CHAPTER - IV

A BRIEF ABOUT MODEL SPECIFICATION TESTS
4.0 INTRODUCTION

One of the important problems in econometric analysis is that the evaluation of an econometric model. It can and should be approached in different ways, and the different approaches complement one another. There does not exist a simple way to evaluate an entity as complex as an econometric model. In practice, when the model is specified, its parameters are estimated, passing the tests of correctness in sign and statistical significance at reasonable levels, and the residuals are not obviously correlated, it is often considered a completed model. The evaluation of the model consists of a variety of tasks: First, one should appeal to economic theory to judge the reasonableness of model specifications. Second, one can compare the model or certain parts of it with alternative formulations, using appropriate statistical criteria for choosing among them. Third, one may devise measures of the degree of model specification. Fourth, one can use the accuracy of forecasts to evaluate a model. Finally, the use of a model for policy analysis and formulation. The above listing is by no means complete, but it does represent a very substantial portion of the tools available for the evaluation of econometric models.

In this chapter, the various model specification tests have been explained by giving their interrelationships. Some tests for functional misspecification in regression analysis have been discussed based on OLS and recursive residuals.
4.1 SIGNIFICANCE TESTS IN MULTIPLE REGRESSION

In multiple regression, there may be some apparent contradiction between the results obtained by applying various significance tests. Writing $\beta_i$ for the partial regression coefficients and $R^2$ for the coefficient of determination, Geary and Leser (1968) have distinguished the following cases at any given level of significance:

(i) $R^2$ and all $\beta_i$ significant;
(ii) $R^2$ and some but not all $\beta_i$ significant;
(iii) $R^2$ but none of the $\beta_i$ significant;
(iv) all $\beta_i$ significant but not $R^2$;
(v) some $\beta_i$ significant but not all nor $R^2$;
and (vi) neither $R^2$ nor any $\beta_i$ significant.

Case (i) and case (vi) do not pose any problem. Case (ii) is very common, particularly with a large number of explanatory variables, and the only problem is to decide which of the non-significant variables, if any, to exclude from the equation. Computer softwares are available to exclude those variables which do not obtain significant regression coefficients. In case (v), one would generally be inclined to give procedure to the test for the regression as a whole and to consider the entire equation as doubtful, bearing in mind the fact that with a
moderately large number of variables, one may well expect some apparently significant individual coefficients to occur. Particularly in econometric applications, in which the exact specification of the relationship is often uncertain, one would then experiment with alternative sets of variables or possibly alternative mathematical formulations. Case (iii) and case (iv) are those which seem quite paradoxical and it may be worth while to inquire in which circumstances they can occur. Geary and Leser (1968) observed from case (iii) that a significant F-ratio does not indicate the significance of any given regression coefficient but merely the existence of at least one linear combination which is significantly different from zero, case (iv) is even more troublesome. The fact that the F-value is sufficiently lower than the value of the lowest $\eta^2$ does not guarantee that case (iv) arises. More likely than not, one can also obtain significance for $R^2$ or non-significance for some of the $t$.

If the situation of case (iv) arise, it is not easy to see what conclusions one would draw. It has been shown that it depends upon a combination of circumstances which is practice will not often be fulfilled, and the possibility of its existence represents little more than a curiosity.
4.2. MISSPECIFICATION IN MULTIPLE REGRESSIONS

Consider the classical linear regression model where all the independent variables are nonstochastic and the error terms are homoscedastic and serially independent. Let two regression equations, of which only one is truth, be:

\[ Y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \epsilon_i \quad \ldots \quad (4.2.1) \]

\[ Y_i = \Gamma_1 x_{1i} + \Gamma_2 x_{2i} + \ldots + \Gamma_k x_{ki} + \Gamma_{k+1} x_{k+1i} + \eta_i \quad \ldots \quad (4.2.2) \]

\[ i = 1, 2, \ldots, n. \]

When equation (4.2.1) is the truth, equation (4.2.2) is a misspecified model because of the presence of the irrelevant variable \( x_{k+1} \). When equation (4.2.2) is truth, equation (4.2.1) is a misspecified model because of the left out variable \( x_{k+1} \). Equations (4.2.1) and (4.2.2) may be written in the matrix form as

\[ Y = X\beta + \epsilon \quad \ldots \quad (4.2.3) \]

\[ Y = \bar{X}\Gamma + \eta \quad \ldots \quad (4.2.4) \]

Where \( Y \) is a vector of observations the dependent variable, and \( X \) and \( \bar{X} \) are matrices of independent variables in the equations (4.2.1)
and (4.2.2) respectively. With out loss of generality the equation (4.2.3) may be rewritten as

$$Y = [X^\top X_{k+1}]^{-1} \cdot \varepsilon$$  \hspace{1cm} (4.2.5)

Where $X_{k+1}$ is a vector of observations on the independent variable $X_{k+1}$. Potluri Rao (1970) proved the following theorems.

**Theorem (4.1)**: In the classical linear regression model, omission of a variable specified by the truth, introduces bias in all the least squares estimates.

**Proof**: Let (4.2.2) be the truth, and let a misspecified model, given by equation (4.2.1) be estimated.

The OLS estimator of $\beta$ from (4.2.1) is given by

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$  \hspace{1cm} (4.2.6)

Since, equation (4.2.2) is the truth, one may rewrite the equation (4.2.6) as

$$\hat{\beta} = (X^\top X)^{-1} X^\top (\bar{X} \Gamma + \eta)$$

$$= (X^\top X)^{-1} X^\top \bar{X} \Gamma + (X^\top X)^{-1} X^\top \eta$$

194
\[ E(\hat{\beta}) = (X'X)^{-1}X'Y \] ... (4.2.7)

or

\[ E(\hat{\beta}) = (X'X)^{-1}X'(X'X_{k+1})' \]

\[ = (\gamma_1 \gamma_2 \cdots \gamma_k)' + \gamma_{k+1}'(X'X)'X'X_{k+1} \] ... (4.2.8)

To facilitate the interpretation of the results, Potluri Rao (1970) introduced an auxiliary regression equation in Yule's notation as

\[ X_{k+1} = h_{k+1,123} \cdots X_1 + h_{k+1,213} \cdots X_2 \]

\[ + \cdots + h_{k+1,k12} \cdots X_k + e_{k+1} \] ... (4.2.9)

Where b's are the OLS estimators, and e_{k+1} is the residual. The auxiliary regression equation (4.2.9) is introduced only an algebraic convenience and need not have any causal interpretation.

Now, we have the expected value of the regression coefficient of the independent variable X, may be written as

\[ E(\hat{\beta}_i) = \gamma_1 + \gamma_{k+1}h_{k+1,123} \cdots k \] ... (4.2.10)
The bias in the regression coefficient \( \hat{\beta}_1 \) is proportional to the auxiliary regression coefficient \( h_{k+1} \) when the auxiliary regression coefficient is zero, then the bias is zero, but this case is rare in empirical work.

**Theorem (4.2)**: In the classical linear regression model, omission of a variable specified by the truth decreases the variance of all the least squares estimates.

**Proof**: Let the truth be given by equation (4.2.2) and the misspecified model be equation (4.2.1), so that the left out variable is \( X_{k+1} \). The least squares estimates of \( \beta \)'s given by equation \( \hat{\beta} = (X'X)^{-1} X'Y \)

The variance of the estimate vector \( \hat{\beta} \) is given by

\[
V(\hat{\beta}) = E[\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})]'
\]

\[
= E\left[ (X'X)^{-1} X'e e' X (X'X)^{-1} \right] = \sigma^2_0 (X'X)^{-1} \quad \text{(4.2.11)}
\]
consider the variance of the regression coefficient corresponding to the independent variable $X_1$. The variance of the regression coefficients may be rewritten in the partitioned matrix form as

$$V(\hat{\beta}) = \sigma^2\eta \begin{bmatrix} X_1'X_1 & X_1'Z \\ Z'X_1 & Z'Z \end{bmatrix}$$ \hspace{1cm} \text{(4.2.12)}$$

Where $X_1$ is a vector of observations on the independent variable $X_1$, and $Z$ is a matrix of observations on the rest of the independent variables in equations in (4.2.1)

We have,

$$V(\hat{\beta}) = \sigma^2\eta \begin{bmatrix} X_1'X_1 & X_1'Z(Z'Z)^{-1}Z'X_1 \\ (Z'Z)^{-1} & (Z'Z)^{-1}A \\ B & C \end{bmatrix}$$ \hspace{1cm} \text{(4.2.13)}$$

Where $A$ and $B$ are vectors, and $C$ is a matrix.

\[ \therefore \] The variance of the regression coefficient $\hat{\beta}_1$ is given by

$$V(\hat{\beta}_1) = \sigma^2\eta \begin{bmatrix} X_1'X_1 & X_1'Z(Z'Z)^{-1}Z'X_1 \\ (Z'Z)^{-1} & (Z'Z)^{-1}A \\ B & C \end{bmatrix}^{-1}$$ \hspace{1cm} \text{(4.2.14)}$$

Consider an auxiliary regression equation

$$x_1 = b_{1.234}x_2 + b_{1.324}x_3 + \ldots + b_{1.k}x_k + \epsilon_1.$$ \hspace{1cm} \text{(4.2.15)}$$
Where the b's are the least squares estimates and $e_i$ is the residual.

The residual sum of squares in the auxiliary regression equation \( \sum e_i^2 \) may be written in Yule's notation as

\[
S^2_{1,2,3,...,k} = \sum e_i^2 = X_1^\dagger X_1 - X_1^\dagger Z (Z^\dagger Z)^{-1} Z^\dagger X_1 \quad \ldots (4.2.16)
\]

\[
\therefore \quad V(\hat{\beta}_1) = \sigma^2_\eta / S^2_{1,2,3,...,k} \quad \ldots (4.2.17)
\]

It is well known that the variance of the least squares estimates of the true equation (4.2.2) are

\[
V(\hat{\beta}) = \sigma^2_\eta (\hat{\beta}_1^\dagger \hat{\beta}_1)^{-1} \quad \ldots (4.2.18)
\]

By using the above analysis, the variance of \( \hat{y}_1 \) is given by

\[
V(\hat{y}_1) = \sigma^2_\eta / S^2_{1,2,3,...,k,k+1} \quad \ldots (4.2.19)
\]

Where \( S^2_{1,2,3,...,k,k+1} \) is the residual sum of squares in an auxiliary regression with \( x_1 \) as the dependent variable and \( (x_2, x_3, \ldots, x_n, x_{n+1}) \) as the independent variables. Since the residual sum of squares can not increase by adding an independent variable to a regression equation, it follows that,
Thus, we have, 

\[ V(\hat{\beta}_1) - V(\hat{\gamma}_1) \] ... \((4.2.20)\)

This inequality becomes an equality only when the partial relation between the independent variable \(x_1\) and the left out variable \(x_{k+1}\) holding \((x_2, x_3, ..., x_k)\) constant zero; this case seldom arises in empirical work. Since the choice of \(x_1\) is arbitrary, the inequality \((4.2.20)\) holds for all coefficients.

**Theorem (4.3):** In the classical linear regression model, discarding an independent variable whose parameter value is smaller than the theoretical standard, deviation of its estimate will decrease the mean square error of all the least squares estimates.

**Proof:** Wallace (1964) proved this theorem in a case of only two independent variables.
It has been shown that when equation (4.2.2) is the truth, and equation (4.2.1) is estimated, the regression coefficients are biased but have smaller variance when compared to the corresponding estimates from the true model. The Mean Square Error (MSE) may, however, increase or decrease depending on whether the gain in variance is compensated by the loss in the bias or not. The MSE's in the two cases are:

\[
\text{MSE} \left( \beta_1 \right) = \gamma_k^2 \cdot \frac{b^2}{\eta_{k+1, 23}} + \frac{\sigma^2}{\eta} \cdot \frac{\bar{y}_1^2}{\bar{y}_{123, k}} \quad \ldots (4.2.21)
\]

\[
\text{MSE} \left( \hat{\gamma}_1 \right) = \frac{\sigma^2}{\eta} \cdot \frac{\bar{y}_1^2}{\eta_{123, k}} \cdot \frac{\bar{y}_1^2}{\eta_{k+1, 23}} \quad \ldots (4.2.22)
\]

By using the following known results from residual analysis.

(i). \[ S_{123, k+1}^2 = S_{123, k}^2 \left( 1 - \frac{r_{1k+1, 23}}{r_{1k, 23}} \right) \quad \ldots (4.2.23) \]

(ii). \[ b_{k+1, 23} = r_{1k, 23} \cdot \frac{S_{1k+1, 23}}{S_{123, k}} \quad (4.2.24) \]

Where \[ r_{1k+1, 23} \] is the partial correlation between the variables \( x_1 \) and \( x_{k+1} \) keeping all the other independent variables \( (x_2, x_3, \ldots, x_k) \) constant; the MSE's of \( \hat{\beta}_1 \) and \( \hat{\gamma}_1 \) may be rewritten as.
MSE (\(\hat{\beta}_1\)) = \gamma^2_{k+1} \frac{r^2_{k+1,23...\kappa}}{1} \left( \frac{S^2_{k-1,23...k}}{S^2_{123...k}} \right) + \frac{\sigma^2_\eta}{S^2_{123...k}} \quad \ldots(4.2.25)

MSE (\(\hat{\gamma}_1\)) = \frac{\sigma^2_\eta}{S^2_{123...k} \left( 1 - r^2_{k+1,23...k} \right)} \quad \ldots(4.2.26)

Consider MSE (\(\hat{\beta}_1\)) \leq MSE (\(\hat{\gamma}_1\))

\gamma^2_{k+1} \frac{r^2_{k+1,23...k}}{1} \left( \frac{S^2_{k-1,23...k}}{S^2_{123...k}} \right) \frac{\sigma^2_\eta}{S^2_{123...k}} \quad \ldots(4.2.27)

\gamma^2_{k+1} \frac{S^2_{k+1,23...\kappa} \left( 1 - r^2_{k+1,23...\kappa} \right)}{S^2_{k+1,123...\kappa}} \cdot \frac{\sigma^2_\eta}{S^2_{123...k}} \quad \ldots(4.2.28)

Since,

\frac{S^2_{k+1,23...\kappa}}{S^2_{k+1,123...\kappa}} = \frac{S^2_{k+1,123...\kappa} \left( 1 - r^2_{k+1,123...\kappa} \right)}{S^2_{123...k}} \quad \ldots(4.2.29)

one may rewrite equation (4.2.28) as

\gamma^2_{k+1} \leq \frac{\sigma^2_\eta}{S^2_{k+1,123...\kappa}} \quad \ldots(4.2.30)
But the variance of the estimate $\hat{y}_{k+1}$ in the true equation (4.2.2) is given by

$$V(\hat{y}_{k+1}) = \frac{\sigma^2}{S^2_{k+1,123..}}.$$  ...(4.2.31)

The condition under which $\text{MSE}(\hat{\beta}_1) \leq \text{MSE}(\hat{\gamma}_1)$, is given by

$$\gamma^2_{k+1} \leq \frac{\sigma^2}{S^2_{k+1,123..}} V(\hat{y}_{k+1})$$

or

$$\gamma_{k+1} \leq \sqrt{V(\hat{y}_{k+1})}.$$  ...(4.2.32)

Thus, when the absolute value of the parameter $\gamma_{k+1}$ is smaller than the theoretical standard deviation of the estimate $\hat{\gamma}_{k+1}$, the mean square error of $\hat{\beta}_1$ would be decreased by discarding the variable $x_{k+1}$.

