Greene 7/e ms 1-1

1. Econometrics

1.1 INTRODUCTION

This book will present an introductory survey of econometrics. We will discuss the fundamental ideas that define the methodology and examine a large number of specific models, tools and methods that econometricians use in analyzing data. This chapter will introduce the central ideas that are the paradigm of econometrics. Section 1.2 defines the field and notes the role that theory plays in motivating econometric practice. Section 1.3 discusses the types of applications that are the focus of econometric analyses. The process of econometric modeling is presented in Section 1.4 with a classic application, Keynes's consumption function. A broad outline of the book is presented in Section 1.5. Section 1.6 notes some specific aspects of the presentation, including the use of numerical examples and the mathematical notation that will be used throughout the book.

1.2 THE PARADIGM OF ECONOMETRICS

In the first issue of *Econometrica*, Ragnar Frisch (1933) said of the Econometric Society that

> *its main object shall be to promote studies that aim at a unification of the theoretical-quantitative and the empirical-quantitative approach to economic problems and that are penetrated by constructive and rigorous thinking similar to that which has come to dominate the natural sciences. But there are several aspects of the quantitative approach to economics, and no single one of these aspects taken by itself, should be confounded with econometrics. Thus, econometrics is by no means the same as economic statistics. Nor is it identical with what we call general economic theory, although a considerable portion of this theory has a definitely quantitative character. Nor should econometrics be taken as synonomous* [sic] *with the application of mathematics to economics. Experience has shown that each of these three viewpoints, that of statistics, economic theory, and mathematics, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the unification of all three that is powerful. And it is this unification that constitutes econometrics.*

The Society responded to an unprecedented accumulation of statistical information. They saw a need to establish a body of principles that could organize what would otherwise become a bewildering mass of data. Neither the pillars nor the objectives of econometrics have changed in the years since this editorial appeared. Econometrics concerns itself with the application of mathematical statistics and the tools of statistical inference to the empirical measurement of relationships postulated by an underlying theory.

The crucial role that econometrics plays in economics has grown over time. The Nobel Prize in Economic Sciences has recognized this contribution with numerous awards to econometricians, including the first which was given to (the same) Ragnar Frisch in 1969, Lawrence Klein in 1980, Trygve Haavelmo in 1989, James Heckman and Daniel McFadden in 2000, and Robert Engle and Clive Granger in 2003. The 2000 prize was noteworthy in that it celebrated the work of two scientists whose research was devoted to the marriage of behavioral theory and econometric modeling.

Example 1.1 Behavioral Models and the Nobel Laureates

The pioneering work by both James Heckman and Dan McFadden rests firmly on a theoretical foundation of utility maximization.

For Heckman's, we begin with the standard theory of household utility maximization over consumption and leisure. The textbook model of utility maximization produces a demand for leisure time that translates into a supply function of labor. When home production (work in the home as opposed to the outside, formal labor market) is considered in the calculus, then desired "hours" of (formal) labor can be negative. An important conditioning variable is the "reservation" wage—the wage rate that will induce formal labor market participation. On the demand side of the labor market, we have firms that offer market wages that respond to such attributes as age, education, and experience. What can we learn about labor supply behavior based on observed market wages, these attributes and observed hours in the formal market? Less than it might seem, intuitively because our observed data omit half the market—the data on formal labor market activity are not randomly drawn from the whole population.

Heckman's observations about this implicit truncation of the distribution of hours or wages revolutionized the analysis of labor markets. Parallel interpretations have since guided analyses in every area of the social sciences. The analysis of policy interventions such as education initiatives, job training and employment policies, health insurance programs, market creation, financial regulation and a host of others is heavily influenced by Heckman's pioneering idea that when participation is part of the behavior being studied, the analyst must be cognizant of the impact of common influences in both the presence of the intervention and the outcome. We will visit the literature on sample selection and treatment/program evaluation in Chapter 18.

Textbook presentations of the theories of demand for goods that produce utility, since they deal in continuous variables, are conspicuously silent on the kinds of discrete choices that consumers make every day—what brand of product to choose, whether to buy a large commodity such as a car or a refrigerator, how to travel to work, whether to rent or buy a home, where to live, what candidate to vote for, and so on. Nonetheless, a model of "random utility" defined over the alternatives available to the consumer provides a theoretically sound platform for studying such choices. Important variables include, as always, income and relative prices. What can we learn about underlying preference structures from the discrete choices that consumers make? What must be assumed about these preferences to allow this kind of inference? What kinds of statistical models will allow us to draw inferences about preferences? McFadden's work on how commuters choose to travel to work, and on the underlying theory appropriate to this kind of modeling, has guided empirical research in discrete consumer choices for several decades. We will examine McFadden's models of discrete choice in Chapter ~~16~~ 17.

The connection between underlying behavioral models and the modern practice of econometrics is increasingly strong. A useful distinction is made between *microeconometrics* and *macroeconometrics*. The former is characterized by its analysis of cross section and panel data and by its focus on individual consumers, firms, and micro-level decision makers. Practitioners rely heavily on the theoretical tools of microeconomics including utility maximization, profit maximization, and market equilibrium. The analyses are directed at subtle, difficult questions that often require intricate formulations. A few applications are as follows:

- What are the likely effects on labor supply behavior of proposed negative income taxes? [Ashenfelter and Heckman (1974).]
- Does attending an elite college bring an expected payoff in lifetime expected income sufficient to justify the higher tuition? [Kreuger and Dale (1999) and Kreuger (2000).]
- Does a voluntary training program produce tangible benefits? Can these benefits be accurately measured? [Angrist (2001).]
- Do smaller class sizes bring real benefits in student performance? [Hanuschek (1999), Hoxby (2000), Angrist and Lavy (1999).]
- Does the presence of health insurance induce individuals to make heavier use of the health care system - is moral hazard a measurable problem? [Riphahn et al. (2003).]

Macroeconometrics is involved in the analysis of time-series data, usually of broad aggregates such as price levels, the money supply, exchange rates, output, investment, economic growth and so on. The boundaries are not sharp. For example, an application that we will examine in this text concerns spending patterns of municipalities, which rests somewhere between the two fields. The very large field of financial econometrics is concerned with long time-series data and occasionally vast panel data sets, but with a

sharply focused orientation toward models of individual behavior. The analysis of market returns and exchange rate behavior is neither exclusively macro- nor microeconometric. (We will not be spending any time in this book on financial econometrics. For those with an interest in this field, I would recommend the celebrated work by Campbell, Lo, and Mackinlay (1997) or, for a more time-series–oriented approach, Tsay (2005).) Macroeconomic model builders rely on the interactions between economic agents and policy makers. For examples:

- Does a monetary policy regime that is strongly oriented toward controlling inflation impose a real cost in terms of lost output on the U.S. economy? [Cecchetti and Rich (2001).]
- Did 2001's largest federal tax cut in U.S. history contribute to or dampen the concurrent recession? Or was it irrelevant?

Each of these analyses would depart from a formal model of the process underlying the observed data.

1.3 THE PRACTICE OF ECONOMETRICS

We can make another useful distinction between *theoretical econometrics* and *applied econometrics*. Theorists develop new techniques for estimation and hypothesis testing and analyze the consequences of applying particular methods when the assumptions that justify those methods are not met. Applied econometricians are the users of these techniques and the analysts of data ("real world" and simulated). The distinction is far from sharp; practitioners routinely develop new analytical tools for the purposes of the study that they are involved in. This book contains a large amount of econometric theory, but it is directed toward applied econometrics. I have attempted to survey techniques, admittedly some quite elaborate and intricate, that have seen wide use "in the field."

