



International Journal of Emerging Technologies in Computational and Applied Sciences(IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

STATISTICAL CRITERIA FOR SELECTION OF FUNCTIONAL FORMS IN REGRESSION ANALYSIS

K. Naga Vihari, T. Suneetha*, and R. Abbaiah

Department of Statistics, Sri Venkateswara University, Tirupati – 517502, Andhra Pradesh, INDIA.

* Asst. Prof. (contract) Department of Statistics, Sri Padmavati Mahila Visvavidyalayam (SPMVV),
Padmavathi Nagar, Tirupati-517502, Andhra Pradesh, INDIA.

Abstract: The purpose of this study is to provide statistical criteria for selection of functional forms in Regression analysis. A model can be misspecified in a number of ways. Two major sources are incorrect functional form and invalid assumptions on the distribution of the disturbance term. In respect of the disturbance term, classical regression analysis is based on the assumptions affects both estimated and inference results. The research also explores specification error in Regression analysis and some test for misspecification of the model. The research concludes that new criteria for selection of functional forms in regression analysis.

Key Words: Regression analysis, specification errors, misspecification, predictors.

I. Introduction

The basic question in econometrics is how to specify a good model for the purpose at hand. The specification of a good model is an art. Once a model is specified, however, there are statistical criteria to judge whether the model is bad. Since many models can explain the same set of data but equally well, a given set of data can be used to screen out bad models but not to generate good models, whatever statistical techniques are used. Most of the techniques presented in the proposed research work are concerned with the choice between alternative model specifications which of two non-nested regression models should be selected; whether a larger or smaller regression models should be selected; and whether a larger or a smaller simultaneous-equation model should be selected.

One of the most important components of econometric model-building is test for specification errors. A model can be misspecified in a number of ways. Two major sources are incorrect functional form and invalid assumptions on the distribution of the disturbance term. Regarding the function form linearity is often assumed for simplicity when a non-linear function would be more appropriate, and this may be accompanied by exclusion (or inclusion) of some relevant (or irrelevant) variables. In respect of the disturbance term, “classical regression analysis is based on the assumption of disturbance normality, homoscedasticity and serial independence violation of these assumptions affects both estimated and inference results.

II. Methodology

2. A BRIEF ABOUT SPECIFICATION ERROR IN REGRESSION ANALYSIS

2.1 SPECIFICATION ERROR

It is often difficult to determine which of many possible variables in $\tilde{x} = (x_1, \dots, x_k)$ to use in estimating the

regression function $g(\tilde{x}) \equiv E(Y / \tilde{x})$ or in predicting y , particularly in cases for which n is relatively small the statistician is form between the wish to keep the model simple and the wish for a good approximation. If a poor choice of a subset $x_{i_1}, \dots, x_{i_5}$ of possible measurements is made what will the penalties be?

For example in trying to determine the regression function $g(\tilde{x})$, should we use a fifth degree polynomial or will a quadratic function suffice? Obviously we can fit the data more closely with a fifty degree polynomial, but may

pay a price in increased complication, poor extrapolation, and as we shall see, a loss of precision. On the other hand. If true regression is cubic and we fit a quadratic function some inaccuracy would seem to result.

To make the discussion precise suppose $y = \theta + \varepsilon$ for $\varepsilon \sim N(0, \sigma^2 I_n)$ and our postulated model is $\theta \in V$ a known subspace of R_n of dimension K . As will be seen by the following analysis. If $\theta \in V_1$ errors may result to see this, let $\theta = \theta_v + \theta_\perp$ for $\theta_v = p(\theta/v)$. Let $\varepsilon_v = p(\varepsilon/v)$ and $\varepsilon_\perp = \varepsilon - \varepsilon_v$. Then the least squares estimator of θ is $\hat{y} = \theta_v + \varepsilon_v$ and the error in the estimation of θ is $d = \hat{y} - \theta = -\theta_\perp + \varepsilon_v$. Thus $\hat{y}$ has bias $-\theta_\perp$. We can assess the expected sizes of the errors made in estimating θ by computing $E(\hat{y} - \theta)(\hat{y} - \theta)^1 = \theta_\perp \theta_\perp^1 + E(\varepsilon_v \varepsilon_v^1) = \theta_\perp \theta_\perp^1 + E(p_v \varepsilon \varepsilon^1 p_v) = \theta_\perp \theta_\perp^1 + \sigma^2 p$ here we have taken advantage of the orthogonality of $\theta_\perp$ and ε_v . To gauge the size of this we can compute the sum of the expected squared errors.

$$E \|d\|^2 = E [\|\theta_\perp\|^2 + \|\varepsilon_v\|^2] = \|\theta_\perp\|^2 + K\sigma^2 \text{ since } v \text{ has dimension } k.$$

We might also study the random variable $Q = \|d\|^2 = \|\theta_\perp\|^2 + \|\varepsilon_v\|^2$ in order to understand the size of their errors. Q is a constant plus σ^2 multiplied by a central χ^2 random variable with expectation given above.

