LINEAR REGRESSION MODEL WITH NEW MIS-SPECIFICATION TESTS IN BIO-STATISTICS

J. Ravi Sankar, G. Mokesh Rayalu*, A. Felix
Assistant Professor, Department of Mathematics, School of Advanced Sciences
VIT University, Tamil Nadu, India – 632014.
Email: mokesh.g@vit.ac.in

Received on 09-08-2016
Accepted on 28-08-2016

Abstract:

In diagnostic testing, model specification tests play an important role. The specification of a good model is an art. Over specification yields unbiased estimates of regression coefficients, but larger variance; under specification yields biased estimates of regression coefficients and understates the variance of these estimates. In case of the left out variables or because of inclusion of variable not specified by the truth; the case of irrelevant variables. Mis-specification is usually interpreted as a case of omitted variables, and many researchers, concerned only with the bias resulting from it, the specification bias. Researchers seldom pay attention to the other aspects of mis-specification of the model. In view of the importance of these aspects of misspecification in empirical research, some major results of misspecification error tests are considered in this Article. In this research article, three types of new tests for the mis-specification of the linear regression model have been developed by using various types of residuals.

I. Introduction:

In diagnostic testing, model specification tests plays an important role. Diagnostic testing is one of the most important areas of research in Biostatistics. To apply the specification tests at various stages. The last four decades have seen a substantial literature on misspecification tests for statistical models. The specification of a good model is an art. Over specification yields unbiased estimates of regression efficient, but larger variance; under specification yields biased estimates of regression coefficients and understates the variance of these estimates. In case of the left out variable or because of inclusion of variable not specified by the truth; the case of irrelevant variable or because of inclusion of variable not specified by the truth, the case of irrelevant variable misspecification and statisticians concerned only with
the bias resulting from it, the specification bias. In view of the importance of these aspects of misspecification in empirical research, some major results of misspecification error tests are considered in the present study.

II. A New Test for Misspecification of the Linear Regression Model:

In statistical model building one often faces the problem of misspecification of the Bio-statistical model. There are mainly four sources of the problem of misspecification namely, omission of a variable specified by truth; the case of the left out variable; inclusion of a variable not specified by the truth and the case of the irrelevant variable.

Potluri Rao (1970) stated the following consequences of omission or inclusion of a variable in the model:

In the classical linear regression model, omission of a variable specified by the truth,

(i) introduces bias in all the least squares estimates;
(ii) decreases the variance of all the least squares estimates; and
(iii) Decrease the mean square error of all the least squares estimates.

Further, inclusion of an irrelevant variable

(i) does not introduce bias in the least squares estimates;
(ii) increases the variance of all the least squares estimates;
(iii) Increases mean squared error of all the least squares estimates.

Consider a linear regression model as

\[ Y_{n1} = X_{n1} \beta_{k1} + \epsilon_{n1} \]  

Such that $\epsilon \sim \text{Multivariate } N \left(0, \sigma^2 I_n\right)$

State the null and alternative hypotheses as

H0: $E[\epsilon | X] = 0$ and H1: $E[\epsilon | X] = Z \neq 0$

Where Z is unknown (nx1) vector, which is unobservable and not orthogonal to X

To test the null hypothesis against a nonzero error mean conditional on X in the regression model, one may consider the augmented regression model as

\[ Y = X \beta + W\Gamma + \epsilon \]  

Where W is a (nxq) matrix of rank ‘q’ of test variables such as powers or functions of the X variables and rank of (X,W)=k+q.
The OLS estimator of $\beta$ from (2.1) is given by

$$\hat{\beta} = (XX)^{-1} XY$$

(2.3)

The OLS estimator of $\beta$ from (2.2) is composed of the first $k$ elements of

$$\left[ (X, W)' (X, W) \right]^{-1} [(X, W)]' Y,$$

which is denoted by $\tilde{\beta}$ under $H_0$, the variances of $\hat{\beta}$ and $\tilde{\beta}$ are given by

$$Var(\hat{\beta}) = \sigma^2 (XX)^{-1}$$

(2.4)

$$Var(\tilde{\beta}) = \sigma^2 (XM_w X)^{-1}$$

$$= \sigma^2 H [WM, W']^{-1} H' + \sigma^2 (XX)^{-1}$$

(2.5)

where,

$$M_w = [I - W(WW)^{-1} W']$$

(2.6)

$$M_*= [I - X(XX)^{-1} X']$$

(2.7)

and

$$H = (XX)^{-1} XW$$

(2.8)

clearly $\tilde{\beta}$ is efficient.