Since, the choice of the independent variable $x_1$ is arbitrary, the relation holds for all the independent variables.
Theorem (4.4): In the classical linear regression model, inclusion of an irrelevant variable does not introduce bias in the least squares estimates.

Proof: Let the truth be given by equation (4.2.1) and let the misspecified model be equation (4.2.2), with the irrelevant variable \( x_{k+1} \) be estimated. The least squares estimate of the misspecified model, equation (4.2.2) is given by:

\[
\hat{\beta} = \left( \tilde{X}^\top \tilde{X} \right)^{-1} \tilde{X}^\top \tilde{Y} \tag{4.2.33}
\]

using equation (4.2.5), one may rewrite equation (4.2.33) as

\[
\hat{\beta} = \left[ X^\top \ X \right]^{-1} X^\top \beta + \epsilon \tag{4.2.34}
\]

\[
\therefore \ E \left( \hat{\beta} \right) = \beta \tag{4.2.35}
\]

Estimation of the true equation (4.2.1) gives

\[
E \left( \hat{\beta} \right) = \beta \tag{4.2.36}
\]

Hence, the least squares estimates from the misspecified model (4.2.2), when equation (4.2.1) is the truth, are unbiased.
Theorem (4.5): In classical linear regression model, inclusion of an irrelevant variable increases the variance of all the least squares estimates.

Proof: When equation (4.2.1) is the truth, and equation (4.2.2) is estimated, the variance of the least squares estimates, given by equation (4.2.34) is:

\[ \nu(\hat{\beta}) = \sigma_e^2 \left( X'X \right)^{-1} \]

Further, from equation (4.2.1), one may have

\[ \nu(\hat{\beta}) = \sigma_e^2 \left( X'X \right)^{-1} \]

From inequality (4.2.20), one can show that

\[ \nu(\hat{\gamma}_1) : \nu(\hat{\beta}_1) \]

Even though an irrelevant variable does not introduce any bias in the regression coefficients, its presence increases the variance of all the estimates of the regression coefficients.

Theorem (4.6): In the classical linear regression model, inclusion of an irrelevant variable increases MSE of all the least squares estimates.

Proof: It can be easily proved from the theorems (4.4) and (4.5).
Potluri Rao (1970) presented the above results, which can be served as theoretical guidelines in empirical research. These results reveal the consequences of omission or inclusion of a variable. For instance, when a researcher is interested in using the regression estimates in decision making, he needs the least MSE estimates rather than the best linear unbiased estimates. In such a case the researcher may gain by excluding a variable even though the truth model specifies the variable as a part of the multiple regression. In certain empirical studies, researchers include variables to maximize $R^2$, or some other summary statistic, even though there are no theoretical reasons for their inclusion in the regression equation. Potluri Rao (1970) concluded that such a procedure may result in loss of efficiency.

4.3 USE OF ORTHOGONAL FACTORS IN PRINCIPAL COMPONENTS FOR SELECTION OF VARIABLES IN A REGRESSION EQUATION

Selection of explanatory variables in a multiple regression analysis has been a prime problem in the analysis of unplanned data. The interdependency among the explanatory variables makes it difficult to determine empirically the condition of each independent variable to the observed variation of the dependent variable. Various alternative selection procedures have been proposed, but the criterion employed in each technique seems quite arbitrary and is known to provide different solutions the same problem.
Daling and Tamura (1970) suggested an approach to the problem of variables in regression analysis. This method purports to come up with a prediction equation in accordance with the principle of parsimony in terms of minimum interdependency among variables. Their technique can be viewed as a use of the principal components regression proposed by Kendall (1957). This technique was first introduced in the article by Wallis (1965). Daling and Tamura (1970) have described the selection technique which makes use of the orthogonality among factors extracted from the correlation matrix. Using the factors not as new variables, but merely as the reference frame, one can identify a near orthogonal subset of explanatory variables.

Let $R$ be the correlation matrix of $k$ explanatory variables $x_i \ [i = 1, 2, \ldots, k]$. Let $C$ be an orthogonal matrix such that $C' RC = D$

Where $D$ is the diagonal matrix of eigenvalues $\lambda_i$.

The principal components are the elements of the vector $G = C'X$. The sample covariance of $G$ is $D$. Thus, the regression of the dependent variable $Y$ on the components $g_i$ of $G$ gives orthogonal regression;
By normalizing each principal component \( g_i \) by dividing it by the square root of the corresponding eigenvalue \( \lambda_i \), an orthonormal set of variables are given by

\[
f_i = \frac{g_i}{\sqrt{\lambda_i}}, \quad i = 1, 2, \ldots, k.
\]  
(4.3.1)

i.e., \( f_i \)'s have the unit variance and are uncorrelated in the data. From,

\[
F = D^{-1/2}G = D^{-1/2}C^T X.
\]  
(4.3.2)

We express,

\[
X = C D^{1/2} F
\]  
(4.3.3)

The matrix \( A = CD^{1/2} \) is the principal factor loading matrix, which is a particular solution of the general factor analysis model

\[
\text{ie., } X = AF
\]  
(4.3.4)

The element \( a_{iy} \) of \( A \) indicates the correlation between the variable \( x_i \) and \( f_y \). Thus, if the loading \( a_{iy} \) is high, i.e., if \( x_i \) and \( f_y \) are highly correlated, and if \( f_y \) contributes significantly to \( Y \), then, in turn, \( x_i \) should be making a significant contribution to \( Y \).

The ideal pattern of the factor loading matrix for selecting the variables is that every factor comes in with the loading of 1 uniquely for
one explanatory variable and with zero loadings for the rest. Such a
dfactor loading matrix is impossible to obtain when the variables are
independent. It should be noted that $A = (D^T D)^{-1} D^T$ is not the unique
solution to (4.3.4) in the sense that there exist alternative solutions $BH$
which also have the same correction matrix $R$ of $X$. For example, let $E$
be an orthogonal matrix. Then

$$H = E^T F$$
is another solution to (4.3.4) with the loading matrix

$$B = AE$$

... (4.3.5)

Thus BH is another expression of $X$ in terms of orthogonal factors
$H$, and $B$ indicates the correlation between $X$ and $H$ as before. By
properly specifying $E$, one might obtain a loading matrix $B$ which is a
best possible approximation to ideal pattern that the data permit. This
approach is referred to as simplification of loading matrix. Several
alternative methods have been developed, among which the Quartimax
and the Varimax methods are two well known methods of orthogonal
rotation of factors. Varimax rotation method is more effective for
selecting variables.

118.
By regressing the dependent variable $Y$ on each varimax component $h_k$ of $H$, one could more easily identify the explanatory variables which, having minimum interdependence among themselves, appear to make a significant contribution to the variation $Y$. It is reasonable to assume that the factors with very small eigenvalues are not expected to make a significant contribution.

On the other hand, when the only factors correlated with the dependent variable are those corresponding to low eigenvalues then a regression equation with high $R^2$ is unlikely; it indicates lack of sufficient information about the dependent variable in the explanatory variables in analysis. Daling and Tamura (1970) have indicated that their approach provides the model builder with the flexibility that is not available in the conventional and purely mechanical selection methods.

4.4 ALLEN'S MEAN SQUARE ERROR OF PREDICTION CRITERION FOR SELECTING VARIABLES

The problem of selecting variables in multiple regression has received a great deal of attention. Most of the procedures for selecting variables given by Garside (1965), Schatzoff, Feinberg and Tsao (1968),
Hocking and Leslie (1967), Beale, Kendall and Mann (1967), La Motte and Hocking (1970) and others have used the residual sum of squares as a criterion for selecting variables. Allen (1971) raised two objections to the use of residual sum of squares as a criterion for selecting variables.

(i) If the residual sum of squares were the sole criterion, then would always use all of the variables. Thus, to delete variables, there must be an additional criterion such as the number of variables used. The degree to which the two criteria are weighted is arbitrary.

(ii) The residual sum of squares is not directly related to the commonly used criteria for good prediction and estimation.

Allen (1971) suggested a criterion along with computing procedure for selecting variables for the purpose of prediction. The selection criterion is based on the mean square error (MSE) criterion. The use of MSE takes into account the expected value of the observation being predicted and estimates the arbitrariness associated with the residual sum of squares. Allen's procedure is a modification of the method given by Anderson, Allen and Cady (1970).

Consider the classical linear regression model as

\[ Y_{(n \times 1)} = X_{(n \times k)} \beta_{(k \times 1)} + \epsilon_{(n \times 1)} \]  \hspace{1cm} (4.4.1)
Where \( Y \) is a \((n \times 1)\) vector of observations on dependent variable;

\( X \) is a known, full rank \((n \times k)\) matrix of nonstochastic variables;

\( \beta \) is the unknown weight \((k \times 1)\) vector corresponding to \( X \);

and \( \varepsilon \) is a normally distributed \((nx1)\) vector of random variables having expected value 0 and dispersion matrix \( \sigma^2 I \).

In this model, \( X \)'s may represent different functional forms of the same basic variables, example, \( X_1 = w \), \( X_2 = v \), \( X_3 = w^2 \), \( X_4 = v^2 \), \( X_5 = uv \).

Let row vector \( x_{1:k} \) contains the values of the \( X \) variables associated with the future observation. Given the data generated by this model, one wishes to predict the value of a future random variable \( y \) having mean \( x_{1:k} \beta \) and variance \( \sigma^2 \). The row vector \( x \) contains the values of the \( X \)-variables associated with the future observation. The mean square error prediction is given by

\[
MSEP = \mathbb{E} [(\hat{y} - y)^2] \quad \text{...(4.4.2)}
\]

The MSE of a predictor \( \hat{y} \) can be expressed as

\[
\mathbb{E} [(\hat{y} - y)^2] = \sigma^2 + \text{Var}(\hat{y}) \cdot \left[ \mathbb{E} (\hat{y}) - x \beta \right]^2 \quad \text{...(4.4.3)}
\]
The last term \([E(\hat{y}) - x\beta]^2\) is the squared bias of prediction.

The last two terms of (4.4.3) are the mean square error (MSE) of \((\hat{y})\), which is given by

\[
\text{MSE} (\hat{y}) = \text{Var} (\hat{y}) + [E(\hat{y}) - x\beta]^2 \quad \cdots (4.4.4)
\]

Consider the least squares predictor is

\[
\hat{y} = x\hat{\beta}
\]

Where \(\hat{\beta} = (X'X)^{-1} X\hat{y}\). The least squares predictor is unbiased and has variance \(x(X'X)^{-1} x'\sigma^2\).

Thus, its MSE of prediction is

\[
\text{MSEP} = \sigma^2 + x(X'X)^{-1} x'\sigma^2 \quad \cdots (4.4.5)
\]

MSEP is the basis for the comparison of other predictors. As alternatives to the least squares predictor, one can consider the class of last squares predictors based on linear functions of the columns of \(X\) and \(x\). From the class, one selects the member having the smallest estimated MSEP.
Write the linear functions of the columns of $X$ and $x$ are

$$XP_* \text{ and } xP_*$$

Where $P_*$ is in $(k \times q)$ matrix of full rank. As an example, suppose $k = 5$, and one wishes to examine $X_1, X_2 + X_3, X_4 + X_5$ as predictor variables. Then

$$P_* = \begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1
\end{bmatrix}$$

If 'q' variables are used as predictors, then $t = k - q$ variables are excluded. A convenient representation for the excluded variables is $XP_*$ and $xP_*$. Where $P_*$ is an $(k \times t)$ matrix of full rank such that $P_*^\top P_* = 0$.

The predictor based $XP_*$ and $xP_*$ is

$$\hat{y}_* = xP_* \left( P_*^\top X P_* \right)^{-1} P_*^\top X Y \quad \ldots (4.4.6)$$

Using the matrix identity,

$$P_* \left( P_* X X P_* \right)^{-1} P_* = (X X)^{-1} = (X X) P_* \left[ P_*(X X) \right]^{-1} = P_*(X X)$$

123
It is observed that \( \hat{y} = (1 - z) \hat{\beta} \)

Where \( Z = x \left( X^t X \right)^{-1} x^t \left[ P_x \left( X^t X \right) P_x^t \right] P_x \). The variance of \( \hat{y} \)
is given by \( \sigma^2 + x \left( X^t X \right)^{-1} x^t \sigma^2 Z \left( X^t X \right)^{-1} Z^t \sigma^2 \).

Since, the first two terms comprise the variance of \( \hat{y}_k \) (From equation (4.4.5)) and the third term is nonnegative, it can be concluded that the variance of \( \hat{y} \) is less than or equal to the variance of \( \hat{y}_k \). The bias of \( \hat{y} \) is \( X \beta \) and thus its mean square error of prediction is

\[
\text{MSEP} = \sigma^2 + x \left( X^t X \right)^{-1} x^t \sigma^2 Z \left( X^t X \right)^{-1} Z^t \sigma^2 + (Z \beta)^2 \quad \text{(4.4.7)}
\]

Since, the first two terms of MSEP are constant, one need to consider only

\[
(Z \beta)^2 \quad \text{(4.4.8)}
\]

It should noted that (4.4.8) is simply \( \text{MSEP}_p \) - \( \text{MSEP} \). If this quantity is negative, then \( \hat{y} \) is better than \( \hat{y}_k \).
Since (4.4.8) depends upon unknown parameters $\beta$ and $\sigma^2$, we base our technique on its estimator

$$ (Z \hat{\beta})^2 - 2Z(X^\prime X)^{-1}Z^\prime S^2 \quad \ldots (4.4.9) $$

Where,

$$ S^2 = (n - i \times 2)^{-1}X^\prime (I - X(X^\prime X)^{-1}X^\prime)W $$

This estimator is the minimum MSE estimator of the class

$$ (Z \hat{\beta})^2 - IS^2 \quad \ldots (4.4.10) $$

Suppose there are a finite number of "reasonable" values of $P_\ast$. Our predictor is $\hat{y}$ evaluated at the $P_\ast$ for which (4.4.9) is minimum.

When feasible, the appropriate $P_\ast$ is found by evaluating (4.4.9) at each potential value of $P_\ast$.

Criterion (4.4.9) provides a comparison between submodels for the same future observation. However, one can not compare the performance of one submodel for two different future observations using (4.4.9).