Error sum of squares is $\|y - \hat{y}\|^2 = \|\theta_\perp + \varepsilon_\perp\|^2$ so that $\|y - \hat{y}\|^2 / \sigma^2 \sim \chi_{n-k}^2(\delta)$ for $\delta = \|\theta_\perp\|^2 / \sigma^2$.

Thus $E(s^2) = \sigma^2 + \|\theta_\perp\|^2 / (n-k)$.

In searching for a good model we might try to choose a subspace V so that $H_v = E \|d\|^2 / \sigma^2 = \|\theta_\perp\|^2 / \sigma^2 + k$ is small. Of course, H_v depends on unknown parameters. It can be estimated if we can find an estimator of pure error variance σ^2 we might for example use a particularly large subspace V_L in which we are quite sure θ lies, and use error sum of squares for this subspace to estimate σ^2 , or we might use past data from another experiment with repeated observations on y for the same x to estimate σ^2 . Let σ^2 be this estimate of pure error variance. Let s^2 be the estimator for the subspace v .

Then $E(s^2 - \sigma^2)^2 (n-k) = \|\theta_\perp\|^2$, so that $\frac{C_v = (s^2 - \sigma^2)^2}{\sigma^4} (n-k) + k$ can be used as an estimator of

$H_{11} \cdot C_v$ is called mallows C_p for the call that $\dim(v) = p$. Since $H_v = \dim(v)$ for $\theta \in v$, we should hope to find a subspace V such that C_v is close to or smaller than $\dim(v)$.

2.2 EFFECTS OF SPECIFICATION ERRORS ON $\eta = c^1 \beta$

In order to assess the effect of specification errors on β or more generally, $\eta = c^1 \beta$ we need to specify θ and v in terms of specific x -vectors. We consider two cases: over specification and under specification.

2.3 SOME TEST FOR MISSPECIFICATION OF THE MODEL

2.3.1 RAMSEY'S (1969) RESET

(REGRESSION SPECIFICATION ERROR TESTS)

Ramsey suggests testing the specification of the linear regression model $Y_t = X_t^1 \beta + u_t$ by augmenting it with a set of regressors Z_t so that the augmented models is

$$Y_t = X_t^1 \beta + Z_t^1 \gamma + u_t$$

If the Z_t 's are available then the specification test would reduce to the F-test for $H_0 : \gamma = 0$. The crucial issue is the choice of Z_t variables. This depends upon the true functional form under the alternative, which is usually unknown. However, this can be often well approximated by higher powers of the initial regressors, as in the case where the true form is quadratic or cubic. Alternatively, one might approximate it with higher moments of $\hat{Y}_t = X_t^1 \hat{\beta}_{OLS}$.

The popular Ramsey RESET test is carried out as follows:

- 1) Regress Y_t on X_t and get $\hat{Y}_t$
- 2) Regress Y_t on $X_t, \hat{Y}_t^2, \hat{Y}_t^3$ and $\hat{Y}_t^4$ and test that the coefficients of all the powers of $\hat{Y}_t$ are zero. This is an $F_{3, T-k-3}$ under the null.
- 3) Note that $\hat{Y}_t$ is not included in the MSE from the regression computed with out the i^{th} observation.

$$s_{(i)}^2 = \frac{(n-k)s^2 - e_i^2 / (1 - h_{ii})}{(n-k-1)} = s^2 \left(\frac{n-k - \tilde{e}_i^2}{n-k-1} \right)$$

Because it would be perfectly multi-collinear with X_t^3 . Different choices of Z_t 's may result in more powerful tests when H_0 is not true. Thursby and Schmidt (1977) carried out an extensive Monte Carlo and concluded that the test based on $z_t = [X_t^2, X_t^3, X_t^4]$ seems to be generally the best choice.

