models and artificial neural network models to forecast the maximum daily surface ozone concentration for the period of 24 hours, within the Greater Athens Area, Greece. The study was based on maximum daily values of surface ozone concentrations (μg/m^3) that were recorded for a five-year period (2001–2005). The model validity was highly based on the use of the variance inflation factor as a measure of multicollinearity and the root mean square error.

The aim of model selection is to minimize the number of predictors which account for the maximum variance in the criterion. In other words, the most efficient model maximizes the value of the coefficient of determination (R-square). This coefficient estimates the amount of variance in the criterion score accounted for by a linear combination of the predictor variables. The higher the value is for R-square, the less error or unexplained variance and, therefore, the better prediction. R-square is dependent on the multiple correlation coefficient (R), which describes the relationship between the observed and predicted criterion scores. If there is no difference between the predicted and observed scores, R equals 1.00. This represents a perfect prediction with no error and no unexplained variance (R-square= 1.00). When R equals 0.00, there is no relationship between the predictor(s) and the criterion and no variance in scores has been explained (R-square= 0.00). The chosen variables cannot predict the criterion. The goal of model selection is, as stated previously, to develop a model that results in the highest estimated value for R-square.

According to Jackson, (1989), lower values of the standard error estimate (SEE) indicate greater accuracy in regression model prediction. Comparison of the SEE for different models using the same sample allows for determination of the most accurate model to use for prediction. SEE % is calculated by dividing the SEE by the mean of the criterion (SEE/mean criterion) and can be used to compare different models derived from different samples.

In summary, the efficiency of every estimated regression model is disputable. Despite this, researchers are still recommending that future research efforts should attempt to examine the aptness of every estimated model. This study focused on using the method of the confidence interval estimation as a measure of assessing regression model predictive power which has received few attention by researchers who uses regression model for prediction.

II. Research Methodology

The data collected for this study were based on the factors that necessitate crime in some randomly selected countries. Hence, the data were secondary in nature. The data were used to fit a regression model with associated parameter estimate confidence interval fitting.

Multiple Regression Analysis

Multiple regression analysis is used for testing hypothesis about the relationship between a dependent variable Y and two or more independent variables X and for prediction. For this study, the regression model used is stated below as;

$$Y_i = \beta_0 + \beta_i X_i + \cdots + \beta_k X_k + e_{ij} \tag{1}$$

Ordinary least squares (OLS) parameter estimates for two predictors model can be obtained by minimizing the sum of the squared errors.

But $e_{ij} = Y_i - \hat{Y}_i$

$$\sum e_{ij} = \sum Y_i - \hat{Y}_i \tag{2}$$

$$\sum e_{ij}{}^2 = \sum (Y_i - \hat{Y}_i)^2 \tag{3}$$

Since $\hat{Y}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$

$$\sum (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})^2 \tag{4}$$

This gives the following three normal equations

$$\sum Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum X_{1i} + \hat{\beta}_2 \sum X_{2i} \tag{5}$$

$$\sum X_{1i} Y_i = \hat{\beta}_0 \sum X_{1i} + \hat{\beta}_1 \sum X_{1i}{}^2 + \hat{\beta}_2 \sum X_{1i} X_{2i} \tag{6}$$

$$\sum X_{2i} Y_i = \hat{\beta}_0 \sum X_{2i} + \hat{\beta}_1 \sum X_{1i} X_{2i} + \hat{\beta}_2 \sum X_{2i}{}^2 \tag{7}$$

Which when expressed in deviation form can be solved simultaneously for $\hat{\beta}_1$ and $\hat{\beta}_2$ giving

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 \tag{8}$$

$$\hat{\beta}_1 = \frac{(\sum X_1 Y)(\sum X_1{}^2) - (\sum X_2 Y)(\sum X_1 X_2)}{(\sum X_1{}^2)(\sum X_2{}^2) - (\sum X_1 X_2)^2} \tag{9}$$

$$\hat{\beta}_2 = \frac{(\sum X_2 Y)(\sum X_1{}^2) - (\sum X_1 Y)(\sum X_1 X_2)}{(\sum X_1{}^2)(\sum X_2{}^2) - (\sum X_1 X_2)^2} \tag{10}$$