Applied econometric methods will be used for estimation of important quantities, analysis of economic outcomes such as policy changes, markets or individual behavior, testing theories, and for forecasting. The last of these is an art and science in itself that is the subject of a vast library of sources. Although we will briefly discuss some aspects of forecasting, our interest in this text will be on estimation and analysis of models. The presentation, where there is a distinction to be made, will contain a blend of microeconometric and macroeconometric techniques and applications. It is also necessary to distinguish between *time-series analysis* (which is not our focus) and methods that primarily use time-series data. The former is, like forecasting, a growth industry served by its own literature in many fields. While we will employ some of the techniques of time-series analysis, we will spend relatively little time developing first principles.

1.4 ECONOMETRIC MODELING

Econometric analysis usually begins with a statement of a theoretical proposition. Consider, for example, a classic application by one of Frisch's contemporaries:

Example 1.2 Keynes's Consumption Function

From Keynes's (1936) *General Theory of Employment, Interest and Money:*

> We shall therefore define what we shall call the propensity to consume as the functional relationship f between X, a given level of income, and C, the expenditure on consumption out of the level of income, so that $C = f(X)$.
>
> The amount that the community spends on consumption depends (i) partly on the amount of its income, (ii) partly on other objective attendant circumstances, and (iii) partly on the subjective needs and the psychological propensities and habits of the individuals composing it. The fundamental psychological law upon which we are entitled to depend with great confidence, both a priori from our knowledge of human nature and from the detailed facts of experience, is that men are disposed, as a rule and on the average, to increase their consumption as their income increases, but not by as much as the increase in their income. That is, dC/dX is positive and less than unity.
>
> But, apart from short period changes in the level of income, it is also obvious that a higher absolute level of income will tend as a rule to widen the gap between income and consumption. These reasons will lead, as a rule, to a greater proportion of income being saved as real income increases.

The theory asserts a relationship between consumption and income, $C = f(X)$, and claims in the second paragraph that the marginal propensity to consume (MPC), dC/dX, is between zero and one.[1] The final paragraph asserts that the average propensity to consume (APC), C/X, falls as income rises, or $d(C/X)/dX$ = (MPC − APC)$/X$ < 0. It follows that MPC < APC. The most common formulation of the consumption function is a linear relationship, $C = \alpha + X\beta$, that satisfies Keynes's "laws" if β lies between zero and one and if α is greater than zero.

These theoretical propositions provide the basis for an econometric study. Given an appropriate data set, we could investigate whether the theory appears to be consistent with the observed "facts." For example, we could see whether the linear specification appears to be a satisfactory description of the relationship between consumption and income, and, if so, whether α is positive and β is between zero and one. Some issues that might be studied are (1) whether this relationship is stable through time or whether the parameters of the relationship change from one generation to the next (a change in the average propensity to save, 1−APC, might represent a fundamental change in the behavior of consumers in the economy); (2) whether there are systematic differences in the relationship across different countries, and, if so, what explains these differences; and (3) whether there are other factors that would improve the ability of the model to explain the relationship between consumption and income. For example, Figure 1.1 presents aggregate consumption and personal income in constant dollars for the U.S. for the 10 years of 2000–2009. (See Appendix Table F1.1.) Apparently, at least superficially, the data (the facts) are consistent with the theory. The relationship appears to be linear, albeit only approximately, the intercept of a line that lies close to most of the points is positive and the slope is less than one, although not by much. (However, if the line is fit by linear least squares regression, the intercept is negative, not positive.)

[1] Modern economists are rarely this confident about their theories. More contemporary applications generally begin from first principles and behavioral axioms, rather than simple observation.

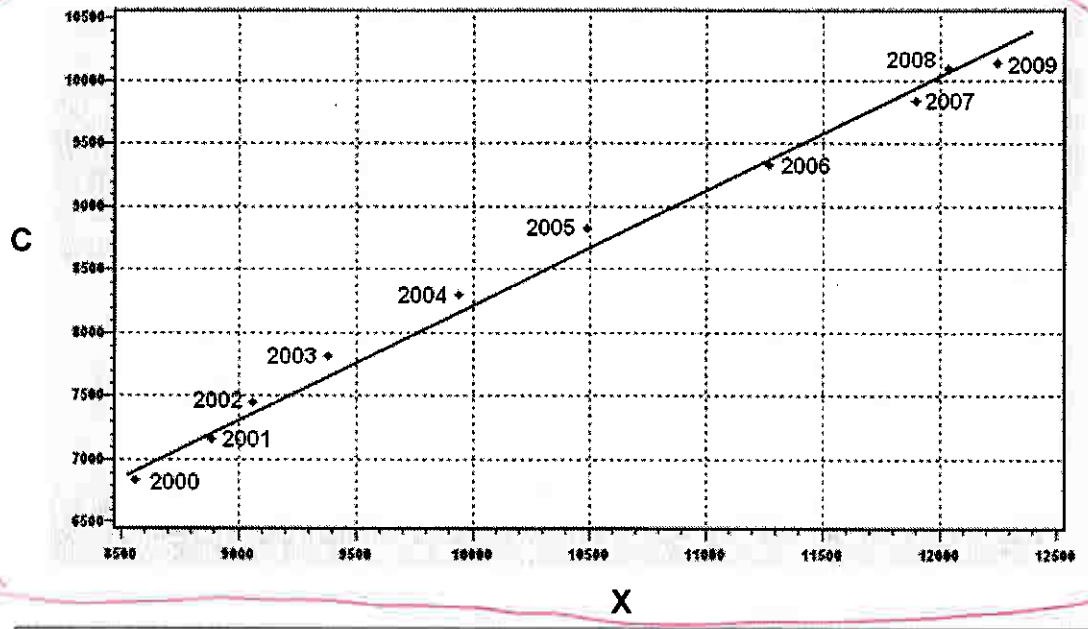

Figure 1.1 Aggregate U.S. Consumption and Income Data, 2000-2009

Economic theories such as Keynes's are typically sharp and unambiguous. Models of demand, production, labor supply, individual choice, educational attainment, income and wages, investment, market equilibrium and aggregate consumption all specify precise, *deterministic* relationships. Dependent and independent variables are identified, a functional form is specified, and in most cases, at least a qualitative statement is made about the directions of effects that occur when independent variables in the model change. The model is only a simplification of reality. It will include the salient features of the relationship of interest, but will leave unaccounted for influences that might well be present but are regarded as unimportant.

Correlations among economic variables are easily observable through descriptive statistics and techniques such as linear regression methods. The ultimate goal of the econometric model builder is often to uncover the deeper causal connections through elaborate structural, behavioral models. Note, for example, Keynes's use of the behavior of a "representative consumer" to motivate the behavior of macroeconomic variables such as income and consumption. Heckman's model of labor supply noted in Example 1.1 is framed in a model of individual behavior. Berry, Levinsohn and Pakes's (1995) detailed model of equilibrium pricing in the automobile market is another.

No model could hope to encompass the myriad essentially random aspects of economic life. It is thus also necessary to incorporate stochastic elements. As a consequence, observations on a ~~dependent~~ variable will display variation attributable not only to differences in variables that are explicitly accounted for, but also to the randomness of human behavior and the interaction of countless minor influences that are not. It is understood that the introduction of a random "disturbance" into a deterministic model is not intended merely to paper over its inadequacies. It is essential to examine the results of the study, in a sort of postmortem, to ensure that the allegedly random, unexplained factor is truly unexplainable. If it is not, the model is, in fact, inadequate. [In the example given earlier, the estimated constant term in the linear least squares regression is negative. Is the theory wrong, or is the finding due to random fluctuation in the data? Another possibility is that the theory is broadly correct, but the world changed between 1936 when Keynes devised his theory and 2000-2009 when the data (outcomes) were generated. Or, perhaps linear least squares is not the appropriate technique to use for this model, and that is responsible for the inconvenient result (the negative intercept).] The stochastic element endows the model with its statistical properties. Observations on the variable(s) under study are thus taken to be the outcomes of random processes. With a sufficiently detailed stochastic structure and adequate data, the analysis will become a matter of deducing the properties of a probability distribution. The tools and methods of mathematical statistics will provide the operating principles.