2.3.2 THE J TEST

Davidson and MacKinnon (1981) have proposed the alternative in a general model that uses a mixing parameter λ in the combined statistical model.

$$Y = (1 - \lambda) X \beta_1 + \lambda Z \beta_2 + e$$

Under the hypothesis H_2 , Davidson and MacKinnon replace the unknown $Z \beta_2$ with its estimate

$$Z \hat{\beta}_2 = (I - M_2) Y. \text{ Therefore, may be rewritten as}$$

$$Y = X \beta_1 (1 - \lambda) + \lambda (Z \hat{\beta}_2) + e$$

$$= X \beta_1 (1 - \lambda) + \lambda \hat{Y}_2 + e$$

Set up in this form, if hypothesis H_1 is true, then $\lambda = 0$. Since $\hat{Y}_2$ is asymptotically independent of e , they suggest testing whether $\lambda = 0$. By using a likelihood ratio test or the conventional t test, which they call a J test. Therefore, if H_1 is true, as Pesaran (1982) notes, this results in the test statistic.

$$J_1 = \hat{\beta}_2^1 Z^1 M_1 Y / \hat{\sigma} \left(\hat{\beta}_2^1 Z^1 M_1 Z \hat{\beta}_2 \right)^{1/2}$$

where $\hat{\sigma}$ is an estimate of the standard error of regression. This test statistic is asymptotically distributed as a standard normal random variable and is asymptotically equivalent to the JA test proposed by Fisher and McAleer (1981). By using different estimates of β_2 , a number of alternative tests may be generated.

III. NEW CRITERIA FOR SELECTION OF FUNCTIONAL FORMS IN REGRESSION ANALYSIS

The general problem in Model selection criteria is to determine the most important or most valid subset of the predictors and the corresponding best fitting regression model for describing the relationship between a response variable Y and a set of K predictor variables $X_1, X_2, \dots, X_k$. The best fitting depends on overall goal for modeling. The goal is to specify a model that provides the best prediction of Y given $X_1, X_2, \dots, X_k$, for some new observation or for a batch of new observations. Using this goal, one may say that the best model is reliable if it predicts well in a new sample.

To select the best regression model, the following five steps may be carried out:

- (i) Specify the maximum model, which is to be the one having the most predictor variables, to be considered.
- (ii) Specify a criterion for selecting a model.
- (iii) Specify a strategy for selecting the variables for models building.
- (iv) Conduct the specified analysis.
- (v) Evaluate the reliability of the chosen model.

A criterion for selecting model is an index that can be computed for each candidate model and may be used to compare regression models. Given one particular selection criterion, candidate models can be ordered from best to worst. It helps automate the process of choosing the best model.

Computing all possible regression models rapidly becomes impractical as the number of variables increases.

In this situation, one may resort to step by step methods to select subsets sequentially and avoid the computational cost of fitting very large numbers of models.

Consider a linear regression model,

$$Y_{n \times 1} = X_{n \times k} \beta_{k \times 1} + \epsilon_{n \times 1}$$

One may be interested to identify the significant explanatory variables having nonzero regression coefficients.

Suppose that one may divide the k variables into two steps: the first set consists of ' r ' regressors and the second contains the remaining $(k-r)$ variables, consists of variables whose regression coefficients one suspect are zero.

Let $e_r^{*1} e_r^*$ be the Internally studentized residual sum of squares based on linear regression model consists of first

set of r regressors; and let $e_k^{*1} e_k^*$ be the Internally studentized residual sum of squares based on full linear regression model.

The F-test statistic can be used to test if the second set contains no significant variables. A modified F-test statistic is given by

$$F^* = \left[\frac{e_r^{*1} e_r^* - e_k^{*1} e_k^*}{e_k^{*1} e_k^*} \right] \left[\frac{n-k}{k-r} \right] \sim F_{[(k-r), (n-k)]}$$

This F^* can be used as a test for discriminating between two models, having k and r variables respectively. In a particular case, if $k=r+1$, then the F^* is given by

$$F^{**} = \left[\frac{e_r^{*1} e_r^* - e_{r+1}^{*1} e_{r+1}^*}{e_{r+1}^{*1} e_{r+1}^*} \right] [n-r] \sim F_{[1, (n-r)]}$$

Which tests if addition of a specified extra regressor is necessary.

The statistic F^* may be compared to critical value of F distribution with $[(k-r), (n-k)]$ degrees of freedom.