In order to test for the statistical significance of the parameters estimates of the multiple regressions, the variance of the estimates is required:

$$var\hat{\beta}_1 = \sigma_v^2 \frac{\sum X_2{}^2}{(\sum X_1{}^2)(\sum X_2{}^2) - (\sum X_1 X_2)^2} \tag{11}$$

$$var\hat{\beta}_2 = \sigma_v^2 \frac{\sum X_1{}^2}{(\sum X_1{}^2)(\sum X_2{}^2) - (\sum X_1 X_2)^2} \tag{12}$$

$\hat{\beta}_0$is usually not of primary concern. Since σ_v^2 is unknown, the residual variance s^2 is used as an unbiased estimate is required:

$$S^2 = \sigma_v^2 = \frac{\sum e_i{}^2}{n-k} \tag{13}$$

Where k is the number of parameter estimates. Unbiased estimates of the variance of $\hat{\beta}_0$ and $\hat{\beta}_1$ are then given by

$$S^2{}_{\hat{\beta}_1} = \frac{\sum e_i{}^2}{n-k} \frac{\sum X_2{}^2}{(\sum X_1{}^2)(\sum X_2{}^2) - (\sum X_1 X_2)^2} \tag{14}$$

$$S^2{}_{\hat{\beta}_2} = \frac{\sum e_i{}^2}{n-k} \frac{\sum X_1{}^2}{(\sum X_1{}^2)(\sum X_2{}^2) - (\sum X_1 X_2)^2} \tag{15}$$

So that $S_{\hat{\beta}_1}$and $S_{\hat{\beta}_2}$are the standard errors of the estimates. Tests of hypothesis can be conducted using the estimators above.

Confidence Interval Estimation

Given a regression model as $Y_i = \beta_0 + \beta_i X_i + \cdots + \beta_k X_k + e_{ij}$, i= 1, 2,…,k, the following assumptions are necessary;

1. e_{ij} is a random variable with mean zero and variance σ^2 (unknown); that is, $E(e_{ij})= 0$, $V(e_{ij}) = \sigma^2$
2. $e_i$ande_j are uncorrelated, $i\neq j$ so that cov(e_i,e_j) = 0. Thus

$$E(Y_i) = \beta_0+\beta_i X_i,\ V(Y_i) = \sigma^2, \tag{16}$$

And Y_i and Y_j, $i \neq j$, are uncorrelated.

3. $e_{ij}$is a normally distributed random variable, with mean zero and variance σ^2 by assumption 1; that is, $e_{ij}{\sim}N(0,\sigma^2)$. Under this assumption, e_i and e_j are not only uncorrelated but are also independent.

The confidence interval for the regression slopes is generally given as;

$$C.I.(\beta_i) = \beta_i \pm t_{(n-2)}\sqrt{\frac{MS_{error}}{SS_{x_i}}} \tag{17}$$

Where

$$SS_{x_i} = \sum(X_i - \bar{X}_i)^2 = \sum X_i^{\,2} - \frac{(\sum X_i)^2}{n} \tag{18}$$

III. Result and Discussion

TABLE 1: Parameter and confidence interval estimates for model including all four predictors

Model		Unstandardized Coefficients		Standardized Coefficients	95.0% Confidence Interval for B	
		B	Std. Error	Beta	Lower Bound	Upper Bound
1	(Constant)	6.553	19.641		-38.172	15.067
	Population of male (14-24)	.642	.523	.239	-.413	1.698
	Number of educated male (14-24) ['000]	-.332	.635	-.110	-1.614	.949
	Average monthly household income (USD)	-.181	.293	-.089	-.772	.410
	Number of unemployed male (14-24) ['000]	.220	.083	.632	.052	.388

TABLE 2: ANOVA for model including all four predictors

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	10543.207	4	2635.802	2.609	.051[b]
	Residual	41424.756	41	1010.360		
	Total	51967.964	45			

a. Dependent Variable: Crime rate

b. Predictors: (Constant), Average monthly household income (USD), Number of unemployed male (14-24) ['000], Population of male (14-24), Number of educated male (14-24) ['000]

In the estimated model parameters presented in table 1, population of male, number of educated male and average monthly household income were found insignificant in contributing to the predictive power of the model since their confidence interval estimates includes zero. This result leaves the predictor, number of unemployed male, to be the only significant predictor variable capable of predicting the rate of crime in the model. This is as result of the contribution stated in the confidence interval estimate having excluded zero. Table 2 attempted the F-global test (ANOVA) approach in assessing the adequacy of the model, which reported that the fitted model is generally inadequate in predicting the occurrence of crime with p-value is 0.051. However, with the report obtained from table 1, the estimated model can be reduced and further assessed for adequacy.