Under $H_1$ one may have

$$E(\hat{\beta}) = \beta + (XX)^{-1} XZ$$

(2.9)

$$E(\tilde{\beta}) = \beta + (XX)^{-1} XZ - H(WM, W)^{-1} WM, Z$$

(2.10)

Where, $Z = E[\varepsilon / X]$

Under $H_0$, $(\hat{\beta} - \tilde{\beta}) \sim N(0, V)$

Where $V = \sigma^2 \left[ (XM_*, X)^{-1} - (XX)^{-1} \right]$

$$= \sigma^2 H [WM, W']^{-1} H'$$

Under $H_1$ one may obtain,
To test the null hypothesis, one may use any of the following two test statistics:

(i) \[ Q_1 = \left( \hat{\beta} - \tilde{\beta} \right)^\top \left[ H (W'M,W)\right]^{-1} H' \left( \hat{\beta} - \tilde{\beta} \right)/\sigma_1^{*2} \quad (2.12) \]

(ii) \[ Q_2 = \left( \hat{\beta} - \tilde{\beta} \right)^\top \left[ H (W'M,W)\right]^{-1} H' \left( \hat{\beta} - \tilde{\beta} \right)/\sigma_2^{*2} \quad (2.13) \]

Where, \( \sigma_1^{*2} = \frac{e_1^* e_1^*}{n-k} \) is the estimator of \( \sigma^2 \) based on Internally studentized residual vector \( e_1^* \) from the regression of \( Y \) on \( X \) and \( W \).

\[ \sigma_2^{*2} = \frac{e_2^* e_2^*}{n-(k+q)} \]

is the estimator of \( \sigma^2 \) based on Internally studentized residual vector \( e_2^* \) from the regression of \( Y \) on \( X \) and \( W \).

Under \( H_0 \): \( Q_1 \) follows asymptotically \( \chi_k^2 \) and \( Q_2 \) follows asymptotically \( \chi_{\text{Min}(k,q)}^2 \).

### III. A Modified Rainbow Test By Using Internally Studentized Residuals:

The main feature of the Rainbow test is that when the true relationship is nonlinear, a good linear fit can still be obtained over the subsets of the sample. The test therefore rejects the null hypothesis of linearity whenever the overall fit is markedly inferior to the fit over a properly selected sub sample of the data. This test is based on comparing two residual sums of squares; one being derived by using the complete data set and the other being obtained by using only a subset of \( n_1 \) observations. The sample data consists of \( n \) observations is assumed to be split into three parts or subsets containing \( n_0, n_1 \) and \( n_2 \) observations with \( k < n_1 < n \). Consider the classical regression model as

\[ Y_{\text{rel}} = X_{n \times k} \beta_{k \times 1} + \epsilon_{\text{rel}} \quad (3.1) \]

Suppose that \( e^* e^* \) be the Internally Studentized residual sum of squares obtained from (3.1) by using the complete data containing \( n \) observations. Also, \( e_M^* e_M^* \) be a Internally studentized residual sum of squares obtained from (3.1) by using only the middle subset of \( n_1 \) observations.

To test for the null hypothesis of no misspecification, the test statistic is given by

\[ F = \left[ \frac{e^* e^* - e_M^* e_M^*}{e_M^* e_M^*/n_1 - k} \right] / \left[ F_{n_1, (n_1-k)} \right] \quad (3.2) \]
The crucial problem of the Rainbow test is the proper choice of the middle subsample.