One can compare the estimated relative efficiencies of $\hat{y}_k$ to $\hat{y}$ for different future observations.
More specifically, 
\[ \frac{-\left( \mathbf{Z} \mathbf{\hat{b}} \right)^2 - 2 \mathbf{Z} (X^\prime X)^{-1} \mathbf{Z}^\prime s^2}{(1 + x(X^\prime X)^{-1} x^\prime) s^2} \] ... (4.4.11)

is an estimator of \( 1 - \frac{\text{MSEP}_k}{\text{MSEP}} \) ... (4.4.12)

The quantity (4.4.12) is the proportionate reduction in MSEP due to having used the submodel rather than the full model. While (4.4.12) has a maximum value of 1, the maximum value of (4.4.11) is 
\[ \frac{2 \mathbf{Z} (X^\prime X)^{-1} \mathbf{Z}^\prime}{1 + x(X^\prime X)^{-1} x^\prime} \]

Thus, if the variance of \( \hat{y}_k \) is large and the variance of \( \hat{y} \) is small, ie., \( \mathbf{Z} (X^\prime X)^{-1} \mathbf{Z}^\prime \) approaches \( (X^\prime X)^{-1} \) in magnitude, then (4.4.11) may be greater than 1.
4.5 RAMSEY'S TESTS FOR SPECIFICATION ERRORS IN CLASSICAL LEAST SQUARES REGRESSION ANALYSIS

Ramsay (1969) first, derived the distributions of the classical linear least squares residuals under a variety of specification errors. The four types of specification errors considered by him are: (i) Omitted variables (ii) Incorrect functional form (iii) Heteroscedasticity and (iv) Simultaneous equations error problems. Later, he derived procedures to test for the presence of the specification errors in the regression models. The tests are developed by comparing the distribution of residuals under the hypothesis that the specification of the model is correct to the distribution of the residuals yielded under the alternative hypothesis that there is a specification error of one of the aforementioned types.

Ramsay (1969) developed four types of tests namely RESET, RASET, KOMSET and BAMSET. RESET is useful to test for the significance of a regression of the residuals or a linear function of vectors \( q_j, j = 1, 2, \ldots \) RASET is a rank correlation coefficient test between \( q_i \) and a ranking of the squared residuals. KOMSET is the application of the Kolomagorov test statistic to discriminate between the central and non-central F distribution. BAMSET is a modification of Bartlett's M test for the heterogeneity of variance.
4.5.1 The Regression Specification Error Test (RESET)

Consider the classical linear model with full ideal conditions as

\[ y_{nm1} = X_{nm1} \beta_{kl1} + \varepsilon_{nm1} \] \hspace{1cm} (4.5.1)

\[ \varepsilon \sim N (0, \sigma^2 I_n) \]

Under the null hypothesis that the full ideal conditions hold, the Theil's BLUS residual vector \( \hat{\varepsilon} \) is distributed as normal with null mean vector and covariance matrix \( \sigma^2 I_{n-k} \). Under the alternative hypothesis, \( \hat{\varepsilon} \) is distributed as normal, but with mean vector given \( A^T \eta \) and the covariance matrix by \( \sigma^2 I_{n-k} \).

In the omitted variables case of specification error, the vector \( A^T \eta \) can be regarded as the net contribution of the omitted variable to the conditional mean of the dependent variable, where 'net' means that one is considering the effect of the omitted variable given the other regressors. Thus, under the alternative hypothesis, one may write \( \varepsilon \) as

\[ \varepsilon = A^T \eta + u \] \hspace{1cm} (4.5.2)

Where the \((n-k) \times 1\) vector \( u \) is normally distributed with null mean vector.
The model formulated in equation (4.5.2) suggests the possibility of testing for specification error under these assumptions by regressing \( \tilde{e} \) on \( \xi = A^1 \eta \) and performing F-test on the estimated regression. However \( \eta \) is unknown. One may approximate the regressor vector \( \xi \) by a linear sum of the vectors \( q_j, j = 1, 2, \ldots \).

Thus one considers the equation:

\[
c_i = \alpha_0 + \alpha_1 q_{i1} + \alpha_2 q_{i2} + \ldots + u_i \quad (4.5.3)
\]

\[i = 1, 2, \ldots, n+1\]

Where \( u_i \) is assumed to be normal with zero mean, and one applies an F-test on the joint significance of the \( \alpha \)'s. No general rule can be specified on number of \( q_j \)'s needed. However, Ramsay (1969) has found that using \( q_1, q_2 \) and \( q_3 \) has been sufficient.

Under the null hypothesis, the regression coefficient estimators have zero means and the covariance matrix of \( u \) is \( \sigma^2 l_{n-k} \). Under the alternative hypothesis that the mean vector of \( e \) is \( A^1 \eta \), one or more of the coefficient estimates will have nonzero means, i.e., multiple correlation coefficient is nonzero. Under the alternative hypotheses of various specification errors, the covariance matrix of \( U \) is given by
\( \sigma^2 I_{n-k} \), where \( \sigma^2 \), is the variance of \( \tilde{e}_i \) about its conditional mean. If the alternative hypothesis is that \( \tilde{e} \) is distributed as \( N(0, \Omega) \), where \( \Omega \) is a diagonal matrix with unequal elements, (i.e., the case heteroscedasticity) then one cannot use the regression test to discriminate against such an alternative. Further, if the alternative hypothesis is that the distribution of \( e \) is not normal then the regression test is no longer well defined.

4.5.2 The Rank Specification Error Test (RASET)

Consider the squared BLUS residuals \( e_i \)'s as independent estimates the error variance. Under the null hypothesis, \( \tilde{e}_i^2, i = 1, 2, \ldots, n - k \), are distributed as \( \sigma^2 X^2_1 \), where as under the three cases of specification error hypotheses, \( \left( \frac{e_i^2}{\sigma^2} \right), i = 1, 2, \ldots, n - k \) are distributed as \( \chi^2 (1, \lambda_i) \), where \( \sigma_i^2 \) is the variance of \( e_i \) about its conditional mean \( \hat{\rho} \). \( \lambda_i = \frac{\eta a_i}{\sigma^2} \). It should be noted that the larger \( \lambda_i, i = 1, 2, \ldots, n - k \), the greater the power of tests discriminating between central non central \( \chi^2 \) distributions. Further, one can show that for the case of an omitted variable.
\[ \sum_{i=1}^{n-k} \frac{\hat{\lambda}_i - E_1^{\text{ARSS}}}{\sigma^2} = 1 \] ... (4.5.4)

Where ARSS is the extra regression sum of squares due to the omitted variable and \( \hat{\sigma}^2 \) is the variance of the error terms.

Under the RASET procedure, one rearranges the \( q_{ij} \) into ascending order permuting the elements of the vector \( \hat{c} \) conformably and one transforms \( q_{ij} \rightarrow j, j = 1, 2, \ldots, n - k \). Where \( j \) denotes the index of the \( j^{\text{th}} \) largest \( q_{ij} \). Now one assigns to each \( c_j^2 \) an integer \( r_j \) such that the value of \( r_j \) indicates that \( c_j^2 \) is the \( r^{\text{th}} \) largest variance estimate. In other words, the set of \( (n-k) \) numbers \( \{r_j\} \) is a permutation of the integers \( 1, 2, \ldots, n-k \). The test statistic to be used is the Spearman's rank correlation coefficient \( \rho \), which is given by

\[ \rho = 1 - \frac{6}{(n-k)(n-k)^2 - 1} \sum_{i=1}^{n-k} (r_i - \overline{r})^2 \] ... (4.5.5)
For values of \((n-k) \leq 10\), tables are readily available. For longer values, the test statistic is given by

\[
I_p = \left[ \frac{(n-k-2)\hat{\sigma}^2}{\chi^2} \right]^{1/2} \left( \frac{1}{\hat{\sigma}^2} \right). 
\]  
(4.5.6)

Which is distributed as student's t with \((n-k-2)\) degrees of freedom under the null hypothesis.

### 4.5.3 The Kolmogorov Specification Error Test (KOMSET)

Ramsay (1969) pointed the main disadvantage of testing on the squared residuals \(e_i^2\) is that their distribution depends upon the unknown scale factor \(\sigma^2\). The RASET is not affected by lack of knowledge of \(\sigma^2\) as the test is invariant to scale changes. An alternative approach suggested by Ramsay is to consider the ratios of the squared residuals so as to eliminate the effect of the common scale factor. One may consider the distribution of

\[
W_i = \frac{e_i^2}{\hat{\sigma}^2}, \quad j \neq k, 
\]

where \(j, k\) are chosen so that the \(W_i\) are independent.
Under the null hypothesis that the full ideal conditions hold,

\[ \frac{\bar{\epsilon}_j^2}{\bar{\sigma}^2}, \quad j=1,2,\ldots, n-k, \] is distributed as central \( \chi^2 \) with 1 degree of freedom. By assuming \( n-k \) is even, so that one has \( t \) \( \chi^2 \) statistics \( W_i \), Thus, the \( W_i, i=1,2,\ldots, t \), form a set of \( t \) independent variates from the central F distribution with \( (1,1) \) degrees of freedom. Under the alternative hypotheses of different specification errors, \( \frac{\epsilon_j^2}{\sigma^2} \) is distributed as non-central \( \chi^2 \) with one degree of freedom, so that \( W_i, i=1,2,\ldots, t \), form a set of \( t \) independent variates from double noncentral F distributions with \( (1,1) \) degrees of freedom and noncentral parameters \( \lambda_1, \lambda_2 \), where

\[ \lambda_1 = (\eta^T \Lambda_1 \Lambda_1^T \eta) / \sigma^2 \] and

\[ \lambda_2 = (\eta^T \Lambda_2 \Lambda_2^T \eta) / \sigma^2 \] and \( \Lambda_1 \) and \( \Lambda_2 \) are the partition matrices of \( \Lambda \). Ramsay approximated the noncentral distributions with central distributions for the purpose of analysis. He used the Kolomagorov test statistic in his study. The Kolomagorov test is based on the sample cumulative distribution function \( S_n(x) \), which is defined by

\[
S_n(x) = \left\{ \begin{array}{ll}
0 & . x < x_{(1)} \\
\frac{r}{n} & . x_{(r)} \leq x < x_{(r+1)} \\
1 & . x_{(n)} \leq x
\end{array} \right.
\]
Where $x_i$'s are the order statistics

$S_n(x)$ simply the proportion of observations on exceeding $x$.

The Kolomogorov statistic $D_n$ is defined by

$$D_n = \sup_x S_n(x) - G_0(x) \quad \ldots \quad (4.5.8)$$

Where, $G_0(x)$ is the true cumulative distribution function, such that

$$\lim_{n \to \infty} P[S_n(x) - G_0(x)] = 1 \quad \ldots \quad (4.5.9)$$

In simple hypothesis testing case, the distribution of $D_n$ is completely distribution free under null hypothesis, the distribution depending only on $n$. For large $n$, the asymptotic distribution has been obtained and tabulated. As sample size increases $D_n$ becomes more efficient relative to the $\chi^2$ test.

4.5.4 Bartlett's M Specification Error Test (BAMSET)

The tests proposed by Ramsey (1969) namely RESET, RASET and KOMSET do not seem to be very sensitive to simple heteroscedasticity. Such a situation is at least potentially an advantage, for one would like to be able to discriminate between specification errors. If the alternative
hypothesis is that $\tilde{e}_i^2, i = 1, 2, \ldots, n-k$ are distributed as $\sigma_i^2 X_i^2$, one can consider as possible test of heterogeneity, a modification of Bartlett's M test for the nonequality of variances. The residuals $\tilde{e}_i^2, i = 1, 2, \ldots, n-k$ provide one with $(n-k)$ independent estimates of the variances $\sigma_i^2, i = 1, 2, \ldots, n-k$. Bartlett's M test is a modification of Pearson's likelihood ratio test for heterogeneity.

Under the null hypothesis, $H_0$, the $\sigma_i^2$ are all equal and under the alternative hypothesis $H_1$, the $\sigma_i^2$ are all different. Pearson's likelihood ratio statistic $\lambda$ is defined by

$$\lambda = \prod_{i=1}^{R} \frac{S_i^2}{S^2} \cdot n_i^{n_i} \quad \ldots \quad (4.5.10)$$

Where,

$$S_i^2 = n_i^{-1} \sum_{i=1}^{R} (x_{it} - \bar{x}_i)$$

$$S^2 = n^{-1} \sum_{i=1}^{R} n_i \cdot S_i^2 \quad \quad n = \sum_{i=1}^{R} n_i$$

135
Pearson assumed that the observations $x_i$ were drawings from $k$ independent samples of size $n_i$ each and that each sample was drawn from a normal population with mean $\mu_i$ and variance $\sigma_i^2$.

Bartlett's modification was to replace the $n_i$ by the corresponding degrees of freedom

$$v_i = (n_i - 1)$$

and where $v \cdot \sum_{i=1}^{R} v_i$ replaces 'n'.

Thus one has $\lambda^*$ defined

$$\lambda^* = \prod_{i=1}^{R} \left( \frac{S_i^2 / \hat{\sigma}^2_i}{S^2} \right)^{v_i / 2}$$

$$\Rightarrow -2 \log \lambda^* = v \log S^2 + \sum_{i=1}^{R} v_i \log S_i^2$$  \hspace{1cm} (4.5.11)

One has a statistic, which is asymptotically distributed as central $\chi^2$ with $R - 1$ degrees of freedom.

Ramsey (1969) considered the following modification to obtain a statistic which is asymptotically distributed as $\chi^2$. 

136
Let \[ S_i^2 = v_i^{-1} \sum_{j=1}^{v_i} c_{ij}^2 \] \( i = 1, 2, 3 \)

\[ S^2 = v^{-1} \sum_{i=1}^{k} v_i S_i^2 = (n - k)v^{-1} \sum_{i=1}^{n} c_i^2 \]

\[ \sum_{i=1}^{k} v_i = n - k; \ R = 3 \]

Thus, one divides the set of \( n \ k \) residuals into three non-intersecting sub-sets in order to form the statistics \( S_i^2, \ i = 1, 2, 3 \).

The distribution of \(-2 \log \lambda^*\) with respect to \( S_i^2 \), \( S^2 \), \( v_i \), \( v \) and \( R \) is under the null hypothesis, asymptotically distributed as central \( \chi^2 \) with two degrees of freedom. Under the null hypothesis, \( c_i^2 \) is distributed as \( \sigma^2 X_1^2, i = 1, 2, \ldots, n \) \( k \) under the alternative hypothesis that \( c_i^2 \) distributed as \( \sigma^2_i X_i^2 \), the \( (n - k) \sigma_i^2 \) not being all equal, the distributions of \( S_i^2 \) and \( S^2 \) are weighted sums of central \( \chi^2 \) distributions.

In order to test for heterogeneity of variance, one needs several variance estimators. \( R = 3 \) was chosen as a useful compromise. In place of
additional information about the distribution of $e_i^2, i = 1, 2, \ldots, n-k$. the
sub-sets are most easily chosen by setting

$$ S_1^2 = \frac{1}{\eta} \sum_{i=1}^{\eta} e_i^2, \quad r_1 \equiv \binom{n}{k} / 3 $$

$$ S_2^2 = \frac{1}{\eta} \sum_{i=\eta+1}^{2\eta} e_i^2, \quad r_2 \equiv 2\eta $$

$$ S_3^2 = \frac{1}{\eta} \sum_{i=\eta+1}^{\eta+k} e_i^2, \quad \eta \equiv n-k \quad 2\eta $$

Where $r_1$ is the integer obtained from the integer division of $(n-k)$ by $3$.

Bartlett's M-test is not defined for the case when the $e_i^2$ are
distributed as noncentral $\chi^2$. However, the test is sensitive to the
alternative hypothesis of simple heteroscedasticity. The main
disadvantage of the test is that it is not robust to non-normality.
4.6 SMALL SAMPLE PROPERTIES OF TESTS FOR SPECIFICATION ERROR

Consider the classical linear statistical model as

\[ Y = X \beta + \varepsilon \]  \hspace{1cm} (4.6.1)

Where \( Y \) is an \((n \times 1)\) regressand vector,

\( X \) is an \((n \times k)\) nonstochastic regressor matrix of rank \( k \),

\( \beta \) is a \((k \times 1)\) vector of coefficients,

and \( \varepsilon \) is an \((n \times 1)\) vector of independent disturbance terms each distributed normally with mean zero and variance \( \sigma^2 \). For each of the specification error tests suggested by Ramsey (1968), the specification of the null hypothesis is that given in (4.6.1). Under the alternative hypotheses, the specification of the true models which would give rise to specification errors when (4.6.1) is used are:

i. \[ Y = X \beta + Z\gamma + u \]  \hspace{1cm} (4.6.2)

ii. \[ Y = Z\gamma + u \]  \hspace{1cm} (4.6.3)

iii. \[ Y = Z\gamma + W\delta + u \]  \hspace{1cm} (4.6.4)

iv. \[ Y = X \beta + u, \ u \sim N(\phi, \Omega) \]  \hspace{1cm} (4.6.5)

Where, \( Z \) is an \((n \times 1)\) nonstochastic regressor vector,

\( \gamma \) is the corresponding coefficient,

and \( u \) is distributed as \( \sim N(\phi, \Sigma^{-1}) \).
If one uses (4.6.1) to estimate the coefficient vector $\beta$, one has a misspecified model where the misspecification is that of omitting the variable $Z$. If one considers a matrix $X$ whose elements are obtained from the elements of the matrix $Z$ by a non-stochastic nonlinear transformation and one uses (4.6.1) as the regression model instead of (4.6.3), then one would have a misspecified model in which the misspecification is that of incorrect functional form of the regressors. In model (4.6.4), let $Z$ be an $n \times k_1$ nonstochastic full rank matrix of regressors and $\gamma$ the corresponding coefficient vector, let $W$ be an $n \times k_2$ stochastic matrix of regressors of rank $k_2$ in observed samples, $\delta$ the corresponding coefficient vector, $u$ the $n \times 1$ vector of disturbance terms distributed as $N(\Phi, \sigma^2 I)$, and where the elements of each row of $W$ and the corresponding $u_i$, $i = 1, 2, \ldots, n$, are statistically dependent. By defining $X_{n \times k} = (Z; W)$,

$$\beta^\top = \begin{pmatrix} \gamma^\top & \delta^\top \end{pmatrix},$$

where $\beta$ is $(k \times 1)$ vector, $k = k_1 + k_2$, if one uses (4.6.1) as the regression model instead of the model (4.6.4) then one has a misspecified model in which the misspecification is denoted the 'simultaneous equation problem'. In the model (4.6.5), all the
specifications made for model (4.6.1) are correct except that \( \Omega \) is assumed to be diagonal, but with unequal elements on the diagonal. If one uses a model such as (4.6.1) the specification error of heteroscedasticity is said to have been made.

Ramsay (1969) discussed the four specification tests by using Theil’s best linear unbiased scalar (BLUS) residual vector \( \tilde{e} \) given by

\[
\tilde{e} = A' Y \text{ where } A \text{ is } n \times n - k \text{ matrix, which has the properties :}
\]

\[
A' X = \Phi, \quad A' A = I_{n-k}, \quad A A' = \left[ I - X' (X' X)^{-1} X' \right] = M.
\]

For each of the above tests, the null hypothesis can be expressed in terms of the Theil residuals as \( H_0 : \tilde{e} \sim N(\Phi, \sigma^2 I_{n-k}) \)

The various alternative hypotheses can be expressed in terms of distributional properties of the Theil’s residuals as

\[
H_{11} : \tilde{e} \sim N(A' \eta, \tilde{\sigma}^2 I), \text{ where } \tilde{\sigma}^2 \text{ is the variance of } \tilde{e}_i
\]

and \( A' \eta = \alpha_0 + \alpha_1 q_1 + \alpha_2 q_2 + \ldots \)
Here, \( q_j = A \hat{y}^{(j+1)} \), \( j = 1, 2, \ldots \) are the \( n - k \times 1 \) vectors and \( \hat{y}^{(j+1)} \) is the \( (n \times 1) \) vector obtained by raising each element of the estimated vector of the conditional mean of the regressand to the \( (j + 1) \) power.

\[
H_{12} : \tilde{e} \sim N(\Phi, \Theta), \Theta \text{ a diagonal positive definite matrix.}
\]

The alternative hypothesis \( H_{11} \) refers to the errors of omitted variables, incorrect functional form of the regressors, and the simultaneous equation problem. \( H_{12} \) refers to the error of heteroscedasticity. In constrast, the distributional properties of the OLS residuals under the various hypotheses are given by

\[
H_0 : e \sim N(\Phi, \sigma^2 M)
\]

\[
H_{11} : e \sim N(M\eta, \sigma^2 M);
\]

\[
H_{12} : e \sim N(\Phi, \psi)
\]

Where \( M \) and \( \psi = M \Omega M \) are positive semi-definite matrices so that in each case \( e \) is distributed as singular normal.

Ramsey and Gilbert (1972) have used five specified models in the Monte Carlo experiments to generate statistically independent observations on the regressand.
The models are given by:

**Model I**: Hypothesis of no specification error

**Model II**: Misspecification of omitted variables

**Model III**: Misspecification of functional form of the regressors

**Model IV**: Misspecification in a simultaneous equation system

**Model V**: Misspecification of heteroscedasticity

In each case, the null hypothesis \( H_0 \) that \( \hat{u} = A' \hat{y} \) is distributed as normal with null mean vector and scalar covariance matrix is tested against the alternative hypothesis that \( H_{11} \) or \( H_{12} \) is true. Ramsey's RESET is a test for \( H_0 \) against \( H_{11} \); BAMSET is a test for \( H_0 \) against \( H_{12} \); and RASET is a test against the union of \( H_{11} \) and \( H_{12} \).

4.7 USE OF AN F-STATISTIC IN FORWARD SELECTION, BACKWARD ELIMINATION AND STEPWISE REGRESSION PROCEDURES

Consider the data for the aspect of multiple regression as in the form \( (Y_r, X_{r1}, X_{r2}, ..., X_{rk}) \), \( r = 1, 2, ..., n > k + 1 \). Each vector contains the information from the \( i^{th} \) unit and it is assumed that a linear equation

143
of a subset of the X's can be determined that will predict Y. Now, the problem is to select the best subset of the X's for the multiple regression.

Let \( Y^1 = (Y_1, Y_2, \ldots, Y_n) \) a (1xn) vector;

\[
X^1_j = (X_{1j} - \overline{X}_j, X_{2j} - \overline{X}_j, \ldots, X_{nj} - \overline{X}_j) \text{ a (1 x n) vector;}
\]

\[
\overline{X}_j = \sum_{r=1}^{n} X_{rj} / n;
\]

\( X_1 \) a \( n \times k - p \) matrix whose columns are \( k - p \) of the \( X_j \); these will be the \( k - p \) independent variables already included in the prediction equation; \( X_2 \) an \( X_j \) that is not in \( X_1 \);

\( X_2 \) a \( n \times p - 1 \) matrix whose columns are the \( X_j \) that are not in \( X_1 \) and not including \( X_p \).

The Forward Selection Procedure keeps the constant term in the prediction equation and then operates in the following manner:
p is set equal to k (which makes $X_2$ a $n \times k - 1$ matrix) and the k ratios (n-2-0) [SSR_{oi}/SSE_{oi}],

Where

$$SSR_{oi} = \left(\begin{array}{c} y_i \\ X_i \end{array}\right)^T X_i$$

and

$$SSE_{oi} = \sum_{r=1}^{n} (y_r - \bar{y})^2 - SSR_{oi}$$

$$= \frac{1}{n} \left( Y^T Y - \frac{1}{n} Y - SSR_{oi} \right),$$

$\frac{1}{n}$ is an (nxn) matrix whose elements are all $\frac{1}{n}$ are compared
to a value of F distribution with $(1, n - 2)$ degrees of freedom. If the
largest of these ratios is greater than the value of $F_{1, n-2}$, the $X_i$ generating
that ratio becomes the first vector in $X_1$. The next step is compute the
(k-1) ratios,

$$(n - 2 - 1)[SSR_{ii}/SSE_{ii}],$$

Where,

$$SSR_{ii} = \left[ Y^T \left( I - X_1 X_1^T \right) X_1 \right] X_1 \frac{1}{n} \left[ X_1^T \left( I - X_1 X_1^T \right) X_1 \right] X_1$$

and

$$SSE_{ii} = \left[ Y^T \left( I - X_1 X_1^T \right) X_1 \right] X_1 \frac{1}{n} \left( Y - SSR_{ii} \right),$$

It should be noted that SSR_{ii} is the sum of squares of regression
which will be increased by adding $X_i$ to the independent variable in $X_1$. If
the largest ratio is greater than the value of $F_{k-1, n-1}$, then this $X_i$ is added to
$X_1$ and the process repeated. Each time a vector is added to $X_1$ and the denominator degrees of freedom is reduced by one. When the maximum ratio at any one step does not exceed the F-value the process stops and the $X$'s in $X_1$ will be the best subset that will be used to predict $Y$. The coefficient of these independent variables are given by $(X_1'X_1)^{-1}X_1'Y$.

Pope and Webster (1972) have pointed out the pseudoness of the F-statistic used in stepwise procedures for determining the independent variables to be used in a linear prediction equation. They considered the linear model in the form,

$$Y = \text{l}{\beta}_0 + X_1{\beta}_1 + X_2{\beta}_2 + \ldots (4.7.1)$$

Where $\text{l}$ is (nx1) vector of 1's;

$\beta_0$ is (1x1); $\beta_1$ is $(k-p \times 1)$; $\beta_i$ is (1x1) and $\beta_2$ is $(p-1 \times 1)$ vectors of unknown parameters; $\epsilon$ is (nx1) vector of independent random variables from a normal distribution with mean zero and variance $\sigma^2$.

At any given step in the forward selection procedure, say when $(k-p)$ independent variables have been included and one of the remaining $p$ is being considered, the increase in the sum of squares of regression is given by,

146
\[ SSR_{k-p_i} = Y' (X_1 X_i) \left[ (X_1 X_i)' (X_1 X_i) \right]^{-1} (X_1 X_i)' Y \]

\[ = Y' Z_i (Z_i' Z_i)^{-1} Z_i' Y \]  \hspace{1cm} \text{... (4.7.2)}

Where \[ Z_i = X_i - X_1 (X_1' X_1)^{-1} X_1' X_i \]

Here, SSR_{k-p} is referred to as the sum of squares of regression due to \( X_i \) adjusted for \( X_1 \). The sum of squares of residuals is given by,

\[ SSE_{k-p_i} = Y' \left\{ I - \left( \frac{1}{n} \right) X_1 (X_1' X_1)^{-1} X_1' - Z_i (Z_i' Z_i)^{-1} Z_i \right\} Y \]  \hspace{1cm} \text{... (4.7.3)}

Since, SSR_{k-p_i} are independent and from the distribution of quadratic form, the ratio

\[ R_{k-p_i} = (n - 2 - k - p) \frac{SSR_{k-p_i}}{SSE_{k-p_i}} \]  \hspace{1cm} \text{... (4.7.4)}

follows a noncentral F-distribution.

Under the condition \( \beta_2 = 0 \) and \( H_0: \beta_i = 0 \), \( R_{k-p_i} \) follows central F-distribution.

147
Stepwise Regression and Backward Elimination Procedure

The stepwise regression procedure is a modification of the forward selection procedure. A step is added in which, after each independent variable enters the included group, each of the included group is then re-investigated to see whether it is still worth including. In other words, an included independent variable may be discarded later if at any future step a subset of the included independent variables contains most of its predictive value.

The stepwise procedure starts off by choosing an equation containing the single best $X$ variable and then attempts to build up with subsequent additions of $X$'s one at a time as long as these additions are worthwhile. The order of addition is determined by using the partial $F$-test values to select which variable should enter next. The highest partial $F$-value is compared to a (selected or default) $F$-to-enter value. After a variable has been added, the equation is examined to see if any variable should be deleted.

Under this procedure, first we select the $Z$ most correlated with $Y$ (suppose it is $Z_1$) and find the first-order, linear regression equation
\[ \hat{y} = f(Z_1) \]. We check if this variable is significant. If it is not, we quit and adopt the model \( y = \hat{y} \) as best; otherwise we search for the second predictor variable to enter regression.

We examine the partial F-values of all the predictor variables not in regression. The \( Z_j \) with the highest such value (suppose this is \( Z_2 \)) is now selected and a second regression equation

\[ \hat{y} = f(Z_1, Z_2) \] is fitted.

The overall regression is checked for significance, the improvement in the \( R^2 \) value is noted, and the partial F-values for both variables now in the equation (not just the one most recently entered) are examined. The lower of these two partial F’s is then compared with an appropriate F percentage point, F-to-remove and the corresponding predictor variable is retained in the equation or rejected according to whether the test is significant or not significant.

This testing of the "least useful predictor currently in the equation" is carried out at every stage of the stepwise procedure. A predictor that may have been the best entry candidate at an earlier stage
may at a later stage, be superfluous because of the relationship between it and other variables now in the regression. To check on this, the partial F criterion for each variable in the regression at any stage of calculation is evaluated, and the lowest of these partial F-values (which may be associated with the most recent entrant or with a previous entrant) is then compared with a pre-selected percentage point of the appropriate F-distribution or a corresponding default F-value. This provides a judgement on the contribution of the least valuable variable in the regression at the stage, treated as though it had been the most recent variable entered irrespective of its actual point of entry into the model. If the tested variable provides a nonsignificant contribution, it is removed from the model and the appropriate fitted regression equation is then computed for all the remaining variables still in the model.

The best of the variables not currently in the model (i.e., the one whose partial correlation with Y, given the predictors already in the equation, is greatest) is then checked to see if it passes the partial F entry test. If it passes, it is entered, and we return to checking all the partial F's for variables in. If it fails, a further removal is attempted.
Eventually (unless the α-levels for entry and removal are badly chosen to provide a cycling effect), when no variables in the current equation can be removed and the next best candidate variable cannot hold its place in the equation, the process stops. As each variable is entered into the regression, its effect on $R^2$, the square of the multiple correlation coefficient, is usually recorded and printed.

The backward elimination procedure starts with every independent variable being in the included set and then goes through a process of throwing the independent variables out one-by-one. The process consists of construction a set of statistics similar to the $F$, and then considering how small the minimum $F$ is.

The backward elimination method is also an economical procedure. It tries to examine only the “best” regressions containing a certain number of variables. The basic steps in the procedure are these:

1. A regression equation containing all variables is computed.
2. The partial $F$-test value is calculated for every predictor variable treated as through it were the last variable to enter the regression equation.
3. The lowest partial F-test value, say, $F_L$, is compared with a preselected or default significance level, say, $F_0$.

a. If $F_L < F_0$, remove the variable $Z_l$, which gave rise to $F_L$, from consideration and recompute the regression equation as calculated.

b. If $F_L > F_0$, adopt the regression equation as calculated.

4.6 HOCKING’S CRITERIA FOR SELECTION OF A SUBSET REGRESSION

Consider the standard linear model as

$$Y = X\beta + \epsilon$$

where, $Y$ is nx1; $X$ is nxk; $\beta$ is kx1; and $\epsilon$ is nx1 matrices.

Let $D_p$ be a k x p matrix consists of zeroes and ones such that $XD_p$ contains only those p columns of $X$ to be included in the subset (p<k). The index r is used to specify which subset of size p is being considered and will not be included in the remaining development. More general $D$ matrices could be considered to allow for linear combinations of the variables.
If we specify the subset by \( D_p \) then the corresponding set of regression coefficients \( \hat{\beta}_p = D_p^\perp \beta \) are estimated by

\[
\hat{\beta}_p = (D_p^\perp X^\perp XD_p)^{-1} D_p^\perp X^\perp Y \quad \ldots \quad (4.8.2)
\]

If \( y \) denotes the actual response for a particular row vector \( x \) in the original model (4.8.1), we have

\[
y = x \beta
\]

and the corresponding predicting value of response is given by

\[
\hat{y}_p = x D_p \hat{\beta}_p \quad \ldots \quad (4.8.3)
\]

The mean squared error of the prediction is given by

\[
\text{MSEP}(\hat{y}_p) = E[\hat{y}_p - y]^2 = \text{VARP}(\hat{y}_p) + [E(\hat{y}_p) X \beta]^2 \quad \ldots \quad (4.8.4)
\]

where the prediction variance is given by

\[
\text{VARP}(\hat{y}_p) = \sigma^2 + x D_p (I_p - x D_p^\perp X D_p^\perp)^{-1} D_p^\perp X^\perp \sigma^2 \quad \ldots \quad (4.8.5)
\]

and

\[
E(\hat{y}_p) = x D_p (D_p^\perp X^\perp XD_p)^{-1} D_p^\perp X^\perp X \beta
\]
Allen (1971) suggested that for predicting at a given row vector \( x \), the subset of regressors which minimizes the mean squared error prediction (4.8.4) should be used.

Hocking (1972) has reviewed a number of criteria suggested by Mallows (1966), Gorman and Toman (1966), Rothman (1968) and Allen (1971) for selecting subset of regressors for the prediction.

Let \( x_i \) denotes the \( i^{th} \) row of \( X \) and \( y_p \) the predicted value according to (4.8.5). The various criteria are given by

i. Rothman (1968) Criterion

\[
J_P = \frac{1}{n} \sum_{i=1}^{n} VARP(\hat{y}_{pi}) \frac{n}{n - p} \sigma^2 \quad \text{... (4.8.6)}
\]

Neglecting the divisor \( n \) and substituting an estimate for \( \sigma^2 \), we obtain the test statistic as

\[
J_P = \frac{n + p}{n - p} c_p c_p \quad \text{... (4.8.7)}
\]
where $c_p \cdot c_p$ is the residual sum of squares for the $p$-variate regression.

Here, Rothman’s statistic $J_p$ uses (4.8.5) rather than (4.8.4), thus neglecting the bias term.

ii. Gorman And Toman (1966) Criterion

\[
\Gamma_p = \frac{1}{\sigma^2} \sum_{i=1}^{n} \text{MSEP}(\hat{y}_p^i)
\]

\[
= \frac{1}{\sigma^2} \left[ E(\hat{y}_p) - X\beta \right] \left[ E(\hat{y}_p) - X\beta \right]^\top, \quad (n + p)
\]

where \( \hat{y}_p = XD_c \hat{\beta}_c \).

By simplification, we obtain

\[
\Gamma_p = \frac{E(c_p^c_p)}{\sigma^2} + 2p
\]

and hence, the test statistic is given by

\[
\hat{\Gamma}_p = \frac{c_p^c_p}{\hat{\sigma}^2} + 2p
\]

... (4.8.8)

Where $\hat{\sigma}^2 = \frac{(c \cdot c)}{n-k}$, $c \cdot c$ is the residual sum of squares for the $k$-variable linear regression.
iii. Mallows (1966) Criterion

Mallows recommends the statistic as

\[ \hat{C}_p = \frac{(c_p^1 c_p)}{\hat{\sigma}^2} + 2p - n \]

\[ = (n - p) \frac{(c_p^1 c_p) \sqrt{(n - p)}}{\hat{\sigma}^2} \cdot \hat{p} \quad \ldots (4.8.9) \]

If the \( p \)-subset adequately describes the data, the quantity in parentheses be near zero and \( \hat{C}_p \simeq p \).

The \( \hat{C}_p \) criterion may often recommend a large value of \( 'p' \) than other criteria.

In addition to \( \hat{J}_p \) and \( \hat{C}_p \), other users recommend the various simple functions of \( c_p^1 c_p \) such as the Residual Mean Square, the Multiple \( R^2 \) and the adjusted \( R^2 \) denoted by \( \bar{R}^2 \).
Allen (1971) recommends the prediction sum of squares PRESS, which is given by

\[
\text{PRESS}_p = \sum_{i=1}^{n} \left[ y_i - \hat{y}_{p(i)} \right]^2 \tag{4.8.10}
\]

PRESS differs from \( e_i^p c_i \) by defining \( \hat{y}_{p(i)} \) as in (4.8.3) except that the \( i^{th} \) observation has been deleted before computing (4.8.2).

We have,

\[
\text{PRESS}_p = \sum_{i=1}^{n} \left[ y_i - \hat{y}_{pi} \right]^2 \tag{4.8.11}
\]

Where

\[
\hat{y}_{pi} = x_i \hat{D}_p \hat{\beta}_p
\]

and \( Q_i = x_i D_p (D_p^t X^t X D_p)^{-1} D_p x_i \)

Thus, \( \text{PRESS}_p \) is weighted sum of squares of residuals where the weights are related to the variance of \( \hat{y}_{pi} \).

ie., \( \text{VAR} (\hat{y}_{pi}) = Q_i \sigma^2 \)

The various functions of PRESS, analogous to (4.8.7) and (4.8.9) may be frequently used in practice.
4.9 CONSEQUENCES OF INCORRECT MODEL SPECIFICATION

Hocking (1976) provided a review of the concepts and methods associated with variable selection in linear regression models. Variable selection reduces the number of independent variables in the final linear regression equation. There are a variety of practical and economical reasons for reducing the number of independent variables. Variable deletion may be desirable in terms of the statistical properties of the parameter estimates and the estimate of the final linear regression equation. Hocking (1976) provided a brief review of the consequences of incorrectly specifying the model either in terms of retaining extraneous variables or deleting relevant variables. The properties described here are dependent on the assumption that the subset of variables under consideration has been selected without reference to the data.

Consider the classical linear regression model in matrix notation as

\[ Y_{n \times 1} = X_{n \times k} \beta_{k \times 1} + \epsilon_{n \times 1} \]  \hspace{1cm} (4.9.1)

158
In the variable selection problem, let \( r \) denote the number of terms which are deleted from model (4.9.1), that is the number of regression coefficients which are set to zero. The number of terms which are retained in the final equation will be denoted by \( p = k - r \).

Let the model (4.9.1) be written in the form as

\[
Y = X_p \beta_p + X_r \beta_r + \epsilon
\]  ... (4.9.2)

Where the \( X \) matrix has been partitioned into \( X_p \) of dimension \( n \times p \) and \( X_r \) of dimension \( n \times r \). The \( \beta \)-vector is partitioned conformably.

If the variables in \( X_r \) are deleted from the model, then we have the OLS estimator of \( \beta_p \) as

\[
\hat{\beta}_p = \left( X_p^\top X_p \right)^{-1} X_p^\top Y
\]  ... (4.9.3)

Also, the OLS estimator of \( \beta \) vector is given by

\[
\hat{\beta} = \left( X^\top X \right)^{-1} X^\top Y
\]  ... (4.9.4)
Further, the corresponding residual mean sum of squares are given by

\begin{align}
\hat{\sigma}_p^2 &= Y^\dagger \left( I - X_p \left[ X_p^\dagger X_p \right] \right) X_p^\dagger Y \sqrt{\frac{n - p}{n - k}} \quad \ldots \quad (4.9.5) \\
\sigma^2 &= Y^\dagger \left( I - X \left( X^\dagger X \right)^{-1} X^\dagger \right) Y \sqrt{\frac{n - k}{n - k}} \quad \ldots \quad (4.9.6)
\end{align}

If the model (4.9.2) is correct, the properties of \( \hat{\beta} \) and \( \hat{\sigma}^2 \) as estimates of \( \beta \) and \( \sigma^2 \) are well known from general linear model theory. In particular they are minimum variance unbiased estimators with

\[ \hat{\beta} \sim N \left( \beta \left( X^\dagger X \right)^{-1} \sigma^2 \right) \text{ and } \hat{\sigma}^2 \sim \chi^2_{n-k} \]

Let

\[ A = \left[ \begin{array}{c}
X_p^\dagger X_p \\
X_p^\dagger X_r
\end{array} \right] \]

Then we have

\[ E \left[ \hat{\beta}_p \right] = \beta_p + A \beta_r, \quad \ldots \quad (4.9.7) \]

and

\[ \text{Var} \left[ \hat{\beta}_p \right] = \left( X_p^\dagger X_p \right)^{-1} \sigma^2 \quad \ldots \quad (4.9.8) \]

Also \( \bar{\beta}_p \) follows multivariate normal distribution.

The Mean Squared Error (MSE) of \( \hat{\beta}_p \) is given by

\[ \text{MSE} \left[ \hat{\beta}_p \right] = E \left[ \hat{\beta}_p - \beta_p \right] \left[ \beta_p - \beta_p \right]^\dagger \]
\[
\begin{align*}
&= \left( x_p^\top x_p \right)^\top \sigma^2 + A \beta_r \beta_r^\top \Lambda^\top \ldots (4.9.9) \\
\text{Also, } & (n - p) \tilde{\sigma}_p^2 \text{ follows noncentral chi-squared distribution with}
\end{align*}
\[
E \left| \tilde{\sigma}_p^2 \right| = \sigma^2 + \beta_r^\top x_r \left( I - x_p (x_p^\top x_p)^\top x_p \right) x_r \beta_r (n - p) \ldots (4.9.10)
\]

**PROPERTIES OF \( \tilde{\beta}_r \)**

(i) \( \tilde{\beta}_p \) is generally biased, interesting exceptional cases being

(a) \( \beta_r = 0 \) and  
(b) \( x_p^\top x_r = 0; \)

(ii) The matrix \( \left[ \text{Var}(\tilde{\beta}_p) - \text{Var}(\beta_p) \right] \) is positive semi - definite matrix,

(iii) If the matrix \( \text{Var}(\tilde{\beta}_r) - \beta_r \beta_r^\top \) is positive semi-definite, then the matrix \( \left[ \text{Var}(\tilde{\beta}_p) - \text{MSE}(\tilde{\beta}_p) \right] \) is positive semi - definite.

(iv) \( \tilde{\sigma}_p^2 \) is generally biased upward.

**4.10 A CRITERION FOR STEPWISE REGRESSION**

The problem of selecting a subset of independent variables in linear regression has led to various 'subset selection' procedures. In common use, the stepwise method which, at a given step, selects the
independent variable which maximizes the squared partial correlation coefficient with the dependent variable, given the variables already selected. The process 'stops' whenever the sample partial correlation is 'non-significant' as shown by the standard F-test.

An alternative method suggested by Bendel and Afifi (1976) which involves testing the hypothesis that the mean squared error of prediction does not decrease from one step to the next.

Suppose \((y_i, x_{i1}, x_{i2}, \ldots, x_{ik}) \), \(i = 1, 2, \ldots, n\), is a random sample of \(n\) observations from a \((k+1)\) variate normal distribution, where \(y\) is the dependent variable. Denoting the sample regression function based on \(q\) of the \(k\)-possible independent variables by

\[
\hat{y}_{(q)}(X) = \hat{y} + \hat{\beta}(q) \left[ X_{(q)} \ X_{(q)}^\top \right]
\]  

\((4.10.1)\)

Where \(\hat{\beta}(q)\) is the vector of least squares estimates of the regression coefficients and \(X_{(q)}\) is the vector of the \(q\) independent variables \(X_{k1}, \ldots, X_{kq}\) which have been chosen by some subset selection procedure.
By denoting \( \hat{y}(q)(x_0) \) as an estimate of \( Y_0 \) the unconditional mean squared error of prediction is defined as

\[
\text{UMSE}_{\psi} = E \left[ \hat{y}(q)(x_0) - y_0 \right]^2 \quad (4.10.2)
\]

Where the expectation is taken over all random variables including the 'new' observation \( (Y_0, X_0) \) as well as the independent variables of the regression sample. The criterion UMSE was used by Stein (1960) and Afifi and Elashoff (1967), to establish relative efficiencies of various missing value techniques.

When all \( k \) independent variables are in the regression equation, Stein (1960) has shown that

\[
\text{UMSE}_{\text{ek}} = \frac{n^2}{n(n-k-2)} \sigma_y^2 \cdot x(k) \quad (4.10.3)
\]

Where \( n > k + 2 \)

Here, \( \sigma_y^2 \cdot x(k) \) = the conditional variance of \( y \) given \( k \) independent variables in the regression equation.

Bandel and Afifi (1976) have discussed a test procedure, which is motivated by the desire to employ a stepwise test which indicates the
value of q for which UMSE\(_{(q)}\) is a minimum. The null hypothesis to be tested here is that

\[ H_0 : \text{UMSE\(_{(q)}\)} \leq \text{UMSE\(_{(q+1)}\)} \]

This procedure produces one or more subsets of variables which are not significantly poorer than the regression equation based on all of the independent variables.

4.11 A TEST FOR FUNCTIONAL MISSPECIFICATION IN REGRESSION ANALYSIS

An important problem which arises frequently in regression analysis is that of testing whether the functional form employed for the k-regressors \(X_1, X_2, \ldots, X_k\), is appropriate. Harvey and Collier (1977) have suggested a test for the functional form of only one of the variables, say the \(k^{th}\) is in doubt. There are two basic approaches to test for functional misspecification. The first is to run a new regression in which an extra term in \(X_k^2\) is included. A test of significance on the co-efficient 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, has considerable appeal and is widely used.
A number of procedures based on residuals have been proposed for testing against functional misspecification. Perhaps the most widely used are the tests which have been developed essentially in connection with serial correlation. The procedure proposed by Durbin and Watson (1971) uses OLS residuals. Alternatively Theil's (1971) BLUS residuals may, because of their independence under the null hypothesis, be used directly in the Von Neumann Ratio. The Recursive residuals, described in Brown, Durbin and Evans (1975) have similar properties to BLUS residuals and so may also be employed to give an exact test with the Von Neumann Ratio.

Consider the general linear regression model as

\[ Y_j = X_j' \beta + \epsilon_j, \ j = 1, 2, ..., n \]  \hspace{1cm} (4.11.1)

Where \( X_j \) is \((k \times 1)\) vector of (fixed) observations on the independent variables;

\( \beta \) is a \((k \times 1)\) vector of coefficients;

\( Y_j \) is the \( j^{th} \) observation on the dependent variable;

\( \epsilon_j \) is a disturbance term.
These disturbances are assumed to be normally and independently distributed with zero expectation and constant variance.

The \((n-k)\) recursive residuals are given by

\[
e_{j}^{***} = \frac{y_{j} - X_{j}^{i}b_{j-1}}{1 + X_{j}^{i}\{X_{j-1}^{i}X_{j-1}^{i}\} X_{j}}, \quad j = k+1, \ldots, n \quad (4.11.2)
\]

Where \(b_{j}\) is the OLS estimate of \(\beta\) obtained from the first \(j\) observations, and \(X_{j}\) is a \(j \times k\) matrix of full rank consisting of the first \(j\) steps of observations on the independent variables. These residuals are easily computed since simple recursive formulae exist for updating \(b_{j}\) and \(X_{j}^{i}X_{j}\).

Under the null hypothesis that model (4.11.1) holds, the recursive residuals have the same properties as the true disturbances.

If \(\bar{c}^{***}\) is the arithmetic mean of the recursive residuals, the statistic

\[
\psi = (n-k-1)^{-\frac{1}{2}} \sum_{j = k+1}^{n} (c_{j}^{***} - \bar{c}^{***})^{\gamma} \left( (n-k)^{1/2} \sum_{j = k+1}^{n} c_{j}^{***} \right)^{-\gamma/2} \quad (4.11.3)
\]

166
follows a student's t-distribution with \((n-k-1)\) degrees of freedom under the null hypothesis.

This \(\psi\)-test consists simply of arranging the observations in ascending (or descending) order according to the variable which is to be tested for functional misspecification.

4.12 A GROUPING TEST FOR MISSPECIFICATION

Consider the model

\[
Y = X\beta + \epsilon \quad \ldots (4.12.1)
\]

Such that \(\epsilon \sim N(0, \sigma^2 I_n)\)

Where \(Y\) is an \((n \times 1)\) vector of observations on the dependent variable;

\(X\) is an \((n \times k)\) matrix of observations on \(k\) regressors;

\(\beta\) is a \((k \times 1)\) vector of parameters;

and \(\epsilon\) is an \((n \times 1)\) vector of disturbances.
If $X$ is of rank $k$, then the ungrouped OLS estimator of $\beta$ is given by

$$\hat{\beta} = (X^\prime X)^{-1} X^\prime Y$$

Let $G$ be an $(m \times n)$ grouping matrix of rank 'm'. We have

$$GY = G X \beta + G \varepsilon$$

and $G \varepsilon \sim N(0, \sigma^2 GG^\prime)$

If $X$ is of rank $k$, then the grouped estimator of $\beta$ is given by

$$\tilde{\beta}_G = \{X^\prime \ G^\prime \left( G G^\prime \right)^{-1} G \ \ G^\prime \left( G G^\prime \right)^{-1} G_Y \}$$

Polinsky (1977) discussed the estimation problems in the context of $\tilde{\beta}_G$. Let $F$ be an $(n-m) \times n$ matrix satisfying $F G^\prime = 0$ and

$$FF^\prime = I_{n-m}, \text{ and let } G_* = \left[G G^\prime \right]^{1/2} G.$$  

The model (4.12.1) is equivalent to

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \end{bmatrix} \begin{bmatrix} \beta' \\ \delta' \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$
\[ \epsilon_1 \sim N(0, \sigma^2 I_{n-m}) \]

\[ \epsilon_2 \sim N(0, \sigma^2_m) \]

When \( \delta = 0 \) where \( \epsilon_1 \epsilon_2^T = 0 \),

\[ Y_1 = F Y, X_1 = F X, \epsilon_1 = F \epsilon, \]

\[ Y_2 = G Y, X_2 = G X, \text{ and } \epsilon_2 = G \epsilon \]

Farebrother (1979) suggested a grouping test statistic for
misspecification of the model as

\[ F = \frac{Q^T \sigma^2 \epsilon}{(Q - Q^T) / (n - 2k)} \]

Which follows \( F_{k, (n-2k)} \) (4.12.3)

Under the null hypothesis \( H_0 : \delta = 0 \)

Where,

\[ Q = \left( \hat{\beta}_{\epsilon \mid \epsilon} - \hat{\beta} \right)^T X G \left( G^T G \right)^{-1} G X \left( X^T X \right)^{-1} \left( X^T \hat{\beta} - \hat{\beta} \right) \]

and \( Q = (y - X\hat{\beta})^T (y - X\hat{\beta}) \)

The rejection of the null hypothesis implies the model is
misspecified.
4.13 A SIMPLE TEST FOR SPECIFICATION ERROR

A major function of econometrics is to test the validity of models put forward by economic theory. Most techniques for hypothesis testing in econometrics, however, simply allow one to test restrictions on a model more general than the one being valid.

Consider the problem of testing nonnested regression models to form compound model from two or more alternative hypothesis and test the restrictions implied by only one of them being true. Let both the null and alternative hypotheses be linear models as follows:

\[ H_0 : \quad Y_i = X_i \beta_1 + W_i \beta_2 + \epsilon_i \]

\[ H_1 : \quad Y_i = Z_i \gamma_1 + W_i \gamma_2 + \epsilon_i \]

Where  \( Y_i \) is the \( i^{th} \) observation on the dependent variable; \( X_i \) and \( Z_i \) denote the vectors of observations on regressors which are unique to \( H_0 \) and \( H_1 \) respectively.

\( W_i \) denotes the vector of observations on the regressors which are common to both hypotheses consider the compound model.

\[ Y_i = (1 - \alpha) X_i \beta_1 + W_i [(1 - \alpha) \beta_2 + \alpha \gamma_2] + \alpha Z_i \gamma_1 + \epsilon_i \quad \ldots \quad (4.13.1) \]
and test whether $\alpha \gamma_1 = 0$ using an F-test, a procedure suggested by Atkinson (1970) and discussed by Pesaran (1974). A simple alternative test was suggested by Davidson and Kinnon (1981) which involves estimating the compound model,

$$Y_i = (1 - \alpha) (X_i \beta_1 + W_i \beta_2) + \alpha (W_i \hat{y}_i + Z_i \hat{y}_i) + \epsilon_i \quad \ldots \quad (4.13.2)$$

and this can be rewritten as

$$Y_i = X_i \beta_1^* + W_i (\beta_2^* + \alpha \hat{y}_2) + \alpha Z_i \hat{y}_1 + \epsilon_i \quad \ldots \quad (4.13.3)$$

Where, 

$$\beta_1^* = (1 - \alpha) \beta_1$$

and 

$$\beta_2^* = (1 - \alpha) \beta_2$$

Davidson and Kinnon (1981) have extended their test for model specification to the nonlinear models.

### 4.14 AMEMIYA’S CRITERION FOR SELECTION OF REGRESSORS

Amemiya (1980) proposed a model selection criterion for nested models based on minimizing the unconditional Mean Square Prediction Error (MSPE) as the risk function. He compared this prediction
criterion with Mallows (1973) Cp criterion. He has shown that, in the context of fixed regressors, prediction criterion and Mallows criterion differ in the amount of information they require. Mallow’s Cp has the practical disadvantage that it requires the knowledge of the error variance of the true model.

On the other hand, Amemiya’s (1980) prediction criterion has the attractive feature that it utilizes information only from the sample.

Consider the true model as

\[ Y_{nx1} = X_{1nxk1} \beta_{1_{k1x1}} + X_{2_{nxk1}} \beta_{2_{k1x2}} + \epsilon_{nx1} \quad \ldots (4.14.1) \]

Where

\[ X = [X_1 \quad X_2] \] is a random sample from

\[ N(0, \Sigma_{xx}) ; \]

\[ \Sigma_{xx} = \begin{bmatrix} \Sigma_{X1X1} & \Sigma_{X1X2} \\ \Sigma_{X2X1} & \Sigma_{X2X2} \end{bmatrix} \]

\[ \epsilon \] is a vector of unobserved disturbances distributed \( N(0, \sigma^2) \) independent of \( X \); and \( n > (k_1 + k_2) \).
For fixed regressors, Amemiya (1980) derived Mallow's criterion as

\[ M_C = \frac{2k_1}{n} \frac{Y'MY}{n-k_1-k_2} + \frac{Y'M_1Y}{n} \quad \text{... (4.14.2)} \]

Where,

\[ M = I - X (X'X)^{-1} X' \]

\[ M_1 = I - X_1 (X_1'X_1)^{-1} X_1' \]

and

\[ \frac{Y'MY}{n-k_1-k_2} = \sigma_i^2 \quad \text{... (4.14.3)} \]

In the case of stochastic regressors, Mallow's criterion is given by

\[ M_C = \frac{n-1}{n-k_1-1} \frac{Y'MY}{n-k_1} \quad \text{(4.14.4)} \]

Under the null hypothesis \( H_0 : \beta_1 = 0 \) in (4.14.1) Kinal and Lahiri (1984) derived the test statistic for the selection of regressors as

\[ F = \frac{Y'[M_1 - M]Y}{Y'M_1Y / n-k_1-k_2} / F_{(k_2, n-k_1-k_2)} \quad \text{(4.14.5)} \]

They have shown that \( H_0 \) is to be accepted if

\[ F < \frac{2(n-k_1-k_2-1)}{n-k_1-k_2-1} = 1 + \frac{n-k}{n-k_1-k_2-1} \quad \text{... (4.14.6)} \]
4.15 SPECIFICATION PRE-TEST ESTIMATOR

Consider the classical linear regression model as

\[ Y = X \beta + \varepsilon \]

or

\[ Y = X_1 \beta_1 + X_2 \beta_2 + \varepsilon \] \hspace{1cm} (4.15.1)

Where \( Y \) is a \((n \times 1)\) vector; \( X, X_1, X_2 \) are non-stochastic design matrices of the respective sizes \((n \times k)\), \((n \times k_1)\) and \((n \times k_2)\) where \( k = k_1 + k_2 \); \( \beta, \beta_1 \) and \( \beta_2 \) are \((k \times 1)\), \((k_1 \times 1)\) and \((k_2 \times 1)\) vectors of unknown parameters, \( \varepsilon \) is a \(n\)-dimensional disturbance which is normally distributed with zero mean and covariance matrix \( \sigma^2 I \).

Further more, it is assumed that the design matrix \( X \) is such that rank of \( \begin{pmatrix} X_1' & X_2' \end{pmatrix} \) is \( k_1 \).

The OLS estimator of \( \beta \) is given by

\[
\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \left( X'X \right)^{-1} X'Y
\]

\( \hat{\beta} \) follows multivariate normal distribution with mean \( \beta \) and covariance matrix \( \sigma^2 \left( X'X \right)^{-1} \).

174
The OLS estimator of $\beta_1$ constrained by $\beta_2 = 0$ is given by

$$\hat{\beta}_{10} = \left( X'_1 X_1 \right)^{-1} X'_1 Y_1 = \hat{\beta}_1 + \left( X'_1 X_1 \right)^{-1} X'_1 X_2 \beta_2 \quad ... (4.15.2)$$

Further, whether $\beta_2 = 0$ or not,

$$\hat{\beta}_{10} \sim N(\beta_1, b, \sigma^2 \left[ X'_1 X_1 \right])$$

where

$$b = \left( X'_1 X_1 \right)^{-1} X'_1 X_2 \beta_2$$

Thus, $\hat{\beta}_{10}$ is an unbiased and efficient estimator of $\beta_1$ when $b = 0$, and otherwise $\hat{\beta}_{10}$ is biased.

However, $\hat{\beta}_1$ is unbiased regardless of $b$.

Hausman (1978) defined a specification test for testing $H_0$: $b = 0$ against $H_1$: $b \neq 0$ based on the statistic

$$\lambda = \left( \hat{\beta}_1 - \hat{\beta}_{10} \right)' \left[ V(\hat{\beta}_1) + V(\hat{\beta}_{10}) \right]^{-1} \left( \hat{\beta}_1 - \hat{\beta}_{10} \right) \quad ... (4.15.3)$$

Under $H_0$: $b = 0$, the distribution of $\lambda$ is a central $\chi^2_{k_1}$.

If $\sigma^2$ is known, the size of the specification test is given by

$$P \left| \chi^2_{k_1} \geq C^2 \right| = \alpha \quad ... (4.15.4)$$

175
If \( \lambda < C^2 \), \( H_0 \) is not rejected; otherwise, \( H_0 \) is rejected in favour of \( H_1 \). Where \( C \) is the critical value.

In other words, one can choose \( \hat{\beta}_{10} \) as an estimator of \( \beta_1 \) if \( \lambda < C^2 \) and \( \hat{\beta}_1 \) if \( \lambda \geq C^2 \).

\[
\beta^*_1 = \hat{\beta}_{10} I_{[\lambda < C^2]} + \hat{\beta}_1 I_{[\lambda \geq C^2]}
\]

\[
= \hat{\beta}_{10} + (\hat{\beta}_1 - \hat{\beta}_{10}) I_{[\lambda \geq C^2]}
\]  

Where \( I_{[\lambda \geq C^2]} = 1 \) if \( \lambda \geq C^2 \)

\[
= 0 \text{ otherwise}
\]

\( \beta^* \) depends upon the size \( \alpha \) through \( C^2 \)

When \( \alpha = 0 \), \( C = + \alpha \) and \( \beta^*_1 = \hat{\beta}_{10} \);

When \( \alpha = 1 \), \( C = 0 \) and \( \beta^*_1 = \hat{\beta}_1 \);

Hence, the unconstrained and the constringed OLS estimates may be seen as limit cases of the specification pre-test estimator.
Hausman and Taylor (1981) have shown that the $\lambda$-statistic inequivalent (if $\sigma^2$ is known) to the standard Wald statistic computed to test the hypothesis that $b=0$, when rank of $\left( X_1'X_2 \right) = \min (k_1, k_2)$.

Gourieroux and Trognon (1984) have derived a specification pre-test estimator when rank of $\left( X_1'X_2 \right) = k_1$.

The specification pre-test estimator of $\beta_1$ is given by:

(i) $\hat{\beta}_{10}$, if the test fails to reject $H_0$: $b = \left( X_1'X_1 \right) X_1'X_2 \hat{\beta}_2 = 0$.

(ii) $\hat{\beta}_1$, if the test rejects $H_0$: $b = 0$.

The classical pre-test estimator of $\beta_1$ is given by:

(i) $\beta^{**}$, the OLS estimator of $\beta_1$ constrained by

$$\left( X_1'X_1 \right)^{-1} X_1'X_2 \hat{\beta}_2 = 0,$$

when the test fails to reject $H_0$: $b = 0$.

(ii) $\hat{\beta}_1$, otherwise.

Where, $\beta^{**} = \begin{bmatrix} \beta_1^{**} \\ \beta_2^{**} \end{bmatrix} = \hat{\beta} - \left( X'X \right)^{-1} R' \begin{bmatrix} R' \left( X'X \right)^{-1} R' \end{bmatrix}^{-1} R \hat{\beta}$.
Here, \[ R = \frac{1}{2} \left( X' \right)^t X' X_2 \]

We deduce that

\[ R \left( X' \right)^t = \left[ X_1' X_1 \right] X_2 [X_2' Q_1 X_2] X_1' X_1 \]

Where, \[ Q_1 = I - X_1 \left[ X_1' X_1 \right] X_1' \] is the orthogonal projector on the orthogonal of the subspace generated by the columns of \( X_1 \).

\[ \therefore \beta_{\text{est}} = \hat{\beta}_1 + R \hat{\beta} = \hat{\beta}_1 + \left( X_1' X_1 \right) X_1' X_2 \hat{\beta}_2 = \hat{\beta}_{10} \quad \ldots (4.15.6) \]

In the case of the linear model, Gourieroux and Trognon (1984) have compared the biases and the mean squared errors of the classical and specification pre-test estimators and observed that the latter is not uniformly better than the former and conversely.
4.16 SELECTING REGRESSORS USING THE MEASURE OF PREDICTION SUM OF SQUARES

Consider the classical general linear model as

\[ y_{nx1} = X_{nxk} \beta_{kx1} + e_{nx1} \]  \hspace{1cm} (4.16.1)

With \( E[\epsilon] = 0, \ E[\epsilon \epsilon'] = \sigma^2 I \) and the design matrix \( X \) has rank \( k \), where \( k < n \).

Defining the predicted residuals as

\[ e_{(i)} = y_i - X_i' \hat{\beta}_{(i)} \]  \hspace{1cm} (4.16.2)

Where \( X_i \) is \( k \times 1 \) vector of \( i^{th} \) data point and \( \hat{\beta}_{(i)} \) is the OLS estimate of \( \beta \) with the \( i^{th} \) case is excluded.

Also, one can have

\[ e_{(i)} = \frac{e_i}{1 - v_{ii}}, \ i = 1, 2, ..., n \]  \hspace{1cm} (4.16.3)

Where \( e_i \)'s are the OLS residuals and \( v_{ii} \)'s are the \( i^{th} \) diagonal elements of matrix \( v = XX' - \hat{\beta}_{(i)}X' \hat{\beta}_{(i)} \)
Now, the predicted residual sum of squares (PRESS) is defined as

\[ \text{PRESS} = \sum_{i=1}^{n} \hat{e}_{(i)}^2 \]  \quad \ldots (4.16.4)

or

\[ \text{PRESS} = \sum_{i=1}^{n} \left(1 - v_{ii}\right) \hat{e}_{(i)}^2 \]  \quad \ldots (4.16.5)

Where

\[ = X \left( \mathbf{X}' \mathbf{X} \right)^{-1} \mathbf{X}' = \left( \lambda_{ii} \right) \]

Consider two competing models:

**Model I** :  \[ Y = X \beta + \epsilon \]

**Model II** :  \[ Y = X' \beta' + \epsilon' = Z \alpha + \epsilon' \]

According to PRESS criterion, Model II can be selected iff

\[ \sum_{i=1}^{n} \frac{\hat{e}_{(i)}^2}{\left(1 - v_{ii}\right)^2} > \sum_{i=1}^{n} \left| \frac{\hat{e}_{(i)}^2}{\left(1 - v_{ii}\right)^2} \right| \]  \quad \ldots (4.16.6)
Where $e_i$ and $v_{ii}$ are analogous to $e_i$ and $v_{ii}$ with $X^*$ replacing $X$.

Thus, the regressors $Z$'s are included along with $X$ if and their inclusion lowers the out-of-sample PRESS. This method involves the calculation of two OLS regressions and two weighted sums of squared residuals.

For the detection of outliers, a modified form of PRESS namely $Q^2$-statistic defined by Wold (1982), which is given by

$$Q^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2(i)}{\sum_{i=1}^{n} [Y_i - \bar{Y}_{(i)}]^2}$$

... (4.16.7)

Where $\sum_{i=1}^{n} [Y_i - \bar{Y}_{(i)}]^2$ is the sum of squares of deviations of observations from the mean $\bar{Y}_{(i)}$ when the $i^{th}$ observation is omitted.

Quan (1988) proposed the PRESS as a measure for the regression diagnostics.
4.17 TESTING LINEAR AND LOGARITHMIC REGRESSION MODELS USING VARIABLE ADDITION AND LAGRANGE MULTIPLIER TESTS

Economic theory typically provides information regarding the variables that should be included in an econometric relationship, but it does not frequently suggest an appropriate functional form. In the literature many tests of linear and log-linear models have been categorized as follows:

(i) tests that explicit the fact that one model is tested against a specific non-nested alternative.

(ii) two types of tests of the linear and log-linear models against the more general model of Box and Cox (1964), one based on the Lagrange Multiplier Principle and the other based on methods proposed by Andrews (1971);

(iii) diagnostic tests of (possible) functional form misspecification against an unspecified alternative.

Godfrey, Mc Aleer and Mc Kenzie (1988) have specified the log-linear and linear models to be tested, which are given by

182
\[ H_0: \log Y_i = \beta_0 + \sum_{j=1}^{k} \hat{\beta}_j \log X_{i,j} + \epsilon_{0i} \quad \ldots \quad (4.17.1) \]

Such that \[ \mathbb{E}(\epsilon_{0i}) = 0, \mathbb{E}(\epsilon_{0i}^2) = \sigma_0^2 \]

and \[ H_1: Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{i,j} + \epsilon_{1i} \quad \ldots \quad (4.17.2) \]

Such that \[ \mathbb{E}(\epsilon_{1i}) = 0, \mathbb{E}(\epsilon_{1i}^2) = \sigma_1^2 \]

Where \[ i=1, 2, \ldots, n. \]

Denote the OLS estimates of the parameters of the log-linear model \( H_0 \) by \( \hat{\beta}_0 \) and \( \hat{\beta}_j \) and the corresponding estimates for the linear model \( H_1 \) by \( \hat{\beta}_0 \) and \( \hat{\beta}_j \) for \( j=1, 2, \ldots, k \). The associated estimates of the error variances will be denote by \( \hat{\sigma}_0^2 \) and \( \hat{\sigma}_1^2 \) respectively.

The OLS predicted value of \( \log Y_i \) under \( H_0 \) and \( H_1 \) are written, respectively as

\[ \hat{\log}(Y_i) = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j \log X_{i,j} \quad \ldots \quad (4.17.3) \]

or

\[ \hat{Y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j X_{i,j} \quad \ldots \quad (4.17.4) \]
A simple approach to test the non-nested (separate) families of hypotheses $H_0$ and $H_1$ can be devised by considering the class of PE (i.e., Extended Projection) tests derived by MacKinnon, White and Davidson (1983). It can be shown that the PE tests of $H_0$ and $H_1$ are equivalent to testing $\alpha_0 = 0$ and $\alpha_1 = 0$ respectively, in

$$\log Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j \log X_{pi} + \alpha_0 \, \hat{Y}_i - \exp \{ \log Y_i \} + \epsilon_i \quad \ldots (4.17.5)$$

and

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j \, X_{pi} + \alpha_1 \, \log y_i \, \log Y_i + \epsilon_i \quad \ldots (4.17.6)$$

Given that (4.17.5) and (4.17.6) simply involve augmenting the original log-linear and linear models with an additional variable.

4.17.1 Exact Tests

Bera and McAleer (1988) proposed tests based on two auxiliary regressions, commencing with

$$\exp \left\{ \log Y_i \right\} = \beta_0 + \sum \beta_j \, X_{pi} + \eta_{yi} \quad \ldots (4.17.7)$$

and

$$\log \hat{Y}_i = \beta_0 + \sum \beta_j \, \log X_{pi} + \eta_{0i} \quad \ldots (4.17.8)$$
Denoting the OLS residuals from (4.17.7) and (4.17.8) as \( \eta_{1i} \) and \( \eta_{0i} \) respectively, the tests of \( H_0 \) and \( H_1 \) are based on OLS estimation of \( \theta_0 \) and \( \theta_1 \) in the auxiliary regressions.

\[
\log Y_i = \beta_0 + \sum_j \beta_j \log X_{pi} + \theta_0 \eta_{1i} + \epsilon_i \quad \cdots (4.17.9)
\]

and

\[
Y_i = \beta_0 + \sum_j \beta_j X_{pi} + \theta_1 \eta_{0i} + \epsilon_i \quad \cdots (4.17.10)
\]

The tests in (4.17.9) and (4.17.10) will be denoted as BM. If the disturbance under \( H_0 \) is normally distributed, the test of \( H_0: \theta_0 = 0 \) is distributed as \( \mathcal{N} \) (n-k-2) under \( H_0 \). It was argued above that the assumption of \( \epsilon_i \), being normally distributed as inconsistent with entertaining the use of a logarithmic transformation of \( Y_i \), since the normality assumption implies a non-zero probability for \( Y_i \) being negative. Consequently, it is difficult to argue that the t-test of \( H_1: \theta_1 = 0 \) in (4.17.10) is distributed exactly as \( t \) (n-k-2) infinite samples. However, as long as the disturbances in (4.17.9) and (4.17.10) satisfy appropriate regularity conditions, the BM statistics will be asymptotically distributed as \( \mathcal{N}(0,1) \) varieties under the respective null hypothesis.

185
4.17.2 RESET Test for Testing Functional Form between Linear and Loglinear Regression Models

Ramsey and Schmidt (1976) have modified the RESET test and suggested for testing functional form between linear and loglinear models, based on the auxiliary regressions

\[(i) \log Y_j = \beta_0 + \sum_{i=1}^{k} \beta_i \log x_{ij} + \sum_{p=2}^{c} \alpha_p (\log Y_j)^p + \epsilon_j \quad \cdots (4.17.11)\]

and \[(ii) Y_j = \beta_0 + \sum_{i=1}^{k} \beta_i x_{ij} + \sum_{p=2}^{c} \alpha_p (Y_j)^p + u_j \quad \cdots (4.17.12)\]

Where \((\log Y_j)^p\) and \((Y_j)^p\) are the predicted values of \(Y_j\) under \(H_0\) and \(H_1\) respectively, raised to the power 'p'. The test of \(H_0\) against \((4.17.11)\) is the test of \(c_{02} = \alpha_{03} = \ldots = \alpha_{0k} = 0\) and is distributed as \(F_{(c-k,n-k-c)}\) under \(H_0\) if the disturbances are normal.

Similarly, the RESET test \(H_1\) in \((4.17.12)\) is test of \(\alpha_{12} = \alpha_{13} = \ldots = \alpha_{1k} = 0\) which is approximately distributed and \(F_{(c-k,n-k-c)}\) under \(H_1\) is the disturbances are approximately normal.
4.17.3 White's Test For Functional Form

White (1980) suggested a functional form test in which the regression parameters are estimated by both OLS and Weighted Least Squares (WLS) procedure, and the difference between these estimators is examined. If the null model is specified correctly, then the OLS and WLS estimators should differ only by amounts attributable to sampling fluctuations.

If the functional form has been misspecified, then the OLS and WLS estimates can be expected to have different probability limits, with their difference tending to a non-zero vector as \( n \) approaches infinity. White suggested that the weights for WLS be the reciprocals of the squared predictions under the null hypothesis, as was the Breusch - Godfrey (1986) algorithm for calculating White's Functional Form by the method of variable addition.

The white's functional form (WFF) tests of

\[
H_0 : \psi_{00} = \psi_{01} = \ldots \psi_{0k} = 0 \quad \text{and} \quad H_1 : \psi_{10} = \psi_{11} = \ldots \psi_{1k} = 0
\]

respectively, in the auxiliary regressions.
\[ \log Y_j = \beta_0 + \sum_{i=1}^{k} \beta_i \log X_i + \psi_{i0} \left( \log Y_j \right)^2 + \sum_{i=1}^{k} \psi_{i1} \left[ \log X_{ij} \left( \log Y_j \right)^2 \right] \]
and
\[ Y_i = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \psi_{10} \left( \bar{Y}_i \right)^2 + \sum_{i=1}^{k} \psi_{11} \left[ X_{ij} \left( \bar{Y}_i \right)^2 \right] + \epsilon_i \]

With the test statistic of \( H_0 \) being distributed as \( F_{(k+1,n-2k-2)} \) if the disturbances are normally distributed and that of \( H_1 \) being approximately distributed as \( F_{(k+1,n-2k-2)} \).

4.17.4 Utts Rainbow Test

Utts (1982) developed a rainbow test which covers a wide spectrum of possible misspecifications. It is based on comparing two sums of squared residuals, and being derived by using the complete data set and the other being calculated using only a subset of \( n_1 \) observations. The sample will be assumed to be split into three parts containing \( n_0, n_1 \), and \( n_2 \) observations with \((k+1) < n_1 < n\). The test statistic is given by Utts (1982) which is equivalent to

\[
F = \frac{(SSE - SSF_1)}{\frac{(\eta_0 + \eta_2)}{(n_0 - k - 1)}} \quad \text{... (4.17.13)}
\]
Where $\text{SSE}$ and $\text{SSE}_0$ are the residual sum of squares obtained when the null model is estimated by OLS using all $n$ observations and only the subset of $n_1$ observations respectively.

The test statistic (4.17.1) is distributed as $F(n_0 + n_1, n_1 - k - 1)$ under the null hypothesis of no misspecification.

**Remarks**: The following three points are to be noted while applying Utts rainbow test.

(i) Utts recommends choosing points with low leverage (i.e., points corresponding to the $n_1$ smaller diagonal elements $X (X^\top X)^{-1} X^\top$, for subsample estimation.

(ii) Utts (1982) also recommends setting $n_1 = n_0 + n_2 = \frac{n}{2}$, so that approximately one-half of the data points are used for subset estimation, the corresponding rainbow test will be referred to as $\text{Utts} \left| \frac{n}{2} \right.$

(iii) The rainbow test is not robust to non-normality.
4.18 SPECIFICATION ANALYSIS OF ECONOMETRIC MODELS

The criteria of model adequacy have been formed in the literature as model adequacy tests; more simply just diagnostic tests since they are based on the residuals of the models. Ullah (1985) suggested that most of the specification tests can be derived by using the principles of likelihood ratios, Wald or Rao's score (or Lagrangian multiplier).

Consider an adequacy test of the linear regression model as,

\[ H_0 : Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I) \]  \quad \ldots (4.18.1)

Against the alternative

\[ H_1 : Y = X\beta + Z\delta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I) \]  \quad \ldots (4.18.2)

Where \( Y \) is an \( n \times 1 \) vector of observations on a dependent variable; \( X \) and \( Z \) are respectively \( n \times k \) and \( n \times q \) matrices of observations on independent variables; \( \beta \) and \( \delta \) are \( k \times 1 \) and \( q \times 1 \) unknown parametric vectors respectively; \( \sigma \) is an unknown scalar.

Thus, the null hypothesis \( H_0 \) is \( \delta = 0 \) and the alternative \( H_1 \) is \( \delta \neq 0 \).
This can be tested by an $F$ statistic,

$$F = \frac{(RSS)_0 - (RSS)}{\hat{\sigma}^2} \cdot \frac{q}{1} \quad \ldots (4.18.3)$$

Where $(RSS)_0$ is the residual sum of squares from $H_0$ and $(RSS)_1$ is the residual sum of squares from $H_1$ and $\hat{\sigma}^2$ is the OLS estimate of $\sigma^2$ from $H_0$.

It should be noted here that $Z$ represents a set of omitted variables causing inadequacy in the model. Quite often, the data on these omitted variables are not available and so the proxy $\hat{Z}$ is used.

The linear model under $H_1$ may be written as

$$Y = X\beta + \hat{Z}\delta + u,$$

where

$$u = (Z - \hat{Z})\delta + \epsilon \quad \ldots (4.18.4)$$

and $H_0: \delta = 0$ is tested again using the $F$-statistic.

It should be again noted that there are three equivalent ways of formulating a test for $\delta = 0$. This can be verified as

(i) $F$-test on $\delta = 0$ in $H_1$ \quad \ldots (4.18.5)
(i) F-test on $\delta = 0$ in $\hat{e} = X\gamma + Z\delta + \epsilon$.

$$\hat{e} = Y - X\hat{\beta} \quad \text{and} \quad \gamma = \beta - \hat{\beta}$$

... (4.18.6)

(iii) F-test on $\delta = 0$ in $\hat{e} = Z\delta + \epsilon^*, \epsilon^* = X(\beta - \hat{\beta}) + \epsilon$

... (4.18.7)

Where $\hat{\beta}$ is the OLS estimator from this linear model under $H_n$.

When using $\hat{Z}$, replace $Z$ by $\hat{Z}$ in (4.18.5) (4.18.6) and (4.18.7) above, and adjust the disturbance term as in (4.18.4).

Ullah (1985) examined the exact nature of $Z$ in (4.18.2) or $\hat{Z}$ in (4.18.4) for various specification tests.

1. Data Coherency

This refers to the notion that a model should adequately reflect the data behaviour, that is, should be data consistent. This is checked by

(i) Goodness of fit $R^2, R^2$, Akaike information criterion
(ii) Specification error test : Ramsey (1969) RESET test

(iii) Autocorrelation test : AR (1) test

(iv) Heteroscedasticity : Engle’s (1982) ARCH test


2. Functional Form Test (Linear Versus Loglinear)

Suppose the unknown functional form has the structure

$$Y = g(X, \lambda, \theta) + \epsilon$$

specialising to a linear model $Y = X\beta + \epsilon$ in (4.18.1) when $\lambda = \lambda_0$. Then

a test of the hypothesis that $\lambda = \lambda_0$ can be found by considering

$$Z = g_1(X, \lambda_0, \theta)$$

in (4.18.6) where $g_1(\cdots)$ is the first derivative of $g(\cdots)$ with respect to $\lambda$ and evaluated at $\lambda = \lambda_0$

4.19 RELATIONSHIP AMONG HAUSMAN, RAMSEY AND CHOW TESTS FOR MODEL SPECIFICATION

Thursby (1985) described a relationship between the tests for specification error in regression models introduced by Ramsey (1969) and Hausman (1978). Further he has shown that the Chow (1960) test
for structural shift can be viewed as a special case of specification error
tests.

Consider the model

\[ Y_{n×1} = X_{n×k} β_{k×1} + ε_{n×1} \]  \hspace{1cm} (4.19.1)

and

\[ ε \sim \text{Multivariate } N (0, σ²I) \]

The hypothesis to be considered are

\[ H₀ : E(ε/X) = 0 \sim H₁ : E(ε/X) = Z \neq 0 \]  \hspace{1cm} (4.19.2)

Where \( Z \) is an unobserved \( n \times 1 \) vector. The nature of \( Z \) is
unknown to the researcher and \( Z \) is nor orthogonal to \( X \).

The alternative hypothesis is diffuse, and the tests of such
hypothesis are 'non constructive' in the sense of Goldfield and Quandt
(1972) or 'General' in the sense of Ramsey (1974).

Hausman (1978) proposed a general approach to test for a nonzero
disturbance mean conditional on \( X \) in the regression analysis. Basically,
the approach is to find an alternative estimator \( \hat{β}_1 \) that is consistent
under both \( H₀ \) and \( H₁ \) and compare this estimator with an estimator \( \hat{β}_0 \)
that is efficient under \( H₀ \) but inconsistent under \( H₁ \).
Hausman assumes that under $H_0, \sqrt{n}(\hat{\beta}_0 - \beta)$ and $\sqrt{n}(\hat{\beta}_1 - \beta)$ are asymptotically $N(0,V_0)$ and $N(0,V_1)$ respectively and considers the difference $\hat{q} = (\hat{\beta}_1 - \hat{\beta}_0)$ and the test statistic

$$m = \hat{q}' \hat{M}(\hat{q})^{-1} \hat{q}, \quad \ldots \text{(4.19.3)}$$

Where $\hat{M}(\hat{q})$ is a consistent estimator of the (asymptotic) covariance matrix of $\hat{q}$.

$$M(\hat{q}) = (V_1 - V_0)/n$$

Under $H_0$, $m$ is asymptotically follows central $\chi^2_k$ and is noncentral $\chi^2$ for local departures from $H_0$.

Hausman and Taylor (1981) have extended the test to the case of singular $\hat{M}(\hat{q})$ by replacing $\hat{M}(\hat{q})^{-1}$, with the generalized inverse $\hat{M}(\hat{q})^-$. 

For the model (4.19.1) and hypothesis (4.19.2) the efficient estimator of $\beta$ under $H_0$ is the OLS estimator $\hat{\beta}_0 = (X'X)^{-1}X'Y$ and

$$\hat{\beta}_0 \sim N[\beta, \sigma^2(X'X)^{-1}]$$
Remarks

In the case of stochastic $X$, these should be considered to be conditional on $X$.

Under stochastic $X$, to obtain the test statistic, one must find an estimator such that $\text{Plim}(\hat{\beta}_0) = \beta$ under $H_0$ and $\text{plim} \hat{\beta}_1 \neq \text{plim} \hat{\beta}_0$ under $H_1$. Numerous possibilities for $\hat{\beta}_1$ exist one of these leads to Ramsey's RESET statistic and one leads to the Chow test statistic.

4.19.1 Relationship Between Hausman's Test Statistic and Ramsey's RESET Statistic

Ramsey's RESET procedure is a test of hypothesis (4.19.2) and amounts to a standard $F$ test of the significance of the $I$ estimates in the augmented regression.

$$Y = X\beta + W\Gamma + \epsilon \quad \ldots (4.19.4)$$

Where $W$ is a $(n \times g)$ matrix of rank 'g' of test variables such as powers of the $X$ variables and $\text{Rank of } (X, W) = k + g$.

Following Hausman's approach one may compute

$$\hat{\beta}_0 = (X'X)^{-1}X'Y \quad \ldots (4.19.5)$$
and \( \hat{\beta}_1 \), which is composed of the first \( k \) elements of

\[
[(X, W)\,^\prime (X, W)]^{-1} (X_1 W)^\prime Y
\]

Under \( H_0 \), both are consistent and unbiased with variances.

\[
V(\hat{\beta}_0) = \sigma^2 (X^\prime X)^{-1}
\]

\[
V(\hat{\beta}_1) = \sigma^2 (X^\prime M_0 X)^{-1}
= \sigma^2 P(W^\prime M_0 W)^{-1} P^\prime + \sigma^2 (X^\prime X)^{-1}
\quad (4.19.6)
\]

Where

\[
M_w = [I - W(W^\prime W)^{-1}W^\prime];
\]

\[
M_0 = [I - X(X^\prime X)^{-1}X^\prime]
\]

and \( p = (X^\prime X)^{-1} X^\prime W \)

Clearly, \( \hat{\beta}_0 \) is efficient. Under \( H_1 \), we have,

\[
E(\hat{\beta}_0) = \beta + (X^\prime X)^{-1} X^\prime Z
\]

\[
E(\hat{\beta}_1) = \beta + (X^\prime X)^{-1} X^\prime Z - P(W^\prime M_0 W)^{-1} W^\prime M_0 Z
\]

Where,

\[
Z = E(\varepsilon/X).
\]

Under \( H_0 \), \( \tilde{q} = (\hat{\beta}_1 - \hat{\beta}_0) \sim N(0, M(\tilde{q})) \)
Where $M(\hat{\sigma}) = \sigma^2 \left[ (X^\top M_0 X)^{-1} - (X^\top X)^{-1} \right]$

$$= \sigma^2 \, P \left( W^\top M_0 W \right)^{-1} P$$

under $H_1$, we have

$$E(\hat{\sigma}) = -P \left( W^\top M_0 W \right)^{-1} W^\top M_0 Z$$

A number of forms of Hausman statistic can be defined for this problem depending on the choice of estimator for $M(\hat{\sigma})$, such as the following two possibilities:

$$m_1 = \hat{\sigma}^2 \left[ P \left( W^\top M_0 W \right)^{-1} P \right] \hat{\sigma} / S_0^2 \quad \ldots \quad (4.19.8)$$

and

$$m_2 = \hat{\sigma}^2 \left[ P \left( W^\top M_0 W \right)^{-1} P \right] \hat{\sigma} / S_1^2 \quad \ldots \quad (4.19.9)$$

Where $S_0^2$ (or $S_1^2$) is the estimator of $\sigma^2$ from the OLS regression of $Y$ as $X$ (or $Y$ on $X$ and $W$).

The statistic $m_2$ most closely follows Hausman's applications of the test. Each of the above statistics follows asymptotically $\sigma^2 \left( J \right)$ under $H_0$. 198
Where \( J = \text{Rank} (X, W) = \text{Min} (k, g) \).

To examine the relationship between \( m_1 \), \( m_2 \) and RESET, write the numerators of \( m_1 \) and \( m_2 \) as

\[
(W^1 M_0 W) (W^1 M_0 W)^{-1} [P^1 \{ P (W^1 M_0 W)^{-1} P^1 \}^{-1} P \\
(W^1 M_0 W)^{-1} W^1 M_0 Y = C_1
\]

The RESET statistic is given by

\[
Y^1 M_0 W (W^1 M_{ow})^{-1} W^1 M_0 Y / g S^2_1 = C_2 / g S^2_1 \quad \ldots (4.19.10)
\]

If \( C_1 = C_2 \) then \( m_1 \) is equivalent to the numerator \( \chi^2 \) of the RESET F-statistic times \( \sigma^2 / S^2_0 \) and \( m_2 \) is precisely the RESET statistic (except for the degrees of freedom adjustment).

Thursby (1985) established that the necessary and sufficient condition for \( C_1 = C_2 \) is the Rank of \( (X^1 W) = g. \) The relationship of Hausman’s test to the RESET procedure is not surprising, given that Hausman’s test can generally be expressed as a test of \( H_0 : \alpha = 0 \) in the regression

\[
Y = X\beta + \bar{X} \alpha + \epsilon
\]

Where \( \bar{X} \) is a suitably transformed version of \( X. \)
Relationship with the Chow Test

The Chow test or a test for the equality of regression populations, is a test of equality of \( \beta_1 \) and \( \beta_2 \) in the regressions

\[
Y_i = X_i \beta_i + \epsilon_i, \quad i = 1, 2
\]

Where the \( Y_i \) are the \( n \times 1 \); then \( X_i \) are regression matrices of order \( n_i \times k \); and the \( \beta_i \) are the unknown parametric vectors of order \( k \times 1 \). The \( \epsilon \) are jointly normal with zero means and equal variances and are independent of the \( X \).

Chow (1960) proposed tests for the cases

\[
\text{Rank of } (X_1) = k < n
\]

and \( \text{R} (X_1) = k < n \), but \( \text{Rank} (X_2) = n_2 < k \)

[ If \( \text{Rank} (X_2) = n_2 = k \) then the tests are equivalent. ]

Consider the pooled regression as

\[
\begin{align*}
Y_1 & = \beta + \epsilon_1 \\
Y_2 & = X_2 \beta + \epsilon_2
\end{align*}
\]
If $\text{SSE}$ is the sum of squared errors from the pooled regression and $\text{SSE}_1$ and $\text{SSE}_2$ are the sum of squares of $Y_1$ on $X_1$ and $Y_2$ on $X_2$ respectively. Then the Chow test statistic is given by

$$R^* = \frac{Q_1/K}{Q_2/(n-2k)}$$

... (4.19.12)

Where $Q_1 = (\text{SSE} - \text{SSE}_1 - \text{SSE}_2)$ and $Q_2 = \text{SSE}_1 + \text{SSE}_2$.

Under the $H_0$, the statistic is distributed as

$$F(k, n-2k), \text{ where } n = n_1 + n_2$$

As a direct application of Hausman's approach we can test the regression model (4.19.11) against the alternative hypothesis $H_1$ in (4.19.2).

We consider two estimators,

$$\hat{\beta}_0 = (X'X)^{-1}X'Y$$

and $$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'Y_1.$$ 

Where $\hat{\beta}_0$ is clearly the efficient under the null hypothesis of no specification error. Again Let \( \hat{\epsilon} = \hat{\beta}_1 - \hat{\beta}_0 \).

under $H_0$, $E(\hat{\epsilon}) = 0$ and under $H_1$, we have
\[ E (\tilde{q}) = \left( x'^{l} x^{l} \right)^{-1} \left( x'^{l} z_{1} \right) - \left( x'^{l} x^{l} \right) x'^{l} z \]

Thurby (1985) defined the following statistics

\[ m_{1} = \tilde{q} \left( x'^{l} x^{l} \right)^{-1} - \left( x'^{l} x^{l} \right)^{l} \tilde{q} / S^{2} \]

\[ m_{2} = \tilde{q} \left( x'^{l} x^{l} \right)^{-1} - \left( x'^{l} x^{l} \right)^{l} \tilde{q} / S^{2} \]

\[ m_{3} = \tilde{q} \left( x'^{l} x^{l} \right)^{-1} - \left( x'^{l} x^{l} \right)^{l} \tilde{q} / S^{2} \]

under \( H_{0} \), each form of the test statistic is asymptotically \( \chi^{2} (k) \).

**Remark**: The term \( \tilde{q} \left( x'^{l} x^{l} \right)^{-1} - \left( x'^{l} x^{l} \right)^{l} \)

appearing in \( m_{3}, m_{4} \) and \( m_{3} \) is exactly the term \( Q_{1} \) of the Chow test statistic. The estimator \( S_{R}^{2} \) is

\[ Q_{p} / (n - 2k) \) and thus \( m_{3} \) is the Chow statistic (except for the degrees of freedom adjustment).