BECE 142

APPLIED ECONOMETRICS

School of Social Sciences
Indira Gandhi National Open University

EXPERT COMMITTEE

Prof. Atul Sarma (retd.) Former Director Indian Statistical Institute, New Delhi	Prof. M S Bhat (retd.) Jamia Millia Islamia New Delhi	Prof. Gopinath Pradhan (retd.) Indira Gandhi National Open University, New Delhi
Dr. Indrani Roy Choudhury CSRD, Jawaharlal Nehru University New Delhi	Dr. S P Sharma Shyamlal College (Evening) University of Delhi	Prof. Narayan Prasad Indira Gandhi National Open University, New Delhi
Sri B S Bagla (retd.) PGDAV College University of Delhi	Dr. Manjula Singh St. Stephens College University of Delhi	Prof. Kaustuva Barik Indira Gandhi National Open University, New Delhi
Dr. Anup Chatterjee (retd.) ARSD College, University of Delhi	Saugato Sen Indira Gandhi National Open University, New Delhi	Prof. B S Prakash (Course Coordinator) Indira Gandhi National Open University, New Delhi

COURSE PREPARATION TEAM

Block 1	**Empirical Issues in Econometric Research**	
Unit 1	Stages in Empirical Research	Prof. B. S. Prakash, SOSS, IGNOU.
Unit 2	Specification Issues	Rimpy Kaushal, Assistant Professor, Dept of Economics, PGDAV College, DU.
Unit 3	Model Selection Criteria	Rimpy Kaushal, Assistant Professor, Dept of Economics, PGDAV College, DU.
Block 2	**Advanced Topics in Regression Analysis**	
Unit 4	Auto Regressive and Distributed Lag Models	Prof. Sushil Haldar, Jadavpur University.
Unit 5	Binary Dependent Variable Models	Prof. Sushil Haldar, Jadavpur University.
Unit 6	Simultaneous Equations Models I	Prof. Sushil Haldar, Jadavpur University.
Unit 7	Simultaneous Equations Models II	Prof. Sushil Haldar, Jadavpur University.
Block 3	**Panel Data Models**	
Unit 8	Introduction to Panel Data	Dr. Poulomi Roy, Jadavpur University.
Unit 9	Estimation of Panel Data Models	Dr. Poulomi Roy, Jadavpur University.
Block 4	**Econometric Software Packages**	
Unit 10	Introduction to R	Dr. Sudip Mukherjee, Asst. Prof. of Economics, Dinabandhu Mahavidyalaya, Bongaon, West Bengal.
Unit 11	Introduction to EViews	Dr. Poulomi Roy, Jadavpur University.
Unit 12	Introduction to Stata	Dr. Poulomi Roy, Jadavpur University.

General Editor
Content, Format and Editing: Prof. B. S. Prakash and Sh. B. S. Bagla

PRINT PRODUCTION

Mr. Tilak Raj
Assistant Registrar
MPDD, IGNOU, New Delhi

June, 2022

ISBN: 978-93-5568-176-8

Further information on the Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi,
New Delhi -110068 or visit our website: http://www.ignou.ac.in

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi, by Ragistrar, MPDD, IGNOU, New Delhi.

Laser Typeset by: Tessa Media & Computers, C-206, Shaheen Bagh, Jamia Nagar, New Delhi

Printed at: S G Print Packs Pvt. Ltd., F-478, Sector-63, Noida-201 301

CONTENTS

COURSE INTRODUCTION

In your first course on Econometrics (BECC 110), you have learnt the various aspects of Classical Linear Regression Models. The data sets to which you would apply the OLS method learnt there is cross section data (e.g. data on variables for different states of India) and time series data (e.g. data for specified time points for different variables). Note that the term Time Series Data has another specific meaning where the variable at time point 't' is dependent on its own value at time point 't –1'. The methods for analysing data sets of this specific type (called Time Series Analysis) are not covered in your present programme at Honours level. Furthering your econometrics skills gathered in the course BECC 110, the present course (BECE 142) exposes you to carry out regression modeling in certain specific types of situations. These include situations where: (i) the dependent variable takes binary values, (ii) the dependent and independent variables influence each other (a situation called simultaneity) and (iii) you have a mix of cross-section and time-series data (called panel data sets). The course also covers aspects of empirical research, model selection, specification errors and software packages for econometrics. Brief outline of the different blocks and units covered in the course is as follows.

The first block (**Block 1**) is on 'Empirical Issues in Econometric Research'. It has three units. **Unit 1** is on Stages in Empirical Research. It first enumerates the different steps in empirical research. It then presents an account of 'research methodology'. Being the introductory unit to your second course on econometrics (first one being the Core Course BECC 110 on Introductory Econometrics), the unit presents a 'review of CLRM (classical linear regression model)'. **Unit 2** is on 'Specification Issues'. This unit covers three important issues viz. consequences of (i) omission of relevant variables, (ii) inclusion of irrelevant variables and (iii) errors in measurement. **Unit 3** is on Model Selection Criteria. The unit first introduces some 'tests for specification errors'. It then introduces many methods of 'model selection criteria'.

Block 2 is on 'Advanced Topics in Regression Analysis'. The block covers four major areas viz. Auto-Regressive and Distributed Lag Models (**Unit 4**), Binary Dependent Variable Models (**Unit 5**), Simultaneous Equations Models I (**Unit 6**) and Simultaneous Equations Models II (**Unit 7**). **Unit 4** discusses the (i) significance of lags in economics, (ii) solution to distributed lag models and (iii) alternative approaches to solving distributed lag models. **Unit 5** discusses three models viz. (i) The Linear Probability Model (LPM), (ii) The Logit Model and (iii) The Probit Model. **Unit 6** introduces you to Simultaneous Equation Models. It particularly discusses: (i) consequences of simultaneity, (ii) problem of identification and (iii) the conditions for identification. **Unit 7** furthers this topic introducing you to three methods of estimation viz. (i) indirect least square (ILS) method, (ii) instrumental variables (IV) method and (iii) two stage least squares (2 SLS) method.

Block 3 is on 'Panel Data Models'. It has two units: Introduction to Panel Data (**Unit 8**) and Estimation of Panel Data Models (**Unit 9**). **Unit 8** discusses the concepts of: (i) linear static panel data models, (ii) fixed effects panel data models and (iii) random effect panel data models. **Unit 9** explains the estimation methods for the fixed and random effect panel data models.

Block 4 is on Introduction to Econometric Software Packages. It has three units: **Unit 10** is on Introduction to R which is a open source software i.e. free to download and use. **Unit 11** is on Introduction to EViews and **Unit 12** is on Introduction to Stata. Each of these units provides you with an exposure to: (i) graphical user interface and (ii) performing basic operations for results (e.g. descriptive statistics, tests of significance, regression, etc.). Unit 11 on EViews also exposes you to obtaining results of panel data regression. Likewise, **Unit 12** on Stata especially covers: (i) regression diagnostics and (ii) testing of hypotheses.

ignou
THE PEOPLE'S
UNIVERSITY

BLOCK 1
EMPIRICAL ISSUES OF ECONOMETRIC RESEARCH

INTRODUCTION TO BLOCK 1

Block 1 is on Empirical Issues in Econometric Research. It has **three** units. **Unit 1** is on Stages in Empirical Research. The unit discusses the three important steps requiring to be paid attention in empirical research viz. (i) logic and experience, (ii) economic relationships and (iii) econometric modelling. It also provides a distinction between 'research design' and 'research methods'. A brief review of CLRM is also given in this introductory unit to the course.

Unit 2 is on Specification Issues. It discusses the consequences of three issues viz. (i) omission of relevant variables, (ii) inclusion of irrelevant variables and (iii) errors in measurement. Under errors in measurement, the unit discusses the consequences of errors in measuring the dependent variable and the explanatory variables.

Unit 3 is Model Selection Criteria. It first explains two situations under which tests for specification errors become important. These are: (i) test for presence of irrelevant variables and (ii) test for omitted variables and incorrect functional form. Under 'model selection criteria', the unit discusses two analytical criteria viz. Akaike Information Criterion (AIC) and Schwarz Information Criterion (SIC).

UNIT 1 STAGES IN EMPIRICAL RESEARCH*

Structure

1.0 OBJECTIVES

After reading this unit, you will be able to:

- explain the different steps of empirical research;
- distinguish between quantitative research and qualitative research;
- differentiate between 'research design' and 'research methods'; and
- present an account of common econometric problems faced in regression analysis delineating the measures which controls their consequences.

1.1 INTRODUCTION

From a general perspective, the term 'research' refers to the search for knowledge. More specifically, it is defined as a scientific and systematic enquiry to either discover new facts or verify known facts. Based on the nature of research studies, some of the classified types of research are the following: (i) theoretical and applied research; (ii) descriptive and explanatory research; (iii) quantitative and qualitative research; and (iv) conceptual and empirical research. The present unit introduces you to the stages in 'empirical research'. This means we are dealing with the type of research studies in which collection and analysis of 'quantitative data' is involved.

* Prof. B. S. Prakash, SOSS, IGNOU, New Delhi.

Empirical research can be based merely on techniques of data analysis limited to summary statistics. It becomes econometric when we pay more attention to the unexplained error or disturbance term. However, the planning stage of any empirical research precedes the data collection and analysis stage. One therefore needs to have ample clarity on what data we need to collect and what purpose the data collected should serve after we analyse the same. In other words, we should know what to expect from our empirical study. This requires that we should be guided both by theory as well as by 'methodology and findings' of other studies conducted before. The latter needs to be accomplished by a 'review of literature'. This also helps in identifying the research issues and to decide on what methodology is best feasible for our purpose. The feasibility has to match the researchers' background and training in applying research techniques. The 'findings of other studies' also help us to specify our hypothesis (which is a tentative statement to be tested empirically). It later helps us to 'compare and contrast' our findings with those of the others. Research is a work in continuity i.e. we draw from the work of others and our work would provide feedback for future studies. Hence, the planning stage comprising of review of literature is crucial for gathering all the background information [including 'a theoretical link'] and in giving a suitable 'orientation to our research study' (in terms of methodology). Since our investigation should add (besides confirming or negating the findings of others) to what others have already revealed before, we must spend considerable time of our total study duration (about 25 percent) on this stage of work to attain clarity. Thus, the literature review stage helps us to: (i) 'collect' the background knowledge and information, (ii) 'ascertain' what data is already available at macro level by secondary sources, (iii) 'answer', in broad terms, some of our research questions identified by secondary sources and (iv) 'identify' the specific questions that need to be answered for which conducting a primary survey is needed (i.e. for answering some research questions that cannot be answered by secondary sources). This background also helps us to decide whether any primary survey needs to be conducted and, if so, in which geographical area it should be done. An empirical survey will cost lot of resources in terms of time and money. Hence, prior work preceding the actual survey is important to sharpen our clarity. Against this background, let us now proceed to identify the specific steps or stages involved in an empirical research.

1.2 STEPS IN EMPIRICAL RESEARCH

In empirical research, we are interested in considering 'relationship between variables'. This is in the sense of influence or impact by a set of explanatory variables (also called exogenous variables, independent variables, regressors, or determinants) on the value of explained variable (also called dependent variable, regress and or endogenous variable). In this sense, we are expecting a group of variables to make an impact or influence any one key variable dependent on all of them. Such a causation could be both ways (a situation

referred to as 'simultaneity'). Though we have used the word 'causation' here, the regression analysis does not relate as much with causation as it does with association in terms of relationship between variables. Based on an idea of the functional form of relationship (obtained from a scattered diagram or graph), the form of the relationship (or model hypothesised) could be linear, quadratic, cubic, exponential, etc. One of the common relationships between economic variables considered in empirical research is the Cobb-Douglas model used in the context of production function. This model is exponential in its original form but linear in its logarithmic form. The following three steps or stages are involved in any empirical study.