Remark: 1. Generally the sub sample size may be considered as \( n_i = \frac{n_0 + n_2}{2} = \frac{n}{2} \)

2. Rainbow test is not robust to non-normality.

IV. New Misspecification of Non-Nested Linear Regression Models:

One of the important problems in the Applied Regression Analysis is that the evaluation of the regression model. The evaluation of the regression model consists of the regression a variety of tasks: First, one should judge the reasonableness of model specifications; second, one can compare the model or certain parts of it with alternative formulations, using appropriate model selection criteria; Third, one may device measures of the degree of model specification; Fourth, one can use the accuracy of forecasts; and Finally, the use of a model for policy making and formulation. This list of tasks is incomplete, but it does represent a very substantial portion of the tools available for the evaluation of regression models.

In the statistical model building, a regression model can be misspecified in a number of ways. The major sources are incorrect functional form and invalid assumptions on distribution of the disturbance term in regression model. One of the most important components of statistical model building is test for misspecification of the regression model. The main problem with statistician is to test whether the functional form employed for the given variables is appropriate. There are two basic approaches to test for the functional misspecification. The first is to fit a new regression model in which an extra term (square or cube of regressor) is included. A test of significance on the coefficient of the variable then provides a test on the specification; of \( X_k \). The second approach which employs test statistics based on residuals from the original regression model. Consider the problem of testing non-nested linear regression models to form augmented model from two or more alternative hypotheses and test the restrictions implied by only one of them being true. Suppose that both the null and alternative hypothesis be linear regression models, which are having some common regressors.

Null Hypothesis: \( H_0: Y_i = X_i'\beta_1 + W_i'\beta_2 + \varepsilon_i \) \hspace{1cm} (4.1)

Alternative Hypothesis: \( H_1: Y_i = Z_i'\eta_1 + W_i'\eta_2 + \varepsilon_2i \) \hspace{1cm} (4.2)

Where \( Y_i \) is the \( i^{th} \) observation on the dependent variable;

\( X_i \) and \( Z_i \) are the (kx1) vectors of observations on regressors which are unique to \( H_0 \) and \( H_1 \) respectively;

\( W_i \) denotes the (qx1) vector of observations on the regressors which are common to both hypotheses;
Consider the augmented linear regression model as
\[
Y_i = \alpha X_i' \beta_1 + (1-\alpha) Z_i' r_1 + W_i \left( \alpha \beta_2 + (1-\alpha) r_2 \right) + \epsilon_i
\]
(4.3)
or
\[
Y_i = \alpha \left[ X_i' \beta_1 + W_i \beta_2 \right] + (1-\alpha) \left[ Z_i' r_1 + W_i r_2 \right] + \epsilon_i
\]
(4.4)
or
\[
Y_i = X_i' \beta_1^* + W_i \left[ \beta_2^* + (1-\alpha) r_2 \right] + (1-\alpha) Z_i' r_1 + \epsilon_i
\]
(4.5)
where \( \beta_1^* = \alpha \beta_1 \) and \( \beta_2^* = \alpha \beta_2 \).

Under H₁, one may estimate the model (4.5) and obtain the OLS estimates for \( r_1 \) and \( r_2 \) as \( \hat{r}_1 \) and \( \hat{r}_2 \) respectively.

Write the augmented linear regression model as
\[
Y_i = X_i' \beta_1^* + W_i \delta_1^* + Z_i' \delta_2^* + \epsilon_i
\]
Where \( \beta_1^* = \alpha \beta_1 \), \( \delta_1^* = \alpha \beta_2 + (1-\alpha) \hat{r}_1 \); \( \delta_2^* = (1-\alpha) \hat{r}_1 \)

one may test that \( \delta_2^* = 0 \), which implies that \( (1-\alpha)r=0 \). For given \( \alpha \), F-test statistic for testing the significance of \( \delta_2^* \) can be considered as a test for misspecification of the regression model.

V. Conclusions:
In practice, the statistical model may not be correctly specified. Over specification yields unbiased estimates of regression coefficients, but larger variance; under specification yields biased estimates of the regression coefficients and underestimates the variance of these estimates. In empirical research, one often faces the problem of estimating a mis-specified statistical model. In this Article, three types of new tests for the mis-specification of the linear regression model have been developed by using various types of residuals.

References


**Corresponding Author:**

G. Mokesh Rayalu*

Email: mokesh.g@vit.ac.in