A model (or theory) can never truly be confirmed unless it is made so broad as to include every possibility. But it may be subjected to ever more rigorous scrutiny and, in the face of contradictory evidence, refuted. A deterministic theory will be invalidated by a single contradictory observation. The introduction of stochastic elements into the model changes it from an exact statement to a probabilistic description about expected outcomes, and carries with it an important implication. Only a preponderance of contradictory evidence can convincingly invalidate the probabilistic model, and what constitutes a "preponderance of evidence" is a matter of interpretation. Thus, the probabilistic model is less precise but at the same time, more robust.[2]

The techniques used in econometrics have been employed in a widening variety of fields, including political methodology, sociology [see, e.g., Long (1997) and DeMaris (2004)], health economics, medical research (how do we handle attrition from medical treatment studies?) environmental economics, economic geography, transportation engineering, and numerous others. Practitioners in these fields and many more are all heavy users of the techniques described in this text.

The process of econometric analysis departs from the specification of a theoretical relationship. We initially proceed on the optimistic assumption that we can obtain precise measurements on all the variables in a correctly specified model. If the ideal conditions are met at every step, the subsequent analysis will be routine. Unfortunately, they rarely are. Some of the difficulties one can expect to encounter are the following:

- The data may be badly measured or may correspond only vaguely to the variables in the model. "The interest rate" is one example.
- Some of the variables may be inherently unmeasurable. "Expectations" is a case in point.
- The theory may make only a rough guess as to the correct form of the model, if it makes any at all, and we may be forced to choose from an embarrassingly long menu of possibilities.
- The assumed stochastic properties of the random terms in the model may be demonstrably violated, which may call into question the methods of estimation and inference procedures we have used.
- Some relevant variables may be missing from the model.
- The conditions under which data are collected leads to a sample of observations that is systematically unrepresentative of the population we wish to study.

The ensuing steps of the analysis consist of coping with these problems and attempting to ~~cull~~ extract whatever information is likely to be present in such obviously imperfect data. The methodology is that of mathematical statistics and economic theory. The product is an econometric model.

[2] See Keuzenkamp and Magnus (1995) for a lengthy symposium on testing in econometrics.

1.5 PLAN OF THE BOOK

Econometrics is a large and growing field. It is a challenge to chart a course through that field for the beginner. Our objective in this survey is to develop in detail a set of tools, then use those tools in applications. The set of applications presented below is large and will include many that readers will use in practice. But, it is not exhaustive. We will attempt to present our results in sufficient generality that the tools we develop here can be extended to other kinds of situations and applications not described here.

One possible approach is to organize (and orient) the areas of study by the type of data being analyzed – cross section, panel, discrete data, then time series being the obvious organization. Alternatively, we could distinguish at the outset between micro- and macro econometrics.[3] Ultimately, all of these will require a common set of tools, including, for example, the multiple regression model, the use of moment conditions for estimation, instrumental variables (IV) and maximum likelihood estimation. With that in mind, the organization of this book is as follows: The first half of the text develops fundamental results that are common to all of the applications. The concept of multiple regression and the linear regression model in particular, constitutes the underlying platform of most modeling, even if the linear model itself is not ultimately used as the empirical specification. This part of the text concludes with developments of IV estimation and the general topic of panel data modeling. The latter pulls together many features of modern econometrics, such as, again, IV estimation, modeling heterogeneity, and a rich variety of extensions of the linear model. The second half of the text presents a variety of topics. Part III is an overview of estimation methods. Finally, Parts IV and V present results from microeconometrics and macroeconometrics, respectively. The broad outline is as follows:

I. Regression Modeling
Chapters 2 through 6 present the multiple linear regression model. We will discuss specification, estimation, and statistical inference. This part develops the ideas of estimation, robust analysis, functional form and principles of model specification. ~~Chapter 7 describes the use of instrumental variables.~~

II. Generalized Regression Instrumental Variables and Panel Data
Chapters 8 and 9 introduce the generalized regression model and systems of regression models. This section ends with Chapter 11 on panel data methods.

III. Estimation Methods
Chapters 12 through 16 present general results on different methods of estimation including GMM, maximum likelihood, and simulation based methods. Various estimation frameworks, including non- and semiparametric and Bayesian estimation are presented in Chapters 12 and 16.

IV. Microeconometric Methods
Chapters 17 through 18 are about microeconometrics, discrete choice modeling and limited dependent variables, and the analysis of data on events—how many occur in a given setting and when they occur. Chapters 17 to 19 are devoted to methods more suited to cross sections and panel data sets.

V. Macroeconometric Methods
Chapters 20 to 23 focus on time series modeling and macroeconometrics.

VI. Background Materials
Appendices A through E present background material on tools used in econometrics including matrix algebra, probability and distribution theory, estimation, and asymptotic distribution theory. Appendix E presents results on computation. Appendices A through E are chapter-length surveys of the tools used in econometrics. Because it is assumed that the reader has some previous training in each of these topics, these summaries are included primarily for those who desire a refresher or a convenient reference. We do not anticipate that these appendices can substitute for a course in any of these subjects. The intent of these chapters is to provide a reasonably concise summary of the results, nearly all of which are explicitly used elsewhere in the book. The data sets used in the numerical examples are described in Appendix F. The actual data sets and other supplementary materials can be downloaded from the author's web site for the text: http://pages.stern.nyu.edu/~wgreene/Text/. Useful tables related to commonly used probability distributions are given in Appendix G.

Chapter 7 extends the regression model to nonlinear functional forms. The method of instrumental variables is presented in Chapter 8.

following

fn 3 is on next page

URL

[3] An excellent reference on the former that is at a more advanced level than this book is Cameron and Trivedi (2005). As of this writing, there does not appear to be available a counterpart, large scale, pedagogical survey of macroeconometrics that includes both econometric theory and applications. The numerous more focused studies include books such as Bårdsen, G., Eitrheim, Ø., Jansen, E. and Nymoen, R., *The Econometrics of Macroeconomic Modelling*, Oxford University Press, 2005 and survey papers such as Wallis, K., "Macroeconometric Models," Published in *Macroeconomic Policy: Iceland in an Era of Global Integration* (M. Gudmundsson, T.T. Herbertsson, and G. Zoega, eds), pp.399-414. Reykjavik: University of Iceland Press, 2000 also at http://www.ecomod.net/conferences/ecomod2001/papers_web/Wallis_Iceland.pdf

1.6 PRELIMINARIES

Before beginning, we note some specific aspects of the presentation in the text.

1.6.1 Numerical Examples

There are many numerical examples given throughout the discussion. Most of these are either self contained exercises or extracts from published studies. In general, their purpose is to provided a limited application to illustrate a method or model. The reader can, if they wish, replicate them with the data sets provided. This will generally not entail attempting to replicate the full published study. Rather, we use the data sets to provide applications that relate to the published study in a limited, manageable fashion that also focuses on a particular technique, model or tool. Thus, the Riphahn, Wambach and Million (2003) provides a very useful, manageable (though relatively large) laboratory data set that the reader can use to explore some issues in health econometrics. The exercises also suggest more extensive analyses, again in some cases based on published studies.