TABLE 3: Parameter and confidence interval estimates for model including only three predictors

Model		Unstandardized Coefficients		Standardized Coefficients	95.0% Confidence Interval for B	
		B	Std. Error	Beta	Lower Bound	Upper Bound
1	(Constant)	-92.651	107.828		-310.257	124.955
	Number of unemployed male (14-24) ['000]	.194	.066	.556	.061	.326
	Population of male (14-24)	.665	.516	.248	-.377	1.707
	Average monthly household income (USD)	-.163	.288	-.081	-.744	.419

TABLE 4: ANOVA for model including three predictors

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10235.801	3	3411.934	3.434	.025[b]
	Residual	41732.163	42	993.623		
	Total	51967.964	45			

a. Dependent Variable: Crime rate

b. Predictors: (Constant), Number of unemployed male (14-24) ['000], Average monthly household income (USD), Population of male (14-24)

TABLE 5: Parameter and confidence interval estimates for model including only two predictors

Model		Unstandardized Coefficients		Standardized Coefficients	95.0% Confidence Interval for B	
		B	Std. Error	Beta	Lower Bound	Upper Bound
1	(Constant)	-120.500	95.110		-312.307	71.308
	Number of unemployed male (14-24) ['000]	.199	.064	.571	.069	.329
	Population of male (14-24)	.736	.497	.274	-.266	1.738

TABLE 6: ANOVA for model including three predictors

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9919.366	2	4959.683	5.072	.011[b]
	Residual	42048.598	43	977.874		
	Total	51967.964	45			

a. Dependent Variable: Crime rate

b. Predictors: (Constant), Population of male (14-24), Average monthly household income (USD)

From table 3, only the number of unemployed male predictor variable stood out as the only significant predictor variable that can foretell the rate of crime rate to be expected with a 95% confidence interval of 0.06 – 0.33. Still the confidence interval for average monthly income and total population of male had confidence interval estimates of -0.74 – 0.49 and -0.38 – 1.71 respectively.

In table 4, the result of the ANOVA overall model fit aptness indicated that the fitted model is generally adequate. Though, the confidence interval approach is clearly spotting that only one of all the predictors is efficient. This there indicates a sign of weakness in the F-global test method of assessing a model adequacy.

With a reduced model presented in table 5, it is proven that the confidence interval of the parameters still did not change its level of significance; that is, only the number of unemployed male remains the only predictor with significance predicting power in the model.

Notwithstanding, the overall model assessment using the ANOVA approach also suggest that the model can be used for prediction, whereas, just one predictor is significant.

IV. Conclusion

The findings from the study shows that the confidence interval approach of assessing model adequacy of the fitted regression model gives a better interpretation and information about the parameters when compared with the global F-test approach (ANOVA). This is because the confidence interval focuses on the possible range of the parameter estimates that can explain the changes that occurred in the response variable rather than the outright rejection or acceptance of the model as applicable in the F-test approach.

References

[1] Draper, N. R., and Harry S. (1998). Applied Regression Analysis (3rd edition). New York: John Wiley & Sons, INC.

[2] Gibbs, Y. K., Janine, C. and brown I. L. (2006).Using regression analysis to establish the relationship between home environment and reading achievement: A case of Zimbabwe. *International Education Journal*, 7(5)

[3] Jackson, A. S. (1989). Application of Regression Analysis to Exercise Science. In: SafritMJ, Wood TM, eds. Measurement Concepts in Physical Education and Exercise Science. Champaign, IL: Human Kinetics Books.

[4] Moustris,K. P., Nastos, P. T., Larissi, I. K. and Paliatsos, A. G. (2012).Application of Multiple Linear Regression Models and Artificial Neural Networks on the Surface Ozone Forecast in the Greater Athens Area, Greece. Advances in Meteorology, 12(1)

[5] Russell and James (2009).Recognizing the Vitality and Energy Inherent in Transience: A Tribute to Carl Milofsky. *International Journal of Research on Emotion*1 (1)

[6] Sarkar, S. K. and Midi, H. (2010).Importance of Assessing the Model Adequacy of Binary Logistic Regression. *Journals of Applied Sciences*, 10 (6): 479-486

[7] Shakil, M. (2010).A Multiple Linear Regression Model to Predict the Student’s Final Grade in a Mathematics Class. Available at https://www.shsu.edu/~wxb001/documents/Amultipleregressionmodelpaper.pdf