If F^* is not significant, one can use the smaller (r -variables) model and achieve roughly the same predictive ability as that yielded by the full regression model. Thus a reasonable rule for selecting regressors is to retain r -variables if F^* is not significant and if r is as small as possible. A special case of F^{**} occurs when $k=r+1$, in which case F^{**} can be used as a test of significance of k^{th} regression coefficient in the full model.

Under modified Forward Selection procedure, one may start with a model containing only the constant term, compute F^{**} with $r=1$ for all the available regressors in turn and pick the variable for which F^{**} is the maximum value. Then repeat the procedure for $r=2,3,\dots$, selecting at each stage, the variable not currently included that gives the maximum value of F^* . One may stop if the maximum F^{**} at any stage not exceed some threshold value F_{IN} (F-to Enter).

Under modified Backward Elimination procedure, as an alternative to Forward Selection, one starts with the full regression model using all k variables and compute F^* with $r=k$ for each of the k variables. Eliminate the variable having the Smallest F^* statistic from the model, provided that F^* is less than some threshold value F_{OUT} (F-to Remove). This procedure is continued until all variables are eliminated or the smallest F^* fails to be less than F_{OUT} . Backward Elimination procedure typically takes much more computation than Forward Selection when a small number of variables explain most of the variation. The thresholds F_{IN} and F_{OUT} may change as the algorithms proceed. A popular choice is to keep the formal significance level fixed, and set the thresholds to be the critical values for the corresponding F^* -test.

The main drawback of these two procedures is that in the case of Backward Elimination (BE), a variable once eliminated can never be reintroduced into the regression model, and in the case of Forward Selection (FS), once a variable included can never be removed. Sometimes, these methods can give very different results on the same data. The Stepwise Regression technique combines both the Forward Selection and Backward Elimination procedures, which is just Forward Selection followed by a Backward Elimination step at each stage.

Under modified Stepwise Regression technique, one starts with a model consisting of constant term alone, then performs an FS step, adding a single variable. It is followed by a BE step, removing a variable if the corresponding F^{**} is less than F_{OUT} . This combination of an FS step followed by a BE step is repeated until no further regressors are introduced at the FS stage.

Suppose that at the beginning of an FS step, there are r variables in the current linear regression model, which has

Internally studentized residual sum of squares $e_r^{*1} e_r^*$. Then either no regressor can be included, and the algorithm terminates or an additional regressor is included, resulting in a new studentized residual sum of squares $e_{r+1}^{*1} e_{r+1}^*$ (say), which must satisfy,

$$F^{**} = \left[\frac{e_r^{*1} e_r^* - e_{r+1}^{*1} e_{r+1}^*}{e_{r+1}^{*1} e_{r+1}^*} \right] [n - r] > F_{IN}$$

Or, equivalently

$$\left[e_r^{*1} e_r^* \right] > e_{r+1}^{*1} e_{r+1}^* \left(1 + \frac{F_{IN}}{n - r} \right) > \left[e_{r+1}^{*1} e_{r+1}^* \right]$$

Now, perform the BE step. Either no regressor is deleted or one may get a new regression model with r variables

and a new Internally studentized residual sum of squares equals to $\left[e_r^{*1} e_r^* \right]_{New}$, say, where

$$\left[e_r^{*1} e_r^* \right]_{New} < e_{r+1}^{*1} e_{r+1}^* \left(1 + \frac{F_{OUT}}{n - r} \right)$$

Thus,

$$\left[e_r^{*1} e_r^* \right]_{New} < e_{r+1}^{*1} e_{r+1}^* \left(1 + \frac{F_{OUT}}{n - r} \right) \leq e_{r+1}^{*1} e_{r+1}^* \left(1 + \frac{F_{IN}}{n - r} \right) < \left[e_r^{*1} e_r^* \right],$$

So that at the end of each FS/BE cycle, either the algorithm has terminated or a new regression model is chosen whose internally studentized residual sum of squares is strictly smaller. In Stepwise Regression technique, a good stopping rule will result in good performance of the regression model.

IV. CONCLUSIONS

Model selection criteria have become exceedingly popular in the specification of regression models. Selecting the best statistical model is an important problem in statistics as well as in any other field that uses regression analysis. The problem of criteria for selection of functional forms in regression has long been of interest of statisticians. In the applied regression analysis, one is faced with a large number of explanatory variables which are potentially important for the specification of the model.

There is considerable literature on selection of functional forms in regression analysis. Many Statisticians and Econometricians have proposed several methods for selecting the best regression model. In deciding on a method, it is necessary to evaluate the criterion of goodness of prediction on which it is based and to some extent, the computational effort involved,

In the present research work, an attempt has been made to develop some new criteria for selection of functional forms in regression analysis.

Firstly, a modified stepwise regression technique for the selection of best linear regression model has been developed by giving modified adjusted coefficients of multiple determinations based on internally studentized

residual sum of squares. Under this technique both forward selection and backward elimination procedures have been used simultaneously for selecting variables for the best linear statistical model.

Secondly, a criterion for model selection has been proposed by using prediction error and model error. The concepts of prediction error and model error are essential to reflect the underlying variability in the data and measuring how well the model predicts the new responses. Here, Total bias and Total variance have been defined for choosing the best subset of regressors.

Thirdly, a new criterion for selecting functional forms in regression analysis has been proposed by giving a test statistic based on coefficients of multiple determinations for reduced model and full regression model.

Finally, a criterion has been suggested test for the specification and misspecification of the models.

BIBLIOGRAPHY

- [1] Basic Econometrics by Damodar N.Gujarati 4th edition
- [2] THE THEORY AND PRACTICE OF ECONOMETRICS by George G. Judge, William E. Griffiths, R. Carter
- [3] Cheng Hsiao (1975), "Some estimation methods for a Random coefficient model", *Econometrics*, V (43), No.sp.(305-323)
- [4] RAO, C.R. "The Theory of least squares when the parameters are stochastic and its Application to the Analysis of Growth curves," *BIOMETRICA*, 52 (1965) Page :447-58.
- [5] THOMAS, F. COOLY and Edward, c. Prescott (1976 JAN). "Estimation in the presence of stochastic parameter variation. " *Econometrics* (44), No. 1 p, (167-184).
- [6] P.A.V.B. Swamy and J.S. Mehta (1975) "Bayesians and Non-Bayesians and random coefficient regression models", *JASA*, V (70) P. (593-602)
- [7] PVVB Swamy & T.S. Mehate (1977,Dec)" Estimation of linear models with varying coefficients" *JASA* (72), p (890-898)
- [8] GARBADE, K (1977 MAR) "Stability of Regression coefficient", *JASA* (1972) P(54-63)
- [9] R.A. BATCHELOR (1977) "A Variable – parameter model of exporting behavior",
- [10] ASHOK PARIKH and Baldev Raj (1979), "Variable coefficients approach to wage-price behavior in united kingdom", *Applied Economics*, V(11)p (389-403)
- [11] BALDEV RAJ, Virendev Kumar, Srivastava and sushma, Upadhyaya (1980) "The efficiency of estimating a random coefficient model, " *Journal of Economics* V (12), p (285-299)
- [12] a. Deaton (1985), "panel data from series of Repeated cross Sections" *J. Econometrics*, V (30) p (109-126)
- [13] ANDREWS, D.F.(1974) "A Robust Method for multiple linear Regression", *Techno-metrics* 16, (523-31)
- [14] DRAPER, N and smith . H (1998) "Applied Regression Analysis " Third Edition, New York, John Wiley & sons
- [15] Lee, L.F. (1983), "On Maximum likelihood estimation of stochastic frontier production Models" *Journal of Econometrics*, 23 page(269-274)
- [16] Zellner, A. (1976), "Bayesian and Non-Bayesian Analysis of the Regression model with Multivariate student 't' Error terms" *JASA*, 71 page: (400-405)
- [17] Gregory, C. Chow (1983), "ECONOMETRICS"
- [18] Judge G.G. W.F. Griffiths, R.C. Hill, H. LUTKEPOHL and T.C. Lee (1985), *The Theory and practice at Econometrics*.
- [19] Quandt, Richhard E, "The Estimation of the parameters of a linear regression system obeying two separate regimes". *JASA*, March 1958. P. (873-880)
- [20] Chow G.C. (1983) : "Econometrics" Mc Graw – Hill Book Company
- [21] Cook R.D. and S. Weisberg (1982) : "Residuals and Influnce in Regression". Chapman and hall, London
- [22] David A. Belsley Edwin Kuh Roy E. Welsch (1980) : "Regression Diagnostics" Johnwiley & sons.
- [23] Donald A. Pierce and Doniel W. Schafer (1986) : Residuals in Generalised Linear models, *JASA* 81, 977-986.
- [24] DRAPER N.R. and H. Smith (1981): "Applied Regression Analysis" John Wiley & sons.