1.6.2 Software and Replication

As noted in the preface, there are now many powerful computer programs that can be used for the computations described in this book. In most cases, the examples presented can be replicated with any modern package, whether the user is employing a high level integrated program such as *NLOGIT*, *Stata* or *SAS*, or writing their own programs in languages such as *R*, *MatLab* or *Gauss*. The notable exception will be exercises based on simulation. Since, essentially, every package uses a different random number generator, it will generally not be possible to replicate exactly the examples in this text that use simulation (unless you are using the same computer program we are). Nonetheless, the differences that do emerge in such cases should be attributable to, essentially, minor random variation. You will be able to replicate the essential results and overall features in these applications with any of the software mentioned. We will return to this general issue of replicability at a few points in the text, including in Section 14.2 where we discuss methods of generating random samples for simulation based estimators. 15

1.6.3 Notation and Conventions

We will use vector and matrix notation and manipulations throughout the text. The following conventions will be used: A scalar variable will be denoted with an italic lower case letter, such as y or x_{nK}, A column vector of scalar values will be denoted by a boldface, lower case letter, such as

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}$$ and, likewise for, $\mathbf{x}$, and $\mathbf{b}$. The dimensions of a column vector are always denoted as those of a matrix with one column, such as $K \times 1$ or $n \times 1$ and so on. A matrix will always be denoted by a boldface upper case letter, such as the $n \times K$ matrix, $\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nK} \end{bmatrix}$. Specific elements in a matrix are always subscripted so that the first subscript gives the row and the second gives the column. Transposition of a vector or a matrix is denoted with a prime. A row vector is obtained by transposing a column vector. Thus, $\boldsymbol{\beta}' = [\beta_1, \beta_2, \ldots, \beta_K]$. The product of a row and a column vector will always be denoted in a form such as $\boldsymbol{\beta}'\mathbf{x} = \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_K x_K$. The elements in a matrix, $\mathbf{X}$, form a set of vectors. In terms of its columns, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K]$ – each column is an $n \times 1$ vector. The one possible, unfortunately unavoidable source of ambiguity is the notation necessary to denote a row of a matrix such as $\mathbf{X}$. The elements of the ith row of $\mathbf{X}$ are the row vector, $\mathbf{x}_i' = [x_{i1}, x_{i2}, \ldots, x_{iK}]$. When the matrix, such as $\mathbf{X}$, refers to a data matrix, we will prefer to use the "i" subscript to denote observations, or the rows of the matrix and "k" to denote the variables, or columns. As we note unfortunately, this would seem to imply that $\mathbf{x}_i$, the transpose of $\mathbf{x}_i'$ would be the ith column of $\mathbf{X}$, which will conflict with our notation. However, with no simple alternative notation available, we will maintain this convention, with the understanding that $\mathbf{x}_i'$ *always* refers to the row vector that is the ith row of an $\mathbf{X}$ matrix. A discussion of the matrix algebra results used in the book is given in Appendix A. A particularly important set of arithmetic results about summation and the elements of the matrix product matrix, $\mathbf{X}'\mathbf{X}$ appears in Section A.2.7.

2. THE LINEAR REGRESSION MODEL

2.1 INTRODUCTION

Econometrics is concerned with *model building*. An intriguing point to begin the inquiry is to consider the question, "What is the model?" The statement of a "model" typically begins with an observation or a proposition that one variable "is caused by" another, or "varies with another," or some qualitative statement about a relationship between a variable and one or more **covariates** that are expected to be related to the interesting one in question. The model might make a broad statement about behavior, such as the suggestion that individuals' usage of the health care system depends on, e.g., perceived health status, demographics such as income, age and education, and the amount and type of insurance they have. It might come in the form of a verbal proposition, or even a picture such as a flowchart or **path diagram** that suggests directions of influence. The econometric model rarely springs forth in full bloom as a set of equations. Rather, it begins with an *idea* of some kind of relationship. The natural next step for the econometrician is to translate that idea into a set of equations, with a notion that some feature of that set of equations will answer interesting questions about the variable of interest. To continue our example, a more definite statement of the relationship between insurance and health care demanded might be able to answer *how* does health care system utilization depend on insurance coverage. Specifically, is the relationship "positive" – all else equal, is an insured consumer more likely to "demand more health care," or is it "negative?" And, ultimately, one might be interested in a more precise statement, "how much more (or less)?" This and the next several chapters will build up the set of tools that model builders use to pursue questions such as these using data and econometric methods.

From a purely statistical point of view, the researcher might have in mind a variable, y, broadly "demand for health care, H," and a vector of covariates, $\mathbf{x}$ (income, I, insurance, T), and a joint probability distribution of the three, $p(H,I,T)$. Stated in this form, the "relationship" is not posed in a particularly interesting fashion – what is the statistical process that produces health care demand, income, and insurance coverage. However, it is true that $p(H,I,T) = p(H|I,T)p(I,T)$, which decomposes the probability model for the joint process into two outcomes, the joint distribution of insurance coverage and income in the population and the distribution of "demand for health care" for a specific income and insurance coverage. From this perspective, the conditional distribution, $p(H|I,T)$, holds some particular interest, while $p(I,T)$, the distribution of income and insurance coverage in the population is perhaps of secondary, or no interest. (On the other hand, from the same perspective, the conditional "demand" for insurance coverage, given income, $p(T|I)$, might also be interesting.) Continuing this line of thinking, the model builder is often interested not in joint variation of all the variables in the model, but in **conditional variation** of one of the variables related to the others.

The idea of the conditional distribution provides a useful starting point for thinking about a relationship between a variable of interest, a "y," and a set of variables, "$\mathbf{x}$," that we think might bear some relationship to it There is a question to be considered now that returns us to the issue of "what is the model?" What feature of the conditional distribution is of interest? The model builder thinking in terms of features of the conditional distribution, often gravitates to the expected value, focusing attention on $E[y|\mathbf{x}]$, i.e., the **regression function**, which brings us to the subject of this chapter. For the example above, this might be natural if y were "doctor visits" as in an example examined at several points in the chapters to follow. If we were studying incomes, I, however, which often have a highly skewed distribution, then the mean might not be particularly interesting. Rather, the **conditional median**, for given ages, $M[I|\mathbf{x}]$, might be a more interesting statistic. On the other hand, still considering the distribution of incomes (and still conditioning on age), other quantiles, such as the 20th percentile, or a poverty line defined as, say, the 5th percentile, might be more interesting yet. Finally, consider a study in finance, in which the variable of interest is asset returns. In at least some contexts, means are not interesting at all – it is variances, and conditional variances in particular, that are most interesting.

The point of the preceding is that we begin the discussion of the regression model with an understanding of what we mean by "the model." For the present, we will focus on the conditional mean which is usually the feature of interest. Once we establish how to analyze the regression function, we will use it as a useful departure point for studying other features, such as quantiles and variances. The **linear**

regression model is the single most useful tool in the econometrician's kit. Although to an increasing degree in contemporary research, it is often only the departure point for the full analysis, it remains the device used to begin almost all empirical research. And, it is the lens through which relationships among variables are usually viewed. This chapter will develop the linear regression model. Here, we will detail the fundamental assumptions of the model. The next several chapters will discuss more elaborate specifications and complications that arise in the application of techniques that are based on the simple models presented here.

2-3

2 THE CLASSICAL MULTIPLE LINEAR REGRESSION MODEL

2.1 INTRODUCTION

An econometric study begins with a set of propositions about some aspect of the economy. The theory specifies a set of precise, deterministic relationships among variables. Familiar examples are demand equations, production functions, and macroeconomic models. The empirical investigation provides estimates of unknown parameters in the model, such as elasticities or the effects of monetary policy, and usually attempts to measure the validity of the theory against the behavior of the observed data. Once suitably constructed, the model might then be used for prediction or analysis of behavior. This book will develop a large number of models and techniques used in this framework.

The **linear regression model** is the single most useful tool in the econometrician's kit. Although to an increasing degree in the contemporary literature, it is often only the departure point for the full analysis, it remains the device used to begin almost all empirical research. This chapter will develop the model. The next several chapters will discuss more elaborate specifications and complications that arise in the application of techniques that are based on the simple models presented here.

2.2 THE LINEAR REGRESSION MODEL

The **multiple linear regression model** is used to study the relationship between a **dependent variable** and one or more **independent variables.** The generic form of the linear regression model is

$$y = f(x_1, x_2, \ldots, x_K) + \varepsilon$$
$$= x_1\beta_1 + x_2\beta_2 + \cdots + x_K\beta_K + \varepsilon \qquad \textbf{(2-1)}$$

where y is the dependent or **explained** variable and $x_1, \ldots, x_K$ are the independent or **explanatory** variables. One's theory will specify $f(x_1, x_2, \ldots, x_K)$. This function is commonly called the **population regression equation** of y on $x_1, \ldots, x_K$. In this setting, y is the **regressand** and x_k, $k = 1, \ldots,$ and K are the **regressors** or **covariates.** The underlying theory will specify the dependent and independent variables in the model. It is not always obvious which is appropriately defined as each of these—for example, a demand equation, $quantity = \beta_1 + price \times \beta_2 + income \times \beta_3 + \varepsilon$, and an inverse demand equation, $price = \gamma_1 + quantity \times \gamma_2 + income \times \gamma_3 + u$ are equally valid representations of a market. For modeling purposes, it will often prove useful to think in terms of "autonomous variation." One can conceive of movement of the independent

Au: Term "covariates" uses bold KT on msp 2-1. Mark if it's ok here?

variables outside the relationships defined by the model while movement of the dependent variable is considered in response to some independent or exogenous stimulus.[1]

The term ε is a random **disturbance,** so named because it "disturbs" an otherwise stable relationship. The disturbance arises for several reasons, primarily because we cannot hope to capture every influence on an economic variable in a model, no matter how elaborate. The net effect, which can be positive or negative, of these omitted factors is captured in the disturbance. There are many other contributors to the disturbance in an empirical model. Probably the most significant is errors of measurement. It is easy to theorize about the relationships among precisely defined variables; it is quite another to obtain accurate measures of these variables. For example, the difficulty of obtaining reasonable measures of profits, interest rates, capital stocks, or, worse yet, flows of services from capital stocks is a recurrent theme in the empirical literature. At the extreme, there may be no observable counterpart to the theoretical variable. The literature on the permanent income model of consumption [e.g., Friedman (1957)] provides an interesting example.

We assume that each observation in a sample $(y_i, x_{i1}, x_{i2}, \ldots, x_{iK}), i = 1, \ldots, n,$ is generated by an underlying process described by

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i.$$

The observed value of y_i is the sum of two parts, a deterministic part and the random part, ε_i. Our objective is to estimate the unknown parameters of the model, use the data to study the validity of the theoretical propositions, and perhaps use the model to predict the variable y. How we proceed from here depends crucially on what we assume about the stochastic process that has led to our observations of the data in hand.

Example 2.1 Keynes's Consumption Function

Example 1.2 discussed a model of consumption proposed by Keynes and his *General Theory* (1936). The theory that consumption, C, and income, X, are related certainly seems consistent with the observed "facts" in Figures 1.1 and 2.1. (These data are in Data Table F2.1.) Of course, the linear function is only approximate. Even ignoring the anomalous wartime years, consumption and income cannot be connected by any simple **deterministic relationship.** The linear model, $C = \alpha + \beta X$, is intended only to represent the salient features of this part of the economy. It is hopeless to attempt to capture every influence in the relationship. The next step is to incorporate the inherent randomness in its real-world counterpart. Thus, we write $C = f(X, \varepsilon)$, where ε is a stochastic element. It is important not to view ε as a catchall for the inadequacies of the model. The model including ε appears adequate for the data not including the war years, but for 1942–1945, something systematic clearly seems to be missing. Consumption in these years could not rise to rates historically consistent with these levels of income because of wartime rationing. A model meant to describe consumption in this period would have to accommodate this influence.

It remains to establish how the stochastic element will be incorporated in the equation. The most frequent approach is to assume that it is *additive*. Thus, we recast the equation in stochastic terms: $C = \alpha + \beta X + \varepsilon$. This equation is an empirical counterpart to Keynes's theoretical model. But, what of those anomalous years of rationing? If we were to ignore our intuition and attempt to "fit" a line to all these data, the next chapter will discuss at length how we should do that, we might arrive at the dotted line in the figure as our best

[1] By this definition, it would seem that in our demand relationship, only income would be an independent variable while both price and quantity would be dependent. That makes sense—in a market, price and quantity *are* determined at the same time, and do change only when something outside the market changes. ~~We will return to this specific case in Chapter 13.~~

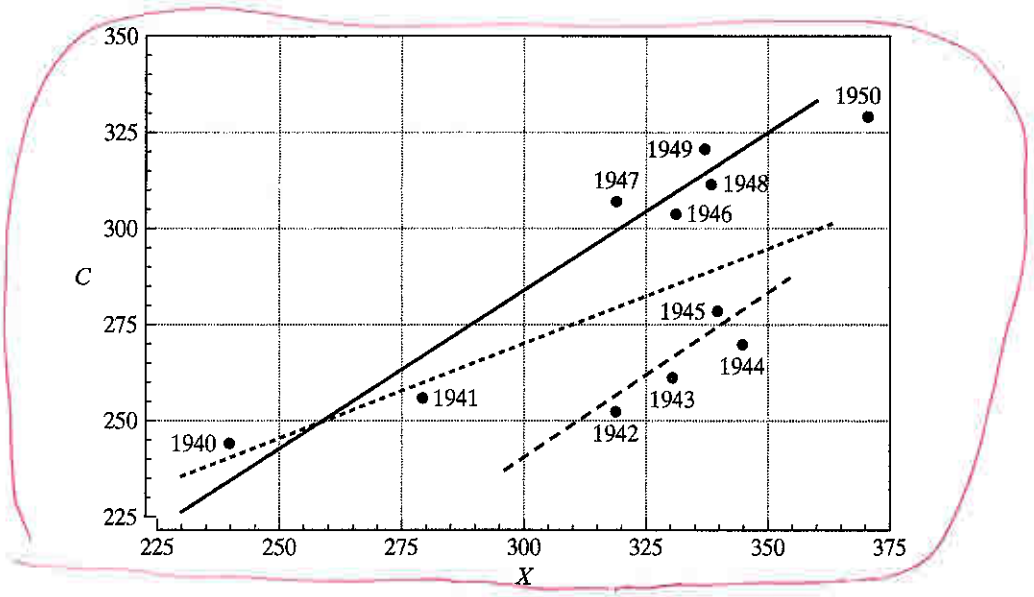

FIGURE 2.1 Consumption Data, 1940–1950.

guess. This line, however, is obviously being distorted by the rationing. A more appropriate specification for these data that accommodates both the stochastic nature of the data and the special circumstances of the years 1942–1945 might be one that shifts straight down in the war years, $C = \alpha + \beta X + d_{waryears}\delta_W + \varepsilon$, where the new variable, $d_{waryears}$ equals one in 1942–1945 and zero in other years and $\delta_W < 0$.

One of the most useful aspects of the multiple regression model is its ability to identify the independent effects of a set of variables on a dependent variable. Example 2.2 describes a common application.

Example 2.2 Earnings and Education

A number of recent studies have analyzed the relationship between earnings and education. We would expect, on average, higher levels of education to be associated with higher incomes. The simple regression model

$$earnings = \beta_1 + \beta_2\, education + \varepsilon,$$

however, neglects the fact that most people have higher incomes when they are older than when they are young, regardless of their education. Thus, β_2 will overstate the marginal impact of education. If age and education are positively correlated, then the regression model will associate all the observed increases in income with increases in education. A better specification would account for the effect of age, as in

$$earnings = \beta_1 + \beta_2\, education + \beta_3\, age + \varepsilon.$$

It is often observed that income tends to rise less rapidly in the later earning years than in the early ones. To accommodate this possibility, we might extend the model to

$$earnings = \beta_1 + \beta_2\, education + \beta_3\, age + \beta_4\, age^2 + \varepsilon.$$

We would expect β_3 to be positive and β_4 to be negative.

The crucial feature of this model is that it allows us to carry out a conceptual experiment that might not be observed in the actual data. In the example, we might like to (and could)

compare the earnings of two individuals of the same age with different amounts of "education" even if the data set does not actually contain two such individuals. How education should be measured in this setting is a difficult problem. The study of the earnings of twins by Ashenfelter and Krueger (1994), which uses precisely this specification of the earnings equation, presents an interesting approach. We will examine this study in some detail in Section 12.5.2.[2] A large literature has been devoted to an intriguing question on this subject. Education is not truly "independent" in this setting. Highly motivated individuals will choose to pursue more education (for example, by going to college or graduate school) than others. By the same token, highly motivated individuals may do things that, on average, lead them to have higher incomes. If so, does a positive β_2 that suggests an association between income and education really measure the effect of education on income, or does it reflect the result of some underlying effect on both variables that we have not included in our regression model? We will revisit the issue in Chapter 24.

2.3 ASSUMPTIONS OF THE ~~CLASSICAL~~ LINEAR REGRESSION MODEL

The ~~classical~~ linear regression model consists of a set of assumptions about how a data set will be produced by an underlying "data-generating process." The theory will specify a deterministic relationship between the dependent variable and the independent variables. The assumptions that describe the form of the model and relationships among its parts and imply appropriate estimation and inference procedures are listed in Table 2.1.

TABLE 2.1 Assumptions of the ~~Classical~~ Linear Regression Model

A1. Linearity: $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i$. The model specifies a linear relationship between y and $x_1, \ldots, x_K$.

A2. Full rank: There is no exact linear relationship among any of the independent variables in the model. This assumption will be necessary for estimation of the parameters of the model.

A3. Exogeneity of the independent variables: $E[\varepsilon_i \mid x_{j1}, x_{j2}, \ldots, x_{jK}] = 0$. This states that the expected value of the disturbance at observation i in the sample is not a function of the independent variables observed at any observation, including this one. This means that the independent variables will not carry useful information for prediction of ε_i.

A4. Homoscedasticity and nonautocorrelation: Each disturbance, ε_i has the same finite variance, σ^2, and is uncorrelated with every other disturbance, ε_j. This assumption limits the generality of the model, and we will want to examine how to relax it in the chapters to follow.

A5. Data generation: The data in $(x_{j1}, x_{j2}, \ldots, x_{jK})$ may be any mixture of constants and random variables. The crucial elements for present purposes are the strict mean independence assumption A3 and the implicit variance independence assumption in A4. Analysis will be done conditionally on the observed $\mathbf{X}$, so whether the elements in $\mathbf{X}$ are fixed constants or random draws from a stochastic process will not influence the results. In later, more advanced treatments, we will want to be more specific about the possible relationship between ε_i and x_j.

A6. Normal distribution: The disturbances are normally distributed. Once again, this is a convenience that we will dispense with after some analysis of its implications.

[2] This model lays yet another trap for the practitioner. In a cross section, the higher incomes of the older individuals in the sample might tell an entirely different, perhaps macroeconomic story (a "cohort effect") from the lower incomes of younger individuals as time and their incomes evolve. It is not necessarily possible to deduce the characteristics of incomes of younger people in the sample *if they were older* by comparing the older individuals in the sample to the younger ones. A parallel problem arises in the analysis of treatment effects that we will examine in Chapter ~~24~~ 18.

(A)

insert on msp 2-6 where indicated

The experiment embodied in the earnings model thus far suggested is a comparison of two otherwise identical individuals who have different years of education. Under this interpretation, the "impact" of education would be $\partial E[Earnings|Age,Education]/\partial Education = \beta_2$. But, one might suggest that the experiment the analyst really has in mind is the truly unobservable impact of the additional year of education on a particular individual. To carry out the experiment, it would be necessary to observe the individual twice, once under circumstances that actually occur, $Education_i$, and a second time under the hypothetical (**counterfactual**) circumstance, $Education_i + 1$. If we consider *Education* in this example as a **treatment**, then the real objective of the experiment is to measure the **impact of the treatment on the treated**. The ability to infer this result from nonexperimental data that essentially compares "otherwise similar individuals" will be examined in Chapter 18.

Av. Term 'treatment' is not in chapter KT list. Add it.

KT KT KT KT

end of insert A

(B)

Insert on msp 2-6 where indicated

(Studies of twins and siblings have provided an interesting thread of research on the education and income relationship. Two other studies are Ashenfelter and Zimmerman (1997) and Bonjour, Cherkas, Haskel, Hawkes and Spector (2003).)

end of insert B

2.3.1 LINEARITY OF THE REGRESSION MODEL

Let the column vector $\mathbf{x}_k$ be the n observations on variable x_k, $k = 1, \ldots, K$, and assemble these data in an $n \times K$ data matrix $\mathbf{X}$. In most contexts, the first column of $\mathbf{X}$ is assumed to be a column of 1s so that β_1 is the constant term in the model. Let $\mathbf{y}$ be the n observations, $y_1, \ldots, y_n$, and let $\boldsymbol{\varepsilon}$ be the column vector containing the n disturbances. The model in (2-1) as it applies to all n observations can now be written

$$\mathbf{y} = \mathbf{x}_1\beta_1 + \cdots + \mathbf{x}_K\beta_K + \boldsymbol{\varepsilon}, \tag{2-2}$$

or in the form of Assumption 1,

$$\boxed{\text{ASSUMPTION:} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.} \tag{2-3}$$

A NOTATIONAL CONVENTION.

Henceforth, to avoid a possibly confusing and cumbersome notation, we will use a boldface $\mathbf{x}$ to denote a column *or* a row of $\mathbf{X}$. Which of these applies will be clear from the context. In (2-2), $\mathbf{x}_k$ is the kth column of $\mathbf{X}$. Subscripts i and k will be used to denote columns (variables). It will often be convenient to refer to a single observation in (2-3), which we would write

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i. \tag{2-4}$$

Subscripts i and t will generally be used to denote rows (observations) of $\mathbf{X}$. In (2-4), $\mathbf{x}_i$ is a column vector that is the transpose of the ith $1 \times K$ row of $\mathbf{X}$.

Our primary interest is in estimation and inference about the parameter vector $\boldsymbol{\beta}$. Note that the simple regression model in Example 2.1 is a special case in which $\mathbf{X}$ has only two columns, the first of which is a column of 1s. The assumption of linearity of the regression model includes the additive disturbance. For the regression to be linear in the sense described here, it must be of the form in (2-1) either in the original variables or after some suitable transformation. For example, the model

$$y = Ax^{\beta}e^{\varepsilon}$$

is linear (after taking logs on both sides of the equation), whereas

$$y = Ax^{\beta} + \varepsilon$$

is not. The observed dependent variable is thus the sum of two components, a deterministic element $\alpha + \beta x$ and a random variable ε. It is worth emphasizing that neither of the two parts is directly observed because α and β are unknown.

The linearity assumption is not so narrow as it might first appear. In the regression context, *linearity* refers to the manner in which the parameters and the disturbance enter the equation, not necessarily to the relationship among the variables. For example, the equations $y = \alpha + \beta x + \varepsilon$, $y = \alpha + \beta \cos(x) + \varepsilon$, $y = \alpha + \beta/x + \varepsilon$, and $y = \alpha + \beta \ln x + \varepsilon$ are all linear in some function of x by the definition we have used here. In the examples, only x has been transformed, but y could have been as well, as in $y = Ax^{\beta}e^{\varepsilon}$, which is a linear relationship in the logs of x and y: $\ln y = \alpha + \beta \ln x + \varepsilon$. The variety of functions is unlimited. This aspect of the model is used in a number of commonly used

functional forms. For example, the **loglinear model** is

$$\ln y = \beta_1 + \beta_2 \ln x_2 + \beta_3 \ln x_3 + \cdots + \beta_K \ln x_K + \varepsilon.$$

This equation is also known as the **constant elasticity** form as in this equation, the elasticity of y with respect to changes in x is $\partial \ln y / \partial \ln x_k = \beta_k$, which does not vary with x_k. The loglinear form is often used in models of demand and production. Different values of β produce widely varying functions.

Example 2.3 The U.S. Gasoline Market

Data on the U.S. gasoline market for the years 1953–2004 are given in Table F2.2 in Appendix F. We will use these data to obtain, among other things, estimates of the income, own price, and cross-price elasticities of demand in this market. These data also present an interesting question on the issue of holding "all other things constant," that was suggested in Example 2.2. In particular, consider a somewhat abbreviated model of per capita gasoline consumption:

$$\ln(G/pop) = \beta_1 + \beta_2 \ln(Income/pop) + \beta_3 \ln price_G + \beta_4 \ln P_{newcars} + \beta_5 \ln P_{usedcars} + \varepsilon.$$

This model will provide estimates of the income and price elasticities of demand for gasoline and an estimate of the elasticity of demand with respect to the prices of new and used cars. What should we expect for the sign of β_4? Cars and gasoline are complementary goods, so if the prices of new cars rise, ceteris paribus, gasoline consumption should fall. Or should it? If the prices of new cars rise, then consumers will buy fewer of them; they will keep their used cars longer and buy fewer new cars. If older cars use more gasoline than newer ones, then the rise in the prices of new cars would lead to higher gasoline consumption than otherwise, not lower. We can use the multiple regression model and the gasoline data to attempt to answer the question.

A **semilog** model is often used to model growth rates:

$$\ln y_t = \mathbf{x}_t'\boldsymbol{\beta} + \delta t + \varepsilon_t.$$

In this model, the autonomous (at least not explained by the model itself) proportional, per period growth rate is $d \ln y / dt = \delta$. Other variations of the general form

$$f(y_t) = g(\mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t)$$

will allow a tremendous variety of functional forms, all of which fit into our definition of a linear model.

The linear regression model is sometimes interpreted as an approximation to some unknown, underlying function. (See Section A.8.1 for discussion.) By this interpretation, however, the linear model, even with quadratic terms, is fairly limited in that such an approximation is likely to be useful only over a small range of variation of the independent variables. The translog model discussed in Example 2.4, in contrast, has proved far more effective as an approximating function.

Example 2.4 The Translog Model

Modern studies of demand and production are usually done with a **flexible functional form.** Flexible functional forms are used in econometrics because they allow analysts to model ~~second-order effects~~ such as elasticities of substitution, which are functions of the second derivatives of production, cost, or utility functions. The linear model restricts these to equal zero, whereas the loglinear model (e.g., the Cobb–Douglas model) restricts the interesting elasticities to the uninteresting values of −1 or +1. The most popular flexible functional form is the **translog model,** which is often interpreted as a second-order approximation to an

complex features of the production function

2-10

unknown functional form. [See Berndt and Christensen (1973).] One way to derive it is as follows. We first write $y = g(x_1, \ldots, x_K)$. Then, $\ln y = \ln g(\ldots) = f(\ldots)$. Since by a trivial transformation $x_k = \exp(\ln x_k)$, we interpret the function as a function of the logarithms of the x's. Thus, $\ln y = f(\ln x_1, \ldots, \ln x_K)$.

Now, expand this function in a second-order Taylor series around the point $\mathbf{x} = [1, 1, \ldots, 1]'$ so that at the expansion point, the log of each variable is a convenient zero. Then

$$\ln y = f(\mathbf{0}) + \sum_{k=1}^{K} [\partial f(\cdot)/\partial \ln x_k]_{|\ln x=0} \ln x_k + \frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{K} [\partial^2 f(\cdot)/\partial \ln x_k \partial \ln x_l]_{|\ln x=0} \ln x_k \ln x_l + \varepsilon.$$

The disturbance in this model is assumed to embody the familiar factors and the error of approximation to the unknown function. Since the function and its derivatives evaluated at the fixed value $\mathbf{0}$ are constants, we interpret them as the coefficients and write

$$\ln y = \beta_0 + \sum_{k=1}^{K} \beta_k \ln x_k + \frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{K} \gamma_{kl} \ln x_k \ln x_l + \varepsilon.$$

This model is linear by our definition but can, in fact, mimic an impressive amount of curvature when it is used to approximate another function. An interesting feature of this formulation is that the loglinear model is a special case, $\gamma_{kl} = 0$. Also, there is an interesting test of the underlying theory possible because if the underlying function were assumed to be continuous and twice continuously differentiable, then by Young's theorem it must be true that $\gamma_{kl} = \gamma_{lk}$. We will see in Chapter 10 how this feature is studied in practice.

replace with (c) next page msp 2-11

Despite its great flexibility, the linear model does not include all the situations we encounter in practice. For a simple example, there is no transformation that will reduce $y = \alpha + 1/(\beta_1 + \beta_2 x) + \varepsilon$ to linearity. The methods we consider in this chapter are not appropriate for estimating the parameters of such a model. Relatively straightforward techniques have been developed for nonlinear models such as this, however. We shall treat them in detail in Chapter 11.

2.3.2 FULL RANK

Assumption 2 is that there are no exact linear relationships among the variables.

ASSUMPTION: $\mathbf{X}$ is an $n \times K$ matrix with rank K.	(2-5)

Hence, $\mathbf{X}$ has full column rank; the columns of $\mathbf{X}$ are linearly independent and there are at least K observations. [See (A-42) and the surrounding text.] This assumption is known as an **identification condition.** To see the need for this assumption, consider an example.

italic

Example 2.5 Short Rank

Suppose that a cross-section model specifies that consumption, C, relates to income as follows:

$$C = \beta_1 + \beta_2 \textit{ nonlabor income} + \beta_3 \textit{ salary} + \beta_4 \textit{ total income} + \varepsilon,$$

Insert on msp 2-10 where indicated

(c)

Example 14.10 and Chapter 19

~~Chapter 14 and 19~~,

2-11

Despite its great flexibility, the linear model will not accommodate all the situations we will encounter in practice. In ~~Example XX.XX~~, we will examine the regression model for doctor visits that was suggested in the introduction to this chapter. An appropriate model that describes the number of visits has conditional mean function $E[y|\mathbf{x}] = \exp(\mathbf{x}_i'\boldsymbol{\beta})$. It is tempting to linearize this directly by taking logs, since $\ln E[y|\mathbf{x}] = \mathbf{x}_i'\boldsymbol{\beta}$. But, $\ln E[y|\mathbf{x}]$ is not equal to $E[\ln y|\mathbf{x}]$. In that setting, y_i can equal zero (and does for most of the sample), so $\mathbf{x}_i'\boldsymbol{\beta}$ (which can be negative) is not an appropriate model for $\ln y_i$ (which does not exist) nor for y_i which cannot be negative. C

end of insert c

where *total income* is exactly equal to *salary* plus *nonlabor income*. Clearly, there is an exact linear dependency in the model. Now let

$$\beta_2' = \beta_2 + a,$$
$$\beta_3' = \beta_3 + a,$$

and

$$\beta_4' = \beta_4 - a,$$

where a is any number. Then the exact same value appears on the right-hand side of C if we substitute β_2', β_3', and β_4' for β_2, β_3, and β_4. Obviously, there is no way to estimate the parameters of this model.