1.2.1 Logic and Experience

The logical part includes both knowledge from theory as also intuition. For instance, in considering a relationship between 'quantity' and 'price', knowledge of economic theory tells us that we must expect an inverse relationship. This tells us that in the estimated model, the coefficient of 'price' (considered as the independent variable) should have a 'minus' sign. But if we are considering the effect of 'education' on family 'income', we may conjecture that 'more education must have a positive influence on the income earning ability of the individual'. We may therefore expect the estimated coefficient of 'education' (considered as independent variable to influence income) to have a positive sign. This intuition is based on the rationale that one would not ordinarily spend time and money to acquire a certificate or a degree, unless it is 'expected to result' in higher earnings or income. Hence, here, intuition is backed by the expectation for higher income. It is for this reason that a government spends more on 'education and health' as both are 'expected' to result in higher 'social benefit'. In fact, this expectation is empirically borne out by many cross-country regressions run for long term time series data.

When we need to test whether expectations like above are borne out by data collected from the field, we conduct an empirical study or investigation. For this, we need to have knowledge on how to conduct an empirical survey. This needs us to be familiar with a logical reasoning for deciding on the choice of the area in which to conduct the study, the method by which to draw a scientific sample choice of issues (or research questions) to focus upon, etc. The relationship we expect between variables are not always apparent and are quite often very complex. This is because we do not know what exactly is happening at the ground level and therefore are not sure of 'what to expect'. This is also the reason why we may feel a quest to investigate 'specific relationships' with empirically verifiable data. For this, we make conjectures, or 'hypotheses', based on logic or theoretical knowledge or our mere understanding of things. Thus, based on how one construes such relationships, there can be many studies conducted on a single phenomenon (e.g. poverty) in different economic settings (like different area, population groups, etc.). Solutions found useful in one area may not be replicable as it is

in another area or may not be as effective. This is due to inherent changes in different population groups and differences in perception. Hence, based on our experience or empirical findings, it is common to come across different persons explaining the same phenomenon differently. Such plural opinions, based on scientific investigations, forms the core of research studies and provide important feedback for policy making.

1.2.2 Economic Relationships

Empirical studies are helpful in proving or upholding economic relationships suggested by economic theory. It however does not often mean that data collected from a particular economic setting upholds what we believe to hold true based on theory or intuitive logic. For instance, consider an expectation like 'enhanced public spending on vocational education brings down unemployment rate'. In a particular survey, we might find that this expectation is not borne out. This means, other 'determinants' of employment creation are not supportive for which appropriate policy (like monetary or fiscal policy) initiative is required. For instance, for employment generation, either more industries should be set up or the capacity in the existing industries should be expanded. This depends upon whether employers (producers) are able to make fresh investment. Investment, in turn, depends on good infrastructure and availability of credit. If such supporting factors do not supplement an effort like 'enhanced public spending on vocational education', then the expectation of employment creation might not bear out. In such situations, empirical findings could provide policy input for creating conditions conducive to employment creation. They might also give rise to counter or alternative hypotheses. Thus, while empirical studies may not always validate (i.e. confirm) theoretical or logical expectation, it may contribute to generating alternative hypotheses. These need testing for validation with new data. In other words, while many economic relationships are validated through empirical findings, sometimes their negation too could reveal useful insights. Empirical studies are thus an exercise in continuum i.e. studies done by others (which one can glean by literature review) provide feedback for new studies. It is therefore important to always conclude a study with statements of 'assertion' or 'negation' in relation to the findings of other studies. And based on the findings arrived at, policy inputs for course correction must be suggested. An example of continued research and relevance is studies on poverty level in developing economies like India. Even though poverty in India has shrunk from the 50+ percent levels at the time of her independence to a level of below 20 percent now, we still find researchers studying the different dimensions of poverty. In a particular region or district, despite implementation of many poverty alleviation programmes, poverty may continue to be relatively high (i.e. non-responsive) compared to other areas. Since tastes and preferences are endemic to a region, but change over time, specific studies are always required to generate specific policy prescriptions.

1.2.3 Econometric Modelling

Unlike other disciplines, in economics many of the variables could be quantified. This feature of the discipline, besides helping in empirical measurement of economic variables and validation of economic theory, has given a stimulus to application of econometric modelling. There is a distinct department of 'economic statistics' (dealing with 'collection, processing and presentation of data') at every district in India. Using such data, researchers are able to apply techniques of econometric modelling in empirical research. In its classical form, a linear model $Y = X\beta + \varepsilon$ is estimated (by a method like 'least squares' or 'maximum likelihood') using data for past time points. The results of testing for the hypothesis of $\beta = 0$ [i.e. there is no contribution of different factors (or determinants) to Y] provide inputs needed for concluding that X influences Y. If the hypothesis is rejected, then the association or contribution by the X_i's to Y is upheld. Such a conclusion is however limited to the power of the model (indicated by R^2, the coefficient of determination) where giving due caution to other factors (discussed in Section 1.4 of this unit) is important. As an extended use, one can simulate different values of exogenous variables (i.e. X_i) to predict their influence on the endogenous variable Y. In light of this, the important steps involved in undertaking an econometric investigation could be enumerated as follows.

- Formulate the hypotheses based on economic theory or intuitive logic or experience;
- Express the hypothesis as a mathematical equation (with a residual or error term);
- Collect data required either from primary or secondary sources;
- Estimate the parameters by a suitable method;
- Test the hypothesis put forth in the first step above; and
- Interpret the results to indicate the implications of the hypotheses tested.

Check Your Progress 1 [answer within the space given in about 50-100 words]

1) Distinguish 'empirical research' from other types of research.

……

……

……

……

……

……

……

……

2) How is 'logic and experience' important in empirical research?

……

……

……

……

……

3) Is it always necessary for an empirical study to confirm a theoretical assertion? Give example in support of your answer.

……

……

……

……

……

4) State the important steps in an econometric study.

……

……

……

……

……

1.3 RESEARCH METHODOLOGY

The methodological choice is often based on the type of data i.e. quantitative data or qualitative data to be collected for the study. As said before, in empirical research, we are mainly concerned with quantitative data. However, some variables may be qualitative (gender, location) in nature. Even in such cases, data is collected in quantitative terms. For instance, gender may be recorded as zero and 1 (zero for male and 1 for female), level of education recorded as 1/2/3/4 (for primary, middle, secondary and graduate respectively), and so on. In light of this, based on the type of data used in the study, a distinction is often made between 'quantitative research' and 'qualitative research'. In some cases, however, it could be a mixed approach since both quantitative and qualitative variables are present in many investigations.

The objective of quantitative research is to develop mathematical or econometric models. Hence, testing of hypothesis related to the estimated parameters is always an objective of quantitative research. In quantitative research, the researcher and the researched are kept separate i.e. they are assumed to be independent of each other. This is same as saying quantitative

research treats 'cause and effect' as explicit. Further, quantitative research deals with large representative sample whereby a generalisation of conclusions drawn (on the basis of the sample to the population) is usually made. In essence, this means that it assumes that the form of the distribution is known and with increased sample size, the distribution tends to a normal distribution. This also means that single reality, as opposed to multiple realities, is expected to prevail in a population.

In contrast, qualitative research treats cause and effect as non-separable (i.e. implicit). Hence, it treats the researcher and the researched as interdependent. In light of this, it gives importance to the observations made by the researcher in the field. Further, qualitative research usually deals with small and purposely chosen samples. In other words, it deals with situations in which nothing is assumed on the nature of distribution in the population. Qualitative research does not therefore generalise research findings beyond the limits of its immediate context. This means, it believes in multiple realities or co-existence of realities. Qualitative research thus deals with distribution-free contexts where tests based on non-parametric methods are employed. The contrasting features of quantitative and qualitative research is summarised in Table 1.1.

Table 1.1: Distinctive Features of Quantitative and Qualitative Research

Quantitative Research	**Qualitative Research**
Believes in a single reality and the pursuit of identifying universal laws beyond the limit of research/social context.	Believes in multiple realities and their co-existence. Its findings are therefore confined to the immediate context of research only.
Treats 'researcher' and 'researched' as separate or distinct or independent.	Treats 'researcher' and 'researched' as inter-dependent.
Works with large, representative samples.	Works with small and purposely chosen samples.
Considers cause and effect are explicit and hence can be separated.	Believes in non-separation of cause and effect i.e. it is implicit and inseparable.
Considers formal hypothesis testing. Tests are based on estimated parameters i.e. performs parametric tests.	Believes in theory building and thereby generates hypotheses for testing. Employs Non-parametric methods of testing.
Assumes that the distribution is not only known but with sample size increasing the distribution tends to a normal distribution.	Works in distribution-free contexts making no assumptions thereon.

1.3.1 Research Design

Research design is a road map on how to conduct the study or enquiry. It deals with issues like what type of evidence is needed (to answer the research questions posed) and how the same should be collected. We can explain this using the analogy of constructing a building. What type of building is to be constructed (whether a school or a hospital, or residence or office complex) is first needed to be known so that a plan or blue print can be sketched before material for construction is ordered. Likewise, a social scientist should know what research questions to be asked to the respondents and in what way the responses should be recorded clearly. Based on this knowledge, in empirical research, a researcher develops a 'research design' comprising of the following four components:

Sampling design i.e. whether a random sampling technique or non-random sampling procedure should be adopted?

Statistical design i.e. decision on what should be the sample size and the particular method by which the sample should be drawn.

Observational design i.e. what specific instruments should be used for collection of data.

Operational design i.e. the specific details by which the procedures in the above three phases of design are to be carried out.

Thus, a research design deals with both the logical and the logistical aspects of the study including the primary survey aspects like sampling. It is different from 'research methods' which is concerned with the two specific aspects of 'techniques and tools'.

1.3.2 Research Methods

Research methods comprise 'research techniques' and 'tools'. The former refers to the practical aspects of collecting data and the way in which the data collected is organised and analysed. Generally, census and survey methods are used in 'quantitative research'. On the other hand, methods like participant observation, semi-structured interview, life histories, experiments, etc. are used in qualitative research. 'Tools' are the instruments used for collecting the data and its analysis. The tools for collecting the data include: questionnaire, check lists, maps, photographs, drawings, etc. Tools for data analysis include statistical techniques for establishing the relationship or association between different variables (e.g. regression analysis, ANOVA, x^2, etc.) as also in evaluating the accuracy of the results (i.e. by testing procedures like 't', 'F', DW, etc.).

Research methodology is a general term commonly used. It encompasses both 'research methods' and 'research design'. It is important to have a clear idea on these terms because no data can be systematically collected without adequate knowledge on the techniques of data collection. Further, no data can

be explained without a comprehension of the philosophy (or the perspective) behind the characteristics underlying the variables to which the data relates.

Check Your Progress 2 [answer within the space given in about 50-100 words]

1) Distinguish between quantitative research and qualitative research.

……

……

……

……

……

2) Is the methodological approach between quantitative and qualitative research strictly non-overlapping? Give an example to justify your answer.

……

……

……

……

……

3) In what way is the treatment for 'cause and effect' separation distinguishable between quantitative and qualitative research designs?

……

……

……

……

……

4) State the four main components of an empirical research design. What does a 'research design' basically deal with?

……

……

……

……

……

5) How is 'research design' different from 'research methods'? How would you relate the two with the commonly used term 'research methodology'?

……

……

……

……

……

1.4 REVIEW OF CLASSICAL LINEAR REGREESION MODEL (CLRM)

You are aware from your earlier course on Introductory Econometrics that, in empirical research employing econometric models, the estimated parameters of the linear model enjoys the best properties only when certain assumptions made are fulfilled. The best property refers to the estimated parameters being 'minimum variance unbiased' i.e. in the class of all linear unbiased estimators, the ordinary least squares (OLS) estimators have the least variance. It is hence called BLUE (best linear unbiased estimator). The assumptions that need to be fulfilled for the OLS estimates to be BLUE are: (i) the regressors are not highly correlated (i.e. a correlation of the order exceeding 0.80 is not there) to ensure the absence of multicollinearity, (ii) the variance of the error terms are constant (i.e. heteroscedasticity is absent), (iii) the disturbance terms are not correlated (i.e. error terms are independent or there is no effect of serial correlation on them) and (iv) the model specified is free from 'specification errors'. Of these, in this section, we shall make a brief review of 'consequences, detection and treatment' (on which you have already studied in your earlier course BECC 110) of the violation of the first three assumptions. You will study the fourth issue in Unit 2 of this course.

Highly Correlated Regressors: There could be two cases viz. (i) regressors have an exact linear relationship (i.e. a case of perfect collinearity) and (ii) regressors have a linear relationship with a random error term (i.e. a case of less than perfect collinearity). In the former case, the regression coefficients are 'indeterminate'. In the latter case, the coefficients are determinate but have large standard errors (implying that coefficients cannot be estimated with greater precision). The estimated coefficients indicate the 'impact' of X on Y (i.e. of X_i on Y, keeping the impact of other X_is constant). Thus, in a 3-variable regression model with X_1 as the intercept term, if we have a situation like $X_3 = 3X_2$, there is no way to estimate the impact of X_3 on Y keeping X_2 fixed (since both move or change together). Thus, the *consequences* of multicollinearity are: (i) though the OLS estimates are unbiased, they have large variances, (ii) because of large variance, confidence intervals will be wide [implying that there is a high probability of accepting the 'null hypothesis' (of 'zero' coefficient) even when the actual parameter is positive], (iii) regression may appear to do well (with high R^2) despite not being able to reject the hypothesis of one or more parameters being equal to

zero and (iv) the OLS estimators, and their standard errors, can be very sensitive to small changes in data.

For *detection* of multicollinearity, some of the major procedures suggested are: (i) high R^2 with too few significant t-statistics, (ii) high pair-wise correlation exceeding 0.8 (a condition which is sufficient but not necessary) and (iii) considering auxiliary regressions (i.e. R_i^2) and computing $F = R_i^2 \div (1 - R_i^2)$ which is distributed as F with $(k-2)$ and $(n - k + 1)$ d.f. Some of the remedial measures i.e. *treatment* are: (i) pooling of data and (ii) transformation of variables. The later part of this course introduces you to 'panel data estimation' which is pooling of data in cross-sectional and time-series forms. Examples of transformation of variables include: (i) considering the first difference of variables and running a regression on the differenced variables and (ii) ratio transformation [e.g. instead of regressing consumption on GDP and population, we can regress consumption on 'per capita income' (PCI)]. Since both GDP and population grow over time, they are likely to be correlated. This situation can be controlled by considering the PCI even though there is the added risk of having the new error term 'serially correlated'. This means removal of one problem sometimes could give rise to a new problem which was earlier not there. The solution to this situation is to run checks and tests repeatedly till we arrive at a database free from the common econometric problems. Note that since the transformations are made on the original data set uniformly for the entire series, no characteristic of the population (in the original database) is lost. In other words, 'properties of the database in the population is maintained undisturbed'.

Variance of Error Terms is Constant: If the error terms have constant variance, they are said to be homoscedastic. If not i.e. if they are different [Var $(U_i) \neq$ Var (U_j)], it is said to be heteroscedastic. Such situations usually occur in cross-section data. For instance, if we are assessing the savings pattern among the people in an area, variance of high-income families would be larger than that of low-income families. This is because while all poor have to spend on their minimum needs and there will be little difference in their income left to save, among the rich there can be wide difference (since there can be a miser and a spend thrift). Such huge differences arise for reasons like: (i) presence of outliers in sample data and (ii) the scale of variables being widely varied within the sample (e.g. presence of large states and small states with hugely varying GDPs). It is clear that the latter can be controlled by considering the per capita GDP, a measure we had seen above to work for controlling collinearity too. Thus, by taking care to appropriately transform the variables before running a regression, we can control for the consequence of heteroscedasticity in many cases.

Consequences of heteroscedasticity are: (i) the OLS estimates are unbiased but are no longer efficient [i.e. they lose the property of being MVUB (minimum variance unbiased)] and (ii) estimated variances of β_i*'s* will be biased. As a result, the tests of significance by *'t'* and *'F'* would be invalid.

Detection methods for heteroscedasticity include: (i) plotting the residuals against the predicted values of the dependent variable *Y* (if the graph shows a linear pattern i.e. increasing or decreasing trend, it indicates the possibility of heteroscedasticity in the data) and (ii) breaking the sample into two or more sub-samples to compute their error terms and testing them for the hypothesis of 'no difference in their error terms' by the chi-squared test. Note that the chi-squared test requires that the 'error terms are normally and independently distributed'. However, since there may not be a clear-cut basis for the selection of break-point (i.e. it may have to be made arbitrarily), the test is indicative and not conclusive.

Treatment methods include: (i) weighting the original data with the reciprocal of σ^2 when the error variance is known (in which case the transformed error term would be homoscedastic), and (ii) applying a test like White's test or Goldfeld-Quandt (GQ) test or Breusch-Pagan test. However, these tests have their own limitations. For instance, White's test has the disadvantage of the power of the test being often bad besides being sensitive to 'specification bias', the GQ test has the problem of deciding on 'how to omit a certain number of observations' required to apply it, etc.

Error Terms are Independent: The assumption of independence between error terms implies that the covariance between e_i and e_j is 'zero'. Therefore, in the presence of autocorrelation we have: $cov\ (e_i\ ,\ e_j\ ) \neq 0$. In cross-section data, the residual would not be much deviant commonly across all units of a survey. For instance, in determining the level of consumption, it could be high for some units (households) due to a reason like 'expenditure on wedding'. But such a reason would not be commonly affecting all households, since, if it is so, the 'intercept' term would capture it. But in time series data, it is more likely to reflect in a cyclical manner. For instance, when income increases, consumption is not automatically increased. After some time, when the income increase is realised to be sustained, consumption would pick up. This means, the consumption pattern changes gradually to such an extent that at several consecutive data points, the additional spending would show up first with low errors but then for several periods with large errors, the pattern continuing cyclically. This gives rise to the possibility that consecutive error terms will be correlated. Another case where autocorrelation could arise is the use of lagged value of a variable as a regressor. In such situations, in almost all cases, it results in auto-correlated error terms. Thus, one needs to be more cautious while using any kind of time series data for auto-correlation effect on estimated parameters even though even in cross sectional data it can arise.

The consequences of auto-correlation on the OLS estimates are that: (i) they continue to be unbiased and consistent, but (ii) they are no longer efficient. Hence, the variance of the estimators would be larger and thereby the confidence intervals would be large. As a result, the results of tests of hypotheses would be unreliable. More specifically, (i) the variance of the

residual would underestimate the true value of σ^2 resulting in over-estimated R^2, and (ii) the results of usual tests of significance would not be valid. Methods of detecting auto-correlation include: (i) a linear pattern in the residual plot of U_t and $U_{(t-1)}$, which confirms the presence of auto-correlation, (ii) computing the DW test statistic *'d'* (defined as the ratio of the sum of squared differences in successive residuals to the overall 'error sum of squares') to see whether *'d'* is closer to 2, (iii) applying the Breusch-Godfrey test, etc. Method of correcting for auto-correlation includes: (i) obtaining an estimate of the coefficient of autocorrelation (ρ) by applying the Cochrane-Orcutt transformation, (ii) re-arranging the Cochrane-Orcutt equation (with the lagged value of Y taken to the right side) and applying the Durbin's method to run the OLS regression on the rearranged model [wherein the errors in the rearranged model possess some classical property (i.e. consistent though not unbiased)], etc.

Check Your Progress 3 [answer within the space given in about 50-100 words]

1) State the two situations when a situation of 'highly correlated regressors' can prevail. What are their effects on OLS estimators?

……

……

……

……

……

2) In the case of multicollinearity, what is meant by the consequence that 'the OLS estimators are highly sensitive to small changes in data'?

……

……

……

……

……

3) State two methods by which the high collinearity consequences can possibly be averted.

……

……

……

……

……

……

4) State two reasons why heteroscedasticity might be present in data. How can such situation be controlled?

……

……

……

……

……

5) Point out the two consequences of 'heteroscedasticity'

……

……

……

……

……

6) Which of the two types of data i.e. cross-section and time-series commonly pose the problem of auto-correlation? Why?

……

……

……

……

……

7) Interpret the phrase 'properties of the database in the population is maintained undisturbed' with justification.

……

……

……

……

……

1.5 LET US SUM UP

Empirical research deals with quantitative data analysis. An exhaustive literature review is a pre-requisite for any type of research – empirical or otherwise. It helps us in obtaining the background information, building up a theoretical framework, formulate hypothesis, choose an appropriate methodology, etc. Research design is a blue print on all stages of research. Besides specific details on the operational aspects, it in particular includes both the sampling and the statistical designs. This is different from 'research methods' which deals with 'tools and techniques'. Both these i.e. research

design and research methods combined is commonly referred to as 'research methodology'. In addition, research methodology also includes knowledge or perspective of characteristics of variables on which data is collected. Hence, research methodology connotes a much wider and over arching connotation. In econometric studies, we must check the data for three types of commonly afflicted problems which makes the estimates obtained from the OLS methods lose its BLUE properties. There are different methods of detecting and treating these problems. A good knowledge of these methods is essential for an empirical researcher.

1.6 KEY WORDS

Empirical Research : It deals with quantitative data collection and analysis. It seeks to explore relationships between a dependent variable and a set of independent variables.

Research Design : It is a blue print for the whole research study. It covers the sampling, statistical and operational aspects/issues. Both the areas of logic and logistics are covered under research design.

Research Methods : Refers only to the 'tools and techniques' of survey. Tools help us in conducting the survey. Techniques help in analysing the data.

Research Methodology : Is a broad term commonly used. It includes both research design and research methods. The philosophy behind research (i.e. the perspective with which data is collected), the characteristics by which the variables are governed (without a knowledge of which interpretation of data collected is not possible), etc. are also covered under Research Methodology.

Multicollinearity : A situation in which regressors are excessively correlated rendering the OLS estimates lose its best properties.

Heteroscedasticity : A situation where the variance of error term is not common to all the U_is. This means $V(U_i) \neq V(U_j)$.

Autocorrelation : A situation in which $cov\,(U_i, U_j) \neq 0$.

1.7 SUGGESTED BOOKS FOR FURTHER READING

1) Cooper, Donald R. and Pamela S. Schindler (2014). Business Research Methods, Twelfth Edition, McGraw Hill Publication.

2) Research Methodology: Conceptual Foundation (2006). Unit 5, MEC 005, IGNOU, ISBN: 81-266-2641-0.

1.8 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) First, it is based on collection and analysis of quantitative data. Second, it concerns itself with an identification of relationship between the variables.

2) Logic includes both theory and intuition. Theory helps us formulate hypothesis, giving also an idea on what to expect from the results of the study. Intuition gives us clarity on what to expect.

3) No. Such studies give the much needed input for policy reorientation. They are also helpful in generating alternative hypotheses.

4) Formulation of hypothesis based on theory or intuitive logic, stating the hypothesis in the form of an equation, collection of data, estimation of parameters, testing of hypothesis and interpretation of results.

Check Your Progress 2

1) In quantitative research, data collected on variables is in numbers. Even for qualitative variables, the response is recorded in numbers. This makes the methodology amenable for application of statistical or econometric techniques. As a result, estimated parameters of the model are tested. In contrast, in qualitative research, data collected are descriptive in nature. Methods of data collection in quantitative research use questionnaires whereas for qualitative research they are unstructured or semi-structured interviews, participant observation, life histories, etc.

2) No. Often, a mixed approach is adopted. For instance, even when questionnaires are used for collection of data, response to closed-ended questions could be quantitative but for open ended questions, the response would be qualitative.

3) Quantitative research treats ‘cause and effect’ as explicit and hence can be separated. In contrast, qualitative research treats it as implicit and hence non-separable.

4) Sampling design, statistical design, observational design and operational design. It deals with issues like type of evidence or data and the methods of its collection.

5) A ‘research design’, in general, deals with both the logical and the logistical aspects of a research study. ‘Research methods’, on the other hand, deals specifically with ‘tools and techniques’ of data collection and

data analysis. Research Methodology = Research Design + Research Methods.

Check Your Progress 3

1) Perfect multicollinearity and less than perfect multicollinearity. In the case of the first situation, OLS estimators are indeterminate. In the case of the second situation, estimators are determinate but they have large standard errors rendering the results of usual tests, t and F, suspect or unreliable.

2) It means that if we drop a variable (where there are say only two regressors like income and wealth), the situation might change (in the sense of the other variable registering significance when earlier neither of the two were registering significance). Alternatively, it could be a small sample problem i.e. small sample size could be posing a situation of higher R^2 but none of the independent variables significant.

3) Pooling of data and transformation of variables.

4) Presence of outliers and the scale of variables being widely varied within the sample (like the presence of a large/rich state and a poor/small state in the sample). A transformation of variable like taking the ratio (e.g. PCI in place of GDP) would solve the problem in many cases.

5) (i) OLS estimators are unbiased but not efficient and (ii) estimated variances of coefficients are biased and hence the results of testing by t and F would be unreliable.

6) Time-series data. Because of a cyclical or lagged effect observed commonly in time series data.

7) This is in the sense of adding, subtracting, multiplying or dividing (or standardising as done in change of origin and scale) by a common factor or performing the process on both sides of an equation or an entire series of data.

UNIT 2 SPECIFICATION ISSUES*

Structure

2.0 OBJECTIVES

After reading this unit, you will be able to:

- state the factors which need to be considered while specifying a model for econometric analysis;
- delineate, theoretically, the consequences of omitting relevant variables in an exercise of econometric modelling;
- explain why it is more important to accord due attention to the underlying theoretical considerations in specifying a model for empirical investigation;
- write a note on the useful 'indicators' in a 'regression output';
- show why it is better to err by including an 'irrelevant variable' as compared to omitting a 'relevant variable' in a regression model;
- outline why the consequences of 'errors in measurement of dependent variable' is less serious as compared to that in the 'independent variables'; and
- discuss the consequences of 'errors in measurement' in 'independent variables'.

2.1 INTRODUCTION

Model specification refers to the very beginning of the process of developing a regression model. Here, we decide which variables should be included for empirical investigation, which of these are justified to be treated as 'independent variables' [to appear on the 'right hand side' (RHS) of the equation (or the regression model)], whether the nature of these variables is

* Rimpy Kaushal, PGDAV College, Delhi.

quantitative or qualitative, etc. We also decide on the 'dependent variable' which appears on the 'left hand side' (LHS) of the equation. In this process, while we technically assume that the regression model is correctly specified, in practice, an exact specification of the model is difficult. Though economic theory helps us in deciding on the specification of regression model, theory can itself be questioned or may prove ambiguous to select suitable variables. Therefore, unknowingly, we commit specification errors like: omitting a variable from the model that should be included, including an irrelevant variable in the model, miss-specification of the functional form of the model or make errors of measurement. Thus, specification errors may occur at two stages viz. (i) when the functional form considered (vis-à-vis the explanatory variables to be included in the regression model) is not close to the true relationship in the population or (ii) when errors are committed while measuring the variables.

We are aware that the properties of the estimated regression coefficients closely depend on the validity of the assumption that there is no 'specification error' in the model. Therefore, if we leave out a variable that is crucial, the estimated regression coefficients would be biased. In this case, the estimated standard errors, the confidence intervals for the estimated coefficients and hence the computed 'test statistic' itself would be incorrect (or far from the unknown true value in the population). As a result, the results of the hypothesis testing would not be reliable. On the other hand, if we include an irrelevant variable in the model, the regression estimates could be unbiased but yet inefficient (i.e. the standard errors and confidence intervals could be unduly large). The objective of the present unit is to learn about the consequences of various types of mis-specification in an estimated regression model and more importantly how to avoid making such errors.

2.2 OMISSION OF RELEVANT VARIABLES

Quite often, unknowingly, we may omit a relevant explanatory variable from the regression model. Consider a regression model in which the dependent variable Y is actually related to two variables X_2 and X_3 ,like:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u_i \qquad (2.1)$$

Not knowing this, suppose we wrongly construe and specify the model as:

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \qquad (2.2)$$

We know that the estimate of β_2, $\hat{\beta}_2$ or the OLS estimator for β_2, can be obtained from Equation 2.2. as:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)Y_i}{\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2} \qquad (2.3)$$

Substituting for Y_i from (2.1) in (2.3) we get:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)\left(\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i\right)}{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2}$$

$$= \frac{\beta_1 \sum_{i=1}^{n}(X_{2i} - \bar{X}_2) + \beta_2 \sum_{i=1}^{n}(X_{2i} - \bar{X}_2) X_{2i} + \beta_3 \sum_{i=1}^{n}(X_{2i} - \bar{X}_2) X_{3i} + \sum_{i=1}^{n}(X_{2i} - \bar{X}_2) u_i}{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2}$$

Note that $\sum_{i=1}^{n}(X_{2i} - \bar{X}_2) = 0$ and

$$\sum_{i=1}^{n}(X_{2i} - \bar{X}_2) X_{2i} = \sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2 \qquad (2.3^*)$$

(see exercise 4, CYP 1, for this step)

Hence $\hat{\beta}_2 = \beta_2 + \frac{\beta_3 \sum_{i=1}^{n}(X_{2i} - \bar{X}_2) X_{3i} + \sum_{i=1}^{n}(X_{2i} - \bar{X}_2) u_i}{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2}$ (2.4)

Taking expectations on both sides of (2.4) we get:

$$E(\hat{\beta}_2) = \beta_2 + \frac{\beta_3 \sum_{i=1}^{n}(X_{2i} - \bar{X}_2) X_{3i}}{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2}$$

Using the assumption from the true model that $E(u|X) = 0$

$$E(\hat{\beta}_2) = \beta_2 + \beta_3 \hat{\delta}_2 \qquad (2.5)$$

where $\hat{\delta}_2 = \frac{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2} = \frac{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2) X_{3i}}{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2}$..

Therefore, the expectation of $\hat{\beta}_2$ is not equal to β_2 or there is a bias in the estimate called *Omitted Variable Bias* that depends on β_3 i.e. the effect of the excluded variable X_3 on the dependent variable Y with δ_2 representing the effect of explanatory variable X_2 included In the omitted variable X_3. Other similar consequences of omitting a variable can be stated as the following.

i) By omitting a variable that is correlated to the dependent variable, and to any of the explanatory variables, the estimates of the regression coefficients will be biased. The nature of the bias depends on the nature of correlations between: (a) the dependent variable and omitted variables and (b) independent variables and omitted variables. If the true regression model is as in (2.1), in which X_3 is the omitted variable, the nature of bias can be summarised as below.

Relationship	**X_2 and X_3 negatively correlated**	**X_2 and X_3 positively correlated**
Y and X_3 negatively correlated	β_3 *is overestimated*	β_3*is underestimated*
Y and X_3 positively correlated	β_3 *is underestimated*	β_3*is overestimated*

ii) The OLS estimates will also be inconsistent to the extent that even in large samples, the estimates would remain biased.

iii) If X_2 and X_3 are not related in the sample, the value of $\hat{\delta}_2$ will be zero and the estimate for slope coefficient will be unbiased as well as consistent. However, the estimate for intercept would remain biased, unless the mean of X_3 is zero.

iv) The error variance estimated from the mis-specified model will also be a biased estimator of the true error variance σ^2 .Consequently, the estimates for the variance of slope coefficient will also be biased and the variance of the estimated slope coefficient $\hat{\beta}_2$ of the mis-specified model will be overestimated.

v) As a result of all these, the confidence interval and the result of hypothesis testing will be seriously compromised to the extent that they are unreliable.

2.2.1 Specification Errors: Illustration

The above explanation on the consequences of 'specification errors' is theoretical. We can understand this better with the help of an illustration. Consider a 'consumer expenditure survey' conducted for a sample of 6334 individuals. We can take data on 'log of annual expenditure on food' (i.e. LGAEOF) as the dependent variable *Y*. We might consider regressing LGAEOF on two explanatory variables viz. (i) log of total annual household expenditure (LGTAHEXP: X_1) and (ii) log of the number of number of persons in the household (LGNOP: X_2). Let us consider the results of two separate regressions: one regressing *Y* on only X_1 (an under-specified model: Model 1) and the second by regressing *Y* on both X_1 and X_2 (Model 2). Note that by taking logarithmic values of both the dependent and the 'independent variables', we are focusing on investigating the impact of the relative changes in X_i over *Y* and the regression coefficients gives us a measure of 'elasticity'. For four present purpose, we consider the estimated values of the two regressions as in Table 2.1.

Table 2.1: Consequence of Specification Error

Dependant Variable: LGAEOF Intercept/ Variables	**Results of Regression Estimation**					
	Model 1			**Model 2**		
	Value	**S. E. (β)**	**t-Ratio**	**Value (β)**	**S. E.**	**t-Ratio**
Intercept	0.70	0.08	8.3	1.2	0.08	14.1
LGTAHEXP (X_1)	0.67	0.01	68.7	0.58	0.01	60.1
LGNOP (X_2)	-	-	-	0.33	0.01	26.2
Adjusted R^2	0.43			0.48		

A standard regression output generated by common software usually presents a number of values. For our present purpose, in this sub-section, we shall confine ourselves to the three basic values viz. estimated value of the

coefficients of 'independent/explanatory variables', the S. E. (standard error) of the estimated value and the value of *t*-statistic (which is a ratio of 'estimated value' and the S. E. i.e. *t* = value ÷ S.E.). For easy comprehension, the values are presented up to one or two decimal points in Table 2.1.

Clearly, in an empirical exercise on estimating the annual expenditure on food, the number of persons in a family is an important explanatory variable (as more persons would mean higher expenditure on food and vice versa). From this angle, Model 2 ought to be a better specified model. Let us examine whether this is actually the case or not? You may observe that 'despite the standard error of both the intercept term and X_1 being the same' (0.08 and 0.01 respectively), there is considerable deviation in the estimated values of both the terms. Specifically, in Model 2, the intercept term is over-estimated while LGTAHEXP is under-estimated. In other words, there is an upward bias in the estimate for the intercept term and a downward bias in that for LGTAHEXP. Further, the overall explanatory power of the Model 2 is only marginally higher by 0.05. The illustration therefore underscores: (i) the importance of carefully specifying a regression model and (ii) the need to consider other variables which too might influence four dependent variable. For instance, we can first calculate the average expenditure on food in a household and then take its logarithmic value as *Y*. Taking such ratios of per capita values quite often makes for a better specification of a Model. You are aware from your study of the course on Introductory Econometrics that many times such 'ratio transformation' helps us in controlling for the problem of multicollinearity in the original variables. The low value of R^2 in the illustration above is indicative of the fact that there are possibly other variables (e.g. total income level of the household) which are important to be included in the model. In other words, understanding the underlying theory is more important in specifying a model for empirical investigation. We must however also note that inclusion of many explanatory variables can introduce 'over-specification errors'. We shall study on this in the subsequent section of this unit. Hence, due care is required to be taken for a judicious choice of explanatory variables in the model. The example considered here is of a cross-section data in which the sample size is usually large. However, when we consider time series data (e.g. annual time series data), our sample size (*n*) will be much smaller. In such cases, we know that as a general rule, '*k*' (the number of parameters estimated) should be less than '$n + k - 1$'.

2.2.2 Regression Output

A standard regression output presented by any software package presents the results up to many decimal points. Besides, the values of estimated value of parameters or coefficient of regression, their standard error and *t*-ratio (both for the intercept term and each one of the independent variables), many other summary statistics are also presented. Specifically, these relate to: (i) mean of the dependent variable, (ii) standard deviation of the dependent variable, (iii) sum of squares of residuals [$\sum e_i^2$ where $e_i = Y_i - \hat{Y}_i$], (iv) standard error of

regression [i.e. (σ_u)], (v) R-square and adjusted R-square, (vi) F-value and P-value of F that indicates the joint significance of the regression model, (vii) log-likelihood value (which indicates fitness of the model such that a model with higher value of log likelihood is preferred,), (viii) Akaike criterion, (ix) Schwarz criterion and (x) Hann-Quinn criterion (which are all criteria used for model selection). An illustration of the results from a standard regression analysis is presented in Table 2.2. Let us now learn the usefulness of some of these values and criterion here.

Table 2.2: Illustration of the Results from a Standard Regression Output

Ordinary Least Square
Number of observation: 6334
Dependent variable: LGAEOF

	Coefficient	Std. Error	t-ratio	p-value
Constant	1.15833	0.0820119	14.12	0.0000***
LGEXP	0.584210	0.00971737	60.12	0.0000***
LGSIZE	0.334348	0.0127587	26.21	0.0000***
Mean dependent variable		6.474297	S.D. dependent var	0.779391
Sum squared residuals		1988.365	S.E. of regression	0.560418
R-squared		0.483136	Adjusted R-squared	0.482973
$F(2, 6331)$		2958.936	P-value(F)	0.000000
Log-likelihood		−5318.209	Akaike criterion	10642.42
Schwarz criterion		10662.68	Hannan-Quinn	10649.43

Check Your Progress 1 [answer within the space given in about 50-100 words]

1) State the four types of specification errors.

……

……

……

……

……

2) What are the consequences of omitting an important variable from inclusion in a regression model?

……

……

……

3) State the consequences of including an irrelevant variable in a regression model.

……

……

……

……

……

4) In Equation (2.3*), show that: $\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)\, X_{2i} = \sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2$.

……

……

……

……

……

5) In what way, taking 'ratio transformation' is helpful in empirical investigation?

……

……

……

……

……

2.3 INCLUSION OF IRRELEVANT VARIABLES

Sometimes, in order to avoid the consequences of omitting a relevant variable, we may include some variables even though the theoretical justification may be lacking. The rationale behind this approach is that over-specification of a model (i.e. inclusion of unnecessary variables not justified by theory), does not harm the basic properties of the model. In other words, (i) the regression estimates still remain unbiased and consistent and (ii) standard confidence intervals and hypothesis testing also remain valid. For a theoretical account of this fact, consider a simple two-variable model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \qquad (2.6)$$

While we assume (2.6) to be correctly specified, let us say, instead of (2.6), we proceed to estimate the following regression equation by including X_3:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u_i \qquad (2.7)$$

where X_3 is not a relevant variable with the theoretical backing of any relationship between Y and X_3 . The consequences of committing this specification error are the following.

i) The OLS estimators of β_1 and β_2 in (2.7) are unbiased [i.e. $E(\hat{\beta}_1) = \beta_1$ and $(\hat{\beta}_2) = \beta_2$]. They are also consistent.

ii) The confidence interval and hypothesis-testing procedure are valid.

iii) The estimator for variance of u_i i.e. σ_u^2 is unbiased and correctly estimated.

However, a negative consequence of (2.7) is that the OLS estimators are inefficient since the variance of $\hat{\beta}_2$ in (2.7) will be larger than that of (2.6). To verify this, consider the variance of $\hat{\beta}_2$ of (2.6):

$$\sigma_{\hat{\beta}_2}^2 = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2}$$

The variance of OLS estimate of β_2 estimated from (2.7) is:

$$\sigma_{\hat{\beta}_2}^2 = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2} \frac{1}{\left(1 - r_{X_2X_3}^2\right)}$$

Thus, the difference between two variances will be large depending on how large and close to 1 or – 1 is the correlation coefficient $r_{X_2X_3}$. This difference will be zero if the above correlation coefficient is also equal to zero. In this case, the variance of estimator of β_2 from both (2.6) and (2.7) will be identical. We can illustrate this by considering the regression result of LGAEOF (i.e. logarithm of total annual household expenditure on food) on LGTAHEXP (i.e. logarithm of total annual household total expenditure) and LGNOP (i.e. logarithm of number of persons in the household) [from the data collected for the 6334 households in the consumer expenditure survey considered in sub-section 2.2.1]. Recall that this model was assumed to be specified correctly. Now, if we include another variable LGHOUS (logarithm of annual expenditure on housing services) without the theoretical justification for its inclusion, we get the following result (Table 2.3). It is just by chance that the coefficient of the variable LGHOUS is statistically significant. Further, despite the inclusion of this variable, although the estimates of the slope coefficient remain unbiased (since the coefficients of LGTAHEXP from the two regressions are not much apart: being 0.58 and 0.64 respectively), their standard errors have increased (from 0.0097 to 0.0126) leading to a loss in efficiency. Note that this is not the case for LGNOP since both the slope coefficient and their 'standard errors' (SEs) are close (i.e. 0.33 & 0.32 and 0.0128 & 0.0129 respectively).

Table 2.3: Results of Regression by Including an Additional Variable

Ordinary Least Square
Number of observations 1-6334 (n = 6223)
Dependent variable: LGAEOF

	Coefficient	Std. error	t-ratio	p-value
Constant	1.04448	0.0839901	12.44	4.36e-35 ***
LGTAHEXP	0.63554	0.0126350	50.30	0.0000 ***
LGHOUS	− 0.0474	0.00803035	−5.897	3.89e-09 ***
LGNOP	0.324457	0.0128652	25.22	1.05e-133 ***

Mean of dependent variable 6.478563 S.D. of dependent variable 0.777861
Sum square of residuals 1935.247 S.E. of regression 0.557838
R-squared value 0.485953 Adjusted R-squared 0.485705
F(3, 6219) 1959.707 *P*-value(*F*) 0.000000

If we omit a variable that is relevant for the model, then the estimates of regression coefficients become biased, inconsistent and inefficient with the result that the usual hypothesis testing procedures (based on *t* and *F*-test) becomes invalid. In other words, the estimates of the model lose their relevance. On the other hand, if we include an irrelevant or unnecessary variable, not only the OLS estimators still remain unbiased and consistent, the hypothesis testing procedures remain valid. However, the efficiency of the estimates of regression coefficients gets highly compromised in the sense that larger variances lead to wider confidence intervals. As a result, in some cases, we may fail to reject the null hypothesis of no significance. We can, therefore, conclude that it is better to include irrelevant variables than to omit a relevant variable. But this approach should not be stretched as there is a cost for such inclusion in terms of both loss in efficiency and the degree of freedom. The best approach is to include only those variables that are theoretically justified.

2.4 ERRORS OF MEASUREMENT

Very often, it happens that while investigating relationship between variables in economics, the variables involved are not measured correctly. Most of the time, macroeconomic data on variables such as gross domestic product, inflation, etc. are measured through sample and hence tend to be approximations. Even microeconomic surveys are based on information collected from individual units and thus might have been measured inaccurately. While a variable is defined in a certain way, the data available through secondary sources may not exactly correspond to such a definition. Thus, in practice, there can be several reasons for errors of measurement in the variables. These could be grouped under: (i) errors in reporting, (ii) missing observations or (iii) human errors. Whatever may be reason for such errors, these errors cause specification errors leading to serious consequences.

2.4.1 Errors of Measurement in Dependent Variable

In case of errors in measuring the dependent variable, the consequences can be thought of as being accounted for in the stochastic term included in the model. Consequently, the model tend to become imprecise i.e. it leads to a loss in the precision of the regression estimates. However, the estimates will remain unbiased. Let us consider the true value of the dependent variable *Y* to be *Z* and its relationship with X_{2i} can be expressed as:

$$Z_i = \beta_1 + \beta_2 X_{2i} + v_i \qquad (2.8)$$

Since Y is the value that is actually sought to be measured empirically, with say η_i as the measurement error, we have:

$$Y_i = Z_i + \eta_i \text{ or } Z_i = Y_i - \eta_i$$

Hence, (2.8) can be re-written as:

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \qquad (2.9)$$

where $u_i = v_i + \eta_i$. Note that (2.9) is different from (2.8) in the sense that the error term u_i has two components: (i) the error term from the original model (v_i) and (ii) the error of measurement (η_i). Since the explanatory variable remain unaffected, the OLS estimates of the regression coefficients remain unbiased [as $E(\hat{\beta}_1) = \beta_1$ and $E(\hat{\beta}_2) = \beta_2$] provided the regressors are non-stochastic. However, there will be a larger variance of the OLS estimates. Specifically, the variance of the slope coefficient will be:

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2} = \frac{\sigma_v^2 + \sigma_\eta^2}{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2}$$

which is larger than the variance when there is no error of measurement of dependent variable. Thus, the consequences of errors in measurement of dependent variable are: (i) the OLS estimates remain unbiased, (ii) the variances of the estimators are also unbiased, but, (iii) the variances of the estimators will be larger, leading to loss in precision. In sum, we may therefore conclude that the errors of measurement in dependent variable do not matter much in practice i.e. it is less serious. Note that this is in a relative sense as the next sub-section (2.4.2) shows.

2.4.2 Errors of Measurement in Explanatory Variables

Unlike the errors of measurement in dependent variable, the errors of measurement in explanatory variables of the model is more serious in nature. This is because, estimators of regression coefficients remains neither unbiased nor consistent. Suppose the true relationship between Y and X' is like:

$$Y_i = \beta_1 + \beta_2 X'_{2i} + v_i \qquad (2.10)$$

where the disturbance term v_i is distributed independently of X' with zeromean and variance σ_v^2. If we assume that X is the inaccurately measured value of X' in which ω_i as the measurement error, then we can write it as:

$$X_i = X'_i + \omega_i \qquad (2.11)$$

Now, if ω_i is also independently distributed of X' and has zero mean and σ_ω^2 variance, we would have $Cov(X'_i, \omega_i) = 0$ and $Cov(v_i, \omega_i) = 0$. Substituting (2.11) in (2.10), we get:

$$Y_i = \beta_1 + \beta_2(X_i - \omega_i) + v_i = \beta_1 + \beta_2 X_i + v_i - \beta_2 \omega_i \qquad (2.12)$$

In (2.12) there are two random components: (i) disturbance term from the original model (v_i) and (ii) the measurement error (ω_i) multiplied by $-\beta_2$.

Indicating the composite disturbance term $(v_i - \beta_2\omega_i)$ asu_i, (2.12) can be re-written as:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad (2.13)$$

Now, for estimating the parameters of (2.13), if we assume that there is no systemic association between X_i and u_i, we would have $Cov(X_i, u_i) = 0$. By assuming further that the disturbance term of (2.10), satisfies the assumptions of classical linear regression, we have:

$$Cov(X_i, u_i) = Cov(-\beta_2\omega_i, \omega_i) = -\beta_2\sigma^2{}_\omega \qquad (2.14)$$

Thus, the CLRM assumption of no systemic association between X_i and u_i, is violated with the consequence that $\hat{\beta}_2$ becomes a biased and inconsistent estimator of β_2. The consequences of the measurement errors in explanatory variables may therefore be summarised as: (i) the OLS estimators are biased, (ii) the estimators are inconsistent with the bias increasing with increase in the sample size. Thus, the error of measurement in explanatory variable is more serious than that in the dependent variable. If the errors of measurement happen in both the dependent as well as the explanatory variables, the problem would be even more serious.

Check Your Progress 2 [answer within the space given in about 50-100 words]

1) State the two basic properties of a regression model.

.....

.....

.....

.....

.....

.....

.....

2) What is the consequence of omitting a relevant explanatory variable in a regression model?

.....

.....

.....

.....

3) State the consequence of inclusion of a irrelevant explanatory variable in a regression model?

.....

.....

.....

.....

.....

4) What are the broad reasons due to which we commonly encounter errors of measurement in variables in economics? What are their consequences?

.....

.....

.....

.....

.....

5) State the consequences of 'measurement errors' in the dependent variable.

.....

.....

.....

.....

.....

6) State the consequences of 'measurement errors' in the independent variables.

.....

.....

.....

.....

2.5 LET US SUM UP

The CLRM assumes that regression model is '*correctly specified'*. The term '*correctly specified'* means that all the theoretically relevant variables are included in the model, irrelevant variables are excluded and there are no errors of measurement. If an important explanatory variable is excluded from the model, the coefficient of the model becomes not only biased but also inconsistent and inefficient. As a result, the hypothesis-testing procedure become invalid. On the other hand, if an irrelevant variable is included in the model, the estimated coefficients remain unbiased and consistent but there will be a loss in the precision of the estimators (since the standard errors of the coefficients will be larger). In case of errors in measurement of variables, while the measurement error in dependent variable is not very serious, the

same in the explanatory variables are more serious [since it destroys the properties of OLS estimators (viz. unbiasedness, consistency and efficiency)].

2.6 KEY WORDS

Relevant Variable : A variable with the theoretical justification for inclusion in the model.

Irrelevant Variable : A variable without the theoretical backing but of which the researcher is unsure and hence prefers to have it included in the regression model. Such inclusion, called over-specification, does not harm the basic properties of the model. This means that the estimates of regression co-efficients are on biased and consistent so that the results of hypothesis testing remain valid.

Omission of a Relevant Variable : The consequences of this are: (i) estimated slope coefficients will be biased, (ii) OLS estimates are inconsistent so much so that even large samples would not eliminate this and (iii) both the confidence interval and the result of hypothesis testing are seriously compromised.

Inclusion of Irrelevant Variable : This results in a situation which, in comparative terms, is far less serious than the omission of a relevant variable. This is because of the sustenance of the three basic properties viz. unbiasedness, consistency and validation of test results.

Measurement Error in Dependent Variable : This is situation in which both the estimates of coefficients and the variance estimates remains unbiased. The latter will be however larger leading to loss in precision. Hence, viewed relatively less serious.

Measurement Error in Independent Variables : This is a situation where estimators would be biased and inconsistent with the extent of bias increasing with increase in sample size. Hence, viewed relatively, this is more serious.

2.7 REFERENCES

1) Gujarati, D. N. and Porter, D. C. (2010). *Essentials of Econometrics,* Fourth Edition), McGraw Hill.

2) Dougherty, C. (2011). *Introduction to Econometrics*, Oxford University Press.

3) Gujarati, D. N. and Porter, D. C. (2009). *Basic Econometrics* (Fifth Edition), McGraw Hill.

4) Wooldridge, J. M. (2009). *Econometrics*, Cengage Learning.

2.8 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) (i) omitting a variable from the model that should be included, (ii) including an irrelevant variable in the model, (iii) miss-specification of the functional form of the model and (iv) errors of measurement.

2) The estimated regression coefficients would be biased and their standard errors (and hence the confidence intervals for the estimated coefficients) would be wide. As a consequence, the computed 'test statistic' would be incorrect and the results of the hypothesis testing would be unreliable.

3) The estimates of regression coefficients could be unbiased but inefficient (i.e. the standard errors and confidence intervals could be unduly large).

4) RHS = $\sum_{i=1}^{n}(X_{2i} - \bar{X_2})\,(X_{2i} - \bar{X_2})$

$= \sum [X_{2i}\big((X_{2i} - \bar{X_2})\big] - \bar{X_2}\sum(X_{2i} - \bar{X_2})]$.

The second term is 'zero' because the term within brackets is 'zero'.

5) It helps in reducing or eliminating the collinearity effect sometimes.

Check Your Progress 2

1) (i) the estimates of regression coefficients should be unbiased and consistent; (ii) confidence interval should be valid so that the results of the hypothesis testing too are valid.

2) The estimates of regression coefficients will be biased, inconsistent and inefficient. As a result, the results of the hypothesis testing (based on *t* and *F* tests) will be invalid.

3) The OLS estimators would be unbiased and consistent. The results of hypothesis testing too remain valid. However, efficiency is compromised with larger variances of the estimates and hence wider confidence intervals.

4) (i) errors in reporting, (ii) missing observations or (iii) human errors. The consequence is that they lead to specification errors with their attendant consequences on estimates and results of testing.

5) (i) OLS estimates remain unbiased, (ii) variances of the estimators are also unbiased, but the variances of the estimators will be larger, leading to loss in precision.

6) Estimators of regression coefficients are neither unbiased nor consistent. Moreover, the magnitude of bias increases with increase in sample size.

Hence, the consequence of errors in measuring explanatory variables is more serious than that of error in measuring the dependent variable.

UNIT 3 MODEL SELECTION CRITERIA*

Structure

3.0 OBJECTIVES

After reading this unit, you will be able to:

- specify the criteria for formulating a regression model;
- state the particulars of test for detecting the presence of 'irrelevant variables';
- discuss the Ramsey's Test (RESET) for identification of 'omitted variables' and 'incorrect functional form';
- distinguish between the terms 'in-sample forecast' and 'out-of-sample forecast';
- outline how R^2 and adjusted R^2 serve as indicators of 'goodness of fit' of a regression model;
- differentiate between the 'Akaike Information Criterion' (AIC) and the 'Schwarz Information Criterion' (SIC) commenting on their usefulness in forecasting the performance of a regression model; and
- illustrate the model selection procedure using the various criteria.

*Rimpy Kaushal, PGDAV College, Delhi.

3.1 INTRODUCTION

One of the assumptions of 'classical linear regression model' (CLRM) is that the regression model is correctly specified. If the regression model is not correctly specified, we encounter the problems of model specification errors. We have studied the consequences of mis-specification of a model in the previous unit (Unit 2).However, if we take a closer look at the assumption of no specification errors, we find that specifying a true model for a given dataset is near to impossible. Therefore, several model selection criterion are suggested in theory. Hendry and Richard (1983) suggest six criteria that should be met while formulating a regression model. These are:

1) **Data Admissibility:** Model predictions are consistent with the data.
2) **Theoretical Consistency:** Model specified is consistent with the existing theory.
3) **Exogenous Regressors:** Regressors are uncorrelated with the error term.
4) **Parameter Constancy:** Estimates of the parameters are stable.
5) **Data Coherency:** The residuals from the model are purely random (i.e. white noise).
6) **Encompassing:** The model is able to explain results from other rival models (in other words, no other model is better than the chosen model).

However, the criteria given above lay down only a theoretical framework. In practice, very often, we commit errors in model specification.

3.2 TESTS FOR SPECIFICATION ERRORS

Let us recall the different types of model specification errors (and their consequences) studied in Unit 2 on 'Specification Issues'. No researcher knowingly commits specification errors. Specification errors arise inadvertently due to researcher's inability to develop the regression model as meticulously as required. There can be several reasons for such specification errors like weak theoretical background, unavailability of adequate data, etc. Therefore, in practice, researchers emphasise on the detection of the specification errors instead of finding out the reasons for specification errors. In this section, we discuss some tests that are helpful in detecting the specification errors.

3.2.1 Test for Presence of Irrelevant Variables

Let us consider a *k*-variable regression model as follows:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \ldots\ldots\ldots\ldots\ldots + \beta_k X_{ki} + u_i \qquad (3.1)$$

In order to test whether a variable, say X_{ki} add real explanation to the model (i.e. it is relevant to the model specified), we test the significance of estimated β_k by the usual *t*-test. Likewise, for testing the relevance of a set of

variables in the model (i.e. all independent variables taken together), we apply the *F*-test. In other words, the detection of irrelevant variables in the model is tested by the usual *t* and *F*-test. But it is important to note that these tests of significance are carried out under the assumption of 'true specification' of the regression model. In view of this, this approach needs to be adopted along with a process called as 'data mining' which is but a process of diagnostic procedures for developing or arriving at a good model. Data mining is a term used to describe the process applied to extract useful data (satisfying the assumptions made for the CLRM) from the raw data. It basically means detecting and removing errors in data set arising from violation of assumptions of CLRM. The primary objective of data mining is therefore to develop a model after conducting several diagnostic tests to finally lead to a regression model that fits the data well.

3.2.2 Test for Omitted Variables and Incorrect Functional Form: Ramsey's Test

A researcher can never be sure that a regression model formulated is the 'true or best' model for empirical investigation. It is only the theoretical framework and prior empirical studies that helps a researcher. These help in designing a model that is assumed to truly reflect the population characteristics sought to be revealed by the regression model estimated on the basis of sample data. In practice, it is only after the design, that a model is subjected to empirical investigation. In other words, it is only after the stage of specification of the model, that the model is tested for its adequacy by the data collected. This is done by examining carefully the broad features of the empirical results such as: (i) the value of coefficient of determination (R^2), (ii) the value of the adjusted R^2, (iii) significance of *t*-ratios estimated, (iv) results of *F*-test, (v) signs of the estimated coefficients, (vi) value of Durbin-Watson test statistic to reveal the presence or absence of serial correlation effect, etc. If the chosen model performs reasonably well in terms of broad features specified above, then the model is considered to be a fair representation of the true relationship in the population. However, if the model fails to satisfy one or the other broad feature, the researcher would have to suspect some specification error. This might be omitted variable bias, wrong functional form, presence of serial correlation, etc. Thus, in order to determine whether the model suffers from one or more specification errors, one can adopt several methods (i.e. diagnosis for detection and treatment procedures). These steps help in finally arriving at a 'cleaned data set' from which the results of tests drawn would be realistically revealing the population characteristics.

Examination of the residual series is a good diagnostic tool helpful in detecting the presence of serial-correlation and heteroscedasticity in the data. The residual examination is also helpful in detecting specification errors in cross-sectional data. This is because in the presence of specification errors, residual series display noticeable patterns. This can be illustrated by the

'cubic total cost function' as follows. Let us begin by assuming the true 'total cost function' as:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_i \quad (3.2)$$

where *Y*= total cost and *X*= output. Let us assume that a researcher fits a quadratic model, ignoring the cubic term, as:

$$Y_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^2 + u_i \quad (3.3)$$

And another researcher considers only a linear relationship between *Y* and *X*, ignoring even the quadratic term, as:

$$Y_i = \gamma_1 + \gamma_2 X_i + u_i \quad (3.4)$$

For illustration, we consider the results of regression estimates for the three models, 3.2 to 3.4, drawn from a secondary source (Gujarati, Basic Econometrics, Fourth Edition, p-519) as follows:

For cubic cost function (Model in 3.2)

$$\hat{Y}_i = 141.767 + 63.478X_i - 12.962X_i^2 + 0.939X_i^3$$

t-statistic: (22.2) (13.3) (–13.2) (15.9)

$$R^2 = 0.9983, \bar{R}^2 = 0.9975, d = 2.70$$

For quadratic cost function (Model in 3.3)

$$\hat{Y}_i = 222.383 - 8.0250X_i + 2.542X_i^2$$

t-statistic: (9.5) (- 0.82) (2.9)

$$R^2 = 0.9284, \bar{R}^2 = 0.9079, d = 1.038$$

For linear cost function (Model in 3.4)

$$\hat{Y}_i = 166.467 + 19.933X_i$$

t-statistic: (8.752) (6.502)

$$R^2 = 0.8409, \bar{R}^2 = 0.8210, d = 0.716$$

Given the true relationship between total cost and output (Model 3.2), model 3.3 and 3.4 suffer from specification errors. The residual series from estimation of model 3.3 and 3.4 also exhibit distinct patterns, indicating presence of specification errors (Fig. 3.1).

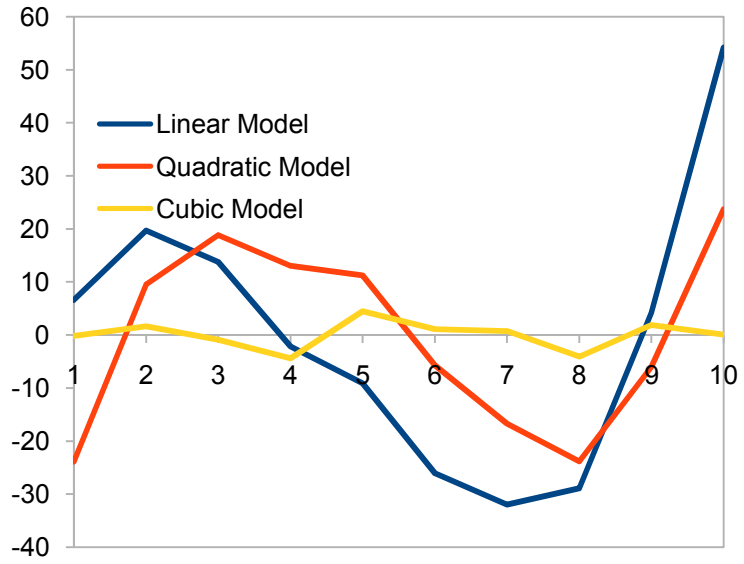

Fig. 3.1: Residual Series from Linear, Quadratic, and Cubic Models

A formal approach was developed by Ramsey (1969) to detect the presence of specification errors. The test is called RESET (regression specification error test). To explain this test, let us revert to our total cost model discussed above. After a visual examination of the residual series from linear, quadratic and cubic models, we saw that model 3.3 and 3.4 are mis-specified in relation to model 3.2. Let us suppose that a researcher estimates the model 3.4 ($Y_i = \gamma_1 + \gamma_2 X_i + u_i$) and proceeds to test for the specification error. Then the steps involved in Ramsey's RESET test are as follows:

a) From incorrectly estimated linear cost model, we first obtain the estimated, or fitted values of total cost$(\hat{Y}_i)$.

b) Now estimate the model again after including the higher powers of the estimated total cost viz.$(\hat{Y}_i^2, \hat{Y}_i^3)$ as:

$$Y_i = \gamma_1 + \gamma_2 X_i + \delta_1 \hat{Y}_i^2 + \delta_2 \hat{Y}_i^3 + u_i \qquad (3.5)$$

c) Now our initially estimated model (3.4) is restricted model and the model (3.5) is the unrestricted model. Consider the R^2 values of both these models [i.e. R^2_{ur} and R^2_r].

d) The null hypothesis of the test says that the restricted model is correctly specified. That is:

$$H_0: \delta_1 = \delta_2 = 0$$

Now, by using the test statistic:

$$F = \frac{(R^2_{ur} - R^2_r)/m}{(1 - R^2_{ur})/(n-k)} \qquad (3.6)$$

Where m is the number of restrictions imposed [i.e. 2 in our case for the two extra regressors included in the unrestricted model], n is the number of observations and k is the number of parameters in the model. The F-statistic above will have m and $(n–k)$ degree of freedom i.e. 2 and 6.

e) If the computed value of *F* is statistically significant, we reject the null hypothesis of correct specification of the restricted model and conclude that the model is mis-specified.

Returning to our total cost model, specified in 3.4, we get the following estimation for restricted and unrestricted model:

$$\text{Restricted Model: } \hat{Y}_i = 166.467 + 19.933X_i : R^2 = 0.8409.$$

$$\text{Unrestricted Model: } \hat{Y}_i = 2140.722 + 476.655X_i - 0.092\hat{Y}_i^2 + 0.0001\hat{Y}_i^3 : \ R^2 = 0.9983.$$

Thus, the computed *F*-statistic will be:

$$F = \frac{(0.9983 - 0.8409)/2}{(1 - 0.9983)/(10 - 4)} = 284.4035$$

For such a high value of the test statistic, we can reject the null hypothesis at all levels of significance. Thus, we can conclude that model 3.4 is mis-specified. The intuition behind this test is that if adding the powers of predicted variable increase the explanatory power of the model then this might be an evidence of specification error. Although easy to apply, the RESET test has some drawbacks. *First* is that the test does not suggest any alternative specification. *Second* is that the test does not provide any guide on the power of the variable included in the unrestricted model.

Check Your Progress 1 [answer within the space given in about 50-100 words]

1) State the six model selection criteria suggested by Hendry and Richard.

……

……

……

……

……

2) How are the *t*-test and the *F*-test helpful in identifying for irrelevant variables in a regression model?

……

……

……

……

……

……

3) What does the term ‘data mining’ connote?

……

……

……

……

……

4) Mention the six broad indicators of a regression result which indicates the relevance of the model estimated.

……

……

……

……

……

5) What feature observed in the graphical residual series in Fig. 3.1 suggest that the cubic model is superior to the quadratic and linear models?

……

……

……

……

……

6) What does the acronym RESET stand for? What specific purpose does this test serve? Specify the steps involved in this test procedure.

……

……

……

……

……

7) Mention the two limitations of the RESET test.

……

……

……

……

……

……

8) Consider the estimated regression:

$$\hat{Y}_i = -21.77 + 0.002X_{2i} + 0.123X_{3i} + 13.85X_{4i}$$
$$(29.475)\quad (0.0006)\quad (0.013)\quad (9.010)$$
$$n = 88; R^2 = 0.672$$

For testing specification error, RESET test was conducted and following subsidiary regression was obtained:

$$\hat{Y}_i = 166.097 + 0.0001X_{2i} + 0.0176X_{3i} + 2.175X_{4i} + 0.000353\hat{Y}_i^2 + 0.00000154\hat{Y}_i^3$$
$$(317.433)\quad (0.00520)\quad (0.299251)\quad (33.8881)\quad (0.0071)\quad (0.0000065)$$
$$n = 88; R^2 = 0.70553$$

Carry out the RESET test for specification error at 5% level of significance stating clearly the null and alternative hypothesis.

……

……

……

……

……

3.3 MODEL SELECTION CRITERIA

Model selection criteria is defined as the set of rules used to select a regression model, from among a set of models, based on observed data. It aims at minimising the expected dissimilarity between the chosen model and the true model. Several criteria are developed for this and these are discussed in this section below. Most of these primarily focus on minimising the 'residual sum of squares' (RSS). It is important to distinguish between the terms 'in-sample forecasting' and 'out-of-sample forecasting' here. An in-sample forecast employs a subset of the dataset to forecast the values within the estimation period and compare them to the actual outcomes. In other words, in-sample forecasting is done to assess how well the chosen model fits the data in a given sample. An out-of-sample forecasting uses all the values in the available data in the sample to predict the future value of the regressand.

3.3.1 R^2

The coefficient of determination (R^2) is one of the measures of goodness of fit of a regression model. Recall that it is defined as:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \qquad (3.7)$$

R^2 lies between 0 and 1. A value of R^2 closer to 1 indicates a good fit. However, R^2 suffers from some drawbacks. *Firstly,* it is an 'estimate' indicating the degree of closeness of a fitted value to that of the actual value. Thus, it is an in-sample measure of goodness of fit which does not essentially

provide an accurate out-of-sample forecasting. *Second,* to compare the R^2 from two models, the dependent variable has to be same. *Third,* R^2 is an increasing function of number of explanatory variables in the model. This means, it can be increased by simply increasing the number of explanatory variables in the model whereas adding more variables may also increase error variance. Therefore, one cannot rely solely on the values of R^2 for choosing the best model.

3.3.2 Adjusted R^2

Henry Theil (1961) developed another measure of goodness of fit called the adjusted R^2 (or $\bar{R}^2$). Recall that it introduces a penalty for each additional variable included in the regression model by reducing the degrees of freedom as follows:

$$\bar{R}^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-k} \qquad (3.8)$$

Thus, for a *k*-variable model, $\bar{R}^2 \leqslant R^2$. However, unlike R^2, adjusted R^2 will increase only if the additional variable included in the model significantly increases the explanatory power of the model. Thus, as a criteria for model selection, adjusted R^2 is a better indicator than R^2. However, here also in order to compare the adjusted R^2 from two models, the dependent variable must be the same.

3.3.3 Akaike Information Criterion

Akaike information criterion (AIC) is an estimator of the relative quality of each model i.e. relative to others. The criteria was developed by the Japanese statistician Hirotugu Akaikein 1970. In AIC criterion also, a penalty is imposed for additional variable included in the regression model. It is defined as:

$$AIC = e^{2k/n}\sum \hat{u}_i^2 = e^{2k/n}\frac{RSS}{n} \qquad (3.9)$$

where *k* is the number of parameters in the model and *n* is the number of observations in the sample. Taking log on both sides of Equation [(3.9)],we get:

$$lnAIC = \frac{2k}{n} + ln\left(\frac{RSS}{n}\right) \qquad (3.10)$$

where *lnAIC* is the 'natural log of AIC' and 2*k*/*n* is the 'penalty factor'. Most of the statistical software package report log transformed AIC. AIC imposes a stronger penalty as compared to the adjusted R^2 for including more variables in the regression model. As a model selection criteria, while choosing between many models, the model with lowest value of AIC, is preferred. There are many advantages of this criteria. One of them is that it can be useful for in-sample as well as out-of-sample forecasting of a regression model. It is also useful in choosing the lag length in an autoregressive model in time series analysis.

3.3.4 Schwarz Information Criterion (SIC)

Also called as the Bayesian information criteria (BIC), SIC is quite similar to that of AIC. It is defined as:

$$SIC = n^{k/n} \frac{\sum \hat{u_i^2}}{n} = n^{k/n} \frac{RSS}{n} \quad (3.11)$$

Log transformation of the above expression yields:

$$lnSIC = \frac{k}{n} lnn + ln\left(\frac{RSS}{n}\right) \quad (3.12)$$

In (3.12), the expression $[k/n\, lnn]$ is the 'penalty factor'. SIC imposes the strongest penalty (stronger than AIC) for adding an additional variable in the regression model. For model selection purposes, models with lower value of SIC is considered better. It is also useful for comparing both the in-sample as well as the out-of-sample forecasting performance of a regression model.

3.3.5 Hannan-Quinn Information Criterion (HQIC)

The Hannan-Quinn information criterion (HQIC) is a measure of the goodness of fit of a regression model. It is often used as a criterion for model selection as an alternative to Schwarz Information Criterion (SIC). It is defined as:

$$HQIC = nln\frac{RSS}{n} + 2kln[lnn] \quad (3.13)$$

where $2kln[lnn]$ is the penalty factor for adding an extra variable in the regression model. This penalty factor is also stronger than the penalty factor in AIC. Given any two estimated models, the model with the lower value of HQIC is preferred. Like AIC and BIC, HQIC is also useful for comparing in-sample and out-of-sample forecasting performance of the regression model.

3.4 ILLUSTRATION FOR MODEL SELECTION USING THE VARIOUS CRITERIA

Let us consider a wage determination model as follows:

$$wage = \beta_1 + \beta_2 \square ours_i + \beta_3 educ_i + \beta_4 exp_i + \beta_5 tenure_i + \beta_6 age_i + \beta_7 married_i + \beta_8 race_i + \beta_9 sibs_i + \beta_{10} sout \square_i \quad (3.12)$$

where *wage* = hourly wages in $, *hours* = number of working hours, *educ* = education in years, *exp* = work experience in years, *tenure* = tenure or period in present occupation, *age* = age in years, *married* = 1 if married, 0 otherwise, *race* = 1 if non-white, 0 otherwise, *sibs* = number of siblings in the family and *south* = 1 if region of residence is south. 0 otherwise. A priori, education, work experience, age, tenure, marital status are expected to be positively related to hourly wages and race and region of residence negatively related to hourly wages. We consider the estimation results based on a dataset of 935 observations drawn from the secondary source ['cps4_small' dataset,

available on Online Resource Centre, Oxford University Press, for Introduction to Econometrics, by Christopher Dougherty, 5th Edition] as in Table 3.1. All the variables in the regression results (Table 3.1) have expected signs but not all are statistically significant. Since estimations are based on cross-sectional data with large number of observations, a low R^2 (0.20) value can be justified. The R^2 value is statistically significant as the computed F value (28) is higher than its p-value (which are all close to zero). Thus, the F test for the joint significance of all the variable in the model is met.

For pedagogical purposes, another model is estimated after dropping some variables, which were statistically insignificant in the earlier estimation i.e. as per Table 3.1. The estimated results of this model are presented in Table 3.2. All the variables considered are statistically significant at 5% level of significance. The value of R^2 and adjusted R^2 is evidently smaller as there are lesser number of variables in Model 2. But the value of BSI criterion is smaller in this model whereas the value of AKI criterion is almost same for both the models.

Table 3.1: Estimated Regression Results by OLS Estimation Procedure

Dependent Variable: Wage				
Independent Variables	**Coefficient**	**Std. Error**	**t-Ratio**	**p-Value**
Constant	-0323.1	200.6	- 1.6	0.11
Hours	- 3.02	2.26	- 1.34	0.18
Education	64.9	6.8	9.6	0.00
Experience	9.7	3.8	2.6	0.01
Tenure	5.4	2.5	2.1	0.03
Age	9.0	4.9	1.9	0.06
Married	172.1	35.2	4.9	0.00
Black	- 132.7	31.3	- 4.2	0.00
Siblings	- 6.7	5.0	- 1.3	0.18
South	- 82.9	26.6	- 3.1	0.00

Mean Dependent Variable: 957.95	S.D. of Dependent Variable: 404.4
SSR: 122414111.3	S.E. of Regression: 363.8
R-squared:0.20	Adjusted R-square: 0.19
F(9, 925): 27.9	P-value of F: 0.00
Log-likelihood:- 6834. 97	Akaike Criterion: 13689.93
Schwarz Criterion: 13738.34	Hannan-Quinn: 13708.39

Table 3.2: Estimated Regression Results by OLS Estimation for Modified Model

Dependent Variable: Wage				
Independent Variables	**Coefficient**	**Std. Error**	***t*-Ratio**	***p*-Value**
Constant	- 480.1	165.1	- 2.9	0.10
Hours	65.9	6.6	9.9	0.00
Education	10.1	3.8	2.7	0.01
Experience	5.8	2.5	2.3	0.02
Tenure	8.7	4.9	1.8	0.08
Age	169.5	35.4	4.8	0.00
Married	- 139.0	30.3	- 4.5	0.00
Black	- 82.1	26.6	- 3.1	0.00

Mean Dependent Variable: 957.95	S.D. of Dependent Variable: 404.4
SSR: 123038014.53	S.E. of Regression: 364.3
R-squared:0.19	Adjusted *R*-square: 0.19
F(9, 925): 34.4	*P*-value of *F*: 0.00
Log-likelihood:- 6837.34	Akaike Criterion: 13690.69
Schwarz Criterion: 13729.41	Hannan-Quinn: 13705.45

Thus, AKI suggests that either of the two models can be chosen whereas the BIS criterion suggests that Model 2 is better than Model 1. Sometimes serial-correlation is caused due to specification errors. Thus, testing Durbin-Watson test statistics can also be helpful in detecting the presence of specification errors. In our example, the Durbin-Watson test value is closer to 2 in both models (1.8172 for Model 1 and 1.7971 for Model 2), suggesting that there is no adequate evidence for auto-correlation and specification errors.

Check Your Progress 2 [answer within the space given in about 50-100 words]

1) Define the term ‘model selection criteria’. What does it basically aim at?

.....

.....

.....

.....

2) Distinguish between the terms 'in sample forecast' and 'out of the sample forecast'.

.....

.....

.....

.....

.....

3) How is R^2 useful in determining the 'choice of a model'? What are its limitations?

.....

.....

.....

.....

.....

4) How is adjusted R^2 superior to R^2?

.....

.....

.....

.....

.....

5) Define AIC. How is AIC superior to adjusted-R^2? What are its advantages?

.....

.....

.....

.....

6) What is Schwartz Information Criteria? What are its advantages?

.....

.....

.....

.....

3.5 LET US SUM UP

The CLRM assumes '*correct specification*' of model [i.e. all the theoretically relevant variables are included in the model, irrelevant variables are excluded and there are no errors of measurement]. Therefore, it is crucial to test for the

presence of specification errors. Econometric theory proposes several tests to detect the presence of the specification error. One of the widely used tests is the Ramsey's regression specification test (RESET). Further, theory also proposes several information criteria (such as R^2, adjusted R^2, Akaike information criteria, Schwarz information criteria) that help in reaching a good model. All of these criteria, except R^2, impose some penalty on including additional variables in the regression model.

3.6 KEYWORDS

Model Specification : Model specification refers to the description of the process by which the dependent variable is estimated by a set of independent variables considered.

Correct Specification : Correct specification of the model is one which represents the true relationship between the regressors and the regressand.

Restricted Model : This is the model which imposes some restrictions on the values of one or more of the coefficients of the model.

In- sample forecasting : An in-sample forecast employ a subset of the dataset to forecast the values within the estimation period and compare them to the actual outcomes. In other words, in-sample forecasting is done to assess how well the chosen model fits the data in a given sample.

Out-of-sample forecasting : An out-of-sample forecasting uses all the values in the available data in the sample to predict the future value of the regressand.

Simple Linear Regression Model : A model with only one independent variable is called as a simple linear regression model. For such a model, $R^2 \geq \bar{R}^2$. In simple linear regression, it is not required to conduct individual significance as well as joint significance because for such models, square of '*t*' value is equal to *F*.

3.7 SUGGESTED BOOKS FOR FURTHER READING

1) Gujarati, D.N. and Porter, D.C. (2010).*Essentials of Econometrics,* Fourth Edition), McGraw Hill.

2) Gujarati, D.N. (2015).*Econometrics by Examples,* (*Second* Edition), Palgrave McMillan.

3) Dougherty, C. (2011). *Introduction to Econometrics*, Oxford, Oxford University Press.

4) Gujarati, D.N. and Porter, D.C. (2009). *Basic Econometrics* (Fifth Edition), McGraw Hill.

5) Wooldridge, J.M. (2009), *Econometrics*, Cengage Learning.

3.8 ANSWERS/HINTSTOCHECKYOUR PROGRESSEXERCISES

Check Your Progress 1

1) Data admissibility, theoretical consistency, independent regressors, parameter constancy, data coherency and competitive or best among the rival models.

2) *t*-test helps in identifying whether an individual independent variable included in a regression model adds real explanation. This is done based on whether the test identifies the estimated parameter value as significant or not. Likewise, the *F*-test serves the purpose of identifying whether all the independent variables included in the model jointly makes for the relevance of the set of variables taken together.

3) In spite of many precautions taken, there still could be many omissions in the model specified and estimated. This basically happens because the data set used would not be satisfying the properties of the CLRM estimates obtained by the OLS method (which itself is based on the assumptions made holding true).In this context, the term 'data mining' is used to indicate the various diagnostic procedures to be applied on the data set or database in order that the database is as free as possible from theproblems resulting from the consequences of violation of assumptions.

4) (i) the value of coefficient of determination (R^2), (ii) the value of the adjusted R^2, (iii) significance of *t*-ratios estimated, (iv) results of *F*-test, (v) signs of the estimated coefficients, (vi) value of Durbin-Watson test statistic.

5) Notice that the curve for the cubic model is more linear. It is this linear character of the cubic curve (which could have been in increasing or decreasing fashion i.e. not necessarily closer or horizontal to the X-axis) which makes it score over the other two forms of the model estimated.

6) RESET stands for 'regression specific error test'. It serves the purpose of empirically or analytically testing for 'mis-specification' of the model. The steps involved are as follows: (i) estimate the fitted values of the dependent variable in step 1, (ii) estimate the model once again by including two higher powers of estimates obtained in step 1 (for the square and the cubic terms) [to be regarded as the unrestricted model], (iii) consider the *R*-square values of the restricted and the unrestricted models and (iv) test for the hypothesis of estimated parameters for the squared and the cubic terms by the *F*-test specified in Equation (3.6).The acceptance of the null establishes the hypothesis of no mis-specification.

7) One, it suggests no alternative specification. Two, the test does not provide any guide for the number of power of the predicted variables (i.e. 2 or 3 or 4, etc.)to be included in the unrestricted model.

8) Computed F is 4.67 which is greater than $F(2,82)$ i.e. 3.11.Therefore, we reject the null hypothesis of correct specification. We conclude that the model is mis-specified.

Check Your Progress 2

1) It is defined as the set of rules used to select a regression model, from among a set of candidate models, based on observed data. Its primary focus is on minimising the 'residual sum of squares' (RSS).

2) The former employs a subset of the dataset to forecast the values within the estimation period and compare them to the actual outcomes. The latter uses all the values in the available data in the sample to predict the future value of the regressand.

3) It is defined as the ratio of 'error sum of squares' to 'total sum of squares' i.e. ESS/TSS. It suffers from three limitations viz. (i) for comparing across models the regressand should be the same, (ii) it increases in its value with the number of explanatory variables, but it would also increase the 'error variance' and (iii) being an estimate, it is an in-sample measure of goodness of fit which does not essentially provide an accurate out-of-sample forecasting.

4) It is always less than R^2. Further, in case of adjusted-R^2, any increase in its value is only when it adds to the explanatory power of the model (and not otherwise like in the case of R^2).

5) AIC is an estimator of the relative quality of each model i.e. relative to others. It is defined as: $AIC = e^{2k/n} \sum \hat{u_i^2} = e^{2k/n} \frac{RSS}{n}$where k is the number of parameters in the model and n is number of observations in the sample. By imposing a stronger penalty for including every additional variable as compared to adjusted-R^2, it is superior to the adjusted-R^2.Its advantages are: (i) it is useful both for in-sample as well as out-of-sample forecasting of a regression model; and (ii)it is also useful in choosing the lag length in an autoregressive model in time series analysis.

6) SIC is defined as: $SIC = n^{k/n} \frac{\sum \hat{u_i^2}}{n} = n^{k/n} \frac{RSS}{n}$. SIC imposes the strongest penalty (stronger than AIC) for adding an additional variable in the regression model. It is useful for comparing both the in-sample as well as the out-of-sample forecasting performance of a regression model.