If there are fewer than K observations, then $\mathbf{X}$ cannot have **full rank.** Hence, we make the (redundant) assumption that n is at least as large as K.

In a two-variable linear model with a constant term, the full rank assumption means that there must be variation in the regressor x. If there is no variation in x, then all our observations will lie on a vertical line. This situation does not invalidate the other assumptions of the model; presumably, it is a flaw in the data set. The possibility that this suggests is that we *could* have drawn a sample in which there was variation in x, but in this instance, we did not. Thus, the model still applies, but we cannot learn about it from the data set in hand.

insert next page (D) msp 2-13

2.3.3 REGRESSION

The disturbance is assumed to have conditional expected value zero at every observation, which we write as

$$E[\varepsilon_i \mid \mathbf{X}] = 0. \tag{2-6}$$

For the full set of observations, we write Assumption 3 as:

$$\text{ASSUMPTION:} \quad E[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \begin{bmatrix} E[\varepsilon_1 \mid \mathbf{X}] \\ E[\varepsilon_2 \mid \mathbf{X}] \\ \vdots \\ E[\varepsilon_n \mid \mathbf{X}] \end{bmatrix} = \mathbf{0}. \tag{2-7}$$

There is a subtle point in this discussion that the observant reader might have noted. In (2-7), the left-hand side states, in principle, that the mean of each ε_i *conditioned on all observations* $\mathbf{x}_i$ is zero. This conditional mean assumption states, in words, that no observations on $\mathbf{x}$ convey information about the expected value of the disturbance. It is conceivable—for example, in a time-series setting—that although $\mathbf{x}_i$ might provide no information about $E[\varepsilon_i \mid \cdot]$, $\mathbf{x}_j$ *at some other observation,* such as in the next time period, might. Our assumption at this point is that there is no information about $E[\varepsilon_i \mid \cdot]$ contained in any observation $\mathbf{x}_j$. Later, when we extend the model, we will study the implications of dropping this assumption. [See Wooldridge (1995).] We will also assume that the disturbances convey no information about each other. That is, $E[\varepsilon_i \mid \varepsilon_1, \ldots, \varepsilon_{i-1}, \varepsilon_{i+1}, \ldots, \varepsilon_n] = 0$. In sum, at this point, we have assumed that the disturbances are purely random draws from some population.

(D)

3.4

insert on msp 2-12 where indicated

Example 2.6 An Inestimable Model

In Example 3.4, we will consider a model for the sale price of Monet paintings. Theorists and observers have different models for how prices of paintings at auction are determined. One (naïve) student of the subject suggests the model

um/aut

$$\ln Price = \beta_1 + \beta_2 \ln Size + \beta_3 \ln Aspect\ Ratio + \beta_4 \ln Height + \varepsilon$$
$$= \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon,$$

where *Size* = *Width* times *Height* and *Aspect Ratio* = *Width/Height*. By simple arithmetic, we can see that this model shares the problem found with the consumption model in Example 2.5 – in this case, $x_2 - x_4 = x_3 + x_4$. So, this model is, like the previous one, not estimable – it is not identified. It is useful to think of the problem from a different perspective here (so to speak). In the linear model, it must be possible for the variables to vary linearly independently. But, in this instance, while it is possible for any pair of the three covariates to vary independently, the three together cannot. The "model," i.e., the theory, is an entirely reasonable model as it stands. Art buyers might very well consider all three of these features in their valuation of a Monet painting. However, it is not possible to learn about that from the observed data, at least not with *this* linear regression model.

that is,

AV: OK to spell out "i.e." in text?

end of insert D

The zero conditional mean implies that the unconditional mean is also zero, since

$$E[\varepsilon_i] = E_{\mathbf{x}}[E[\varepsilon_i \mid \mathbf{X}]] = E_{\mathbf{x}}[0] = 0.$$

Since, for each ε_i, $\text{Cov}[E[\varepsilon_i \mid \mathbf{X}], \mathbf{X}] = \text{Cov}[\varepsilon_i, \mathbf{X}]$, Assumption 3 implies that $\text{Cov}[\varepsilon_i, \mathbf{X}] = 0$ for all i. (Exercise: Is the converse true?)

In most cases, the zero mean assumption is not restrictive. Consider a two-variable model and suppose that the mean of ε is $\mu \neq 0$. Then $\alpha + \beta x + \varepsilon$ is the same as $(\alpha + \mu) + \beta x + (\varepsilon - \mu)$. Letting $\alpha' = \alpha + \mu$ and $\varepsilon' = \varepsilon - \mu$ produces the original model. For an application, see the discussion of frontier production functions in Section 18.2.3. But, if the original model does not contain a constant term, then assuming $E[\varepsilon_i] = 0$ could be substantive. This suggests that there is a potential problem in models without constant terms. As a general rule, regression models should not be specified without constant terms unless this is specifically dictated by the underlying theory.[3] Arguably, if we have reason to specify that the mean of the disturbance is something other than zero, we should build it into the systematic part of the regression, leaving in the disturbance only the unknown part of ε. Assumption 3 also implies that

$$E[\mathbf{y} \mid \mathbf{X}] = \mathbf{X}\boldsymbol{\beta}. \tag{2-8}$$

Assumptions 1 and 3 comprise the *linear regression model*. The **regression** of $\mathbf{y}$ on $\mathbf{X}$ is the conditional mean, $E[\mathbf{y} \mid \mathbf{X}]$, so that without Assumption 3, $\mathbf{X}\boldsymbol{\beta}$ is *not* the conditional mean function.

The remaining assumptions will more completely specify the characteristics of the disturbances in the model and state the conditions under which the sample observations on $\mathbf{x}$ are obtained.

2.3.4 SPHERICAL DISTURBANCES

The fourth assumption concerns the variances and covariances of the disturbances:

$$\text{Var}[\varepsilon_i \mid \mathbf{X}] = \sigma^2, \qquad \text{for all } i = 1, \ldots, n,$$

and

$$\text{Cov}[\varepsilon_i, \varepsilon_j \mid \mathbf{X}] = 0, \qquad \text{for all } i \neq j.$$

Constant variance is labeled **homoscedasticity.** Consider a model that describes the profits of firms in an industry as a function of, say, size. Even accounting for size, measured in dollar terms, the profits of large firms will exhibit greater variation than those of smaller firms. The homoscedasticity assumption would be inappropriate here. Also, survey data on household expenditure patterns often display marked **heteroscedasticity,** even after accounting for income and household size.

Uncorrelatedness across observations is labeled generically **nonautocorrelation.** In Figure 2.1, there is some suggestion that the disturbances might not be truly independent across observations. Although the number of observations is limited, it does appear that, on average, each disturbance tends to be followed by one with the same sign. This

[3] Models that describe first differences of variables might well be specified without constants. Consider $y_t - y_{t-1}$. If there is a constant term α on the right-hand side of the equation, then y_t is a function of αt, which is an explosive regressor. Models with linear time trends merit special treatment in the time-series literature. We will return to this issue in Chapter 20.

(E)

Insert on msp2-14 where indicated

The converse is not true; $E[\varepsilon_i] = 0$ does not imply that $E[\varepsilon_i | \mathbf{x}_i] = 0$. Example 2.7 illustrates the difference.

Example 2.7 Nonzero Conditional Mean of the Disturbances

FIG 2.2

Figure 2.2 illustrates the important difference between $E[\varepsilon_i] = 0$ and $E[\varepsilon_i | x_i] = 0$. The overall mean of the disturbances in the sample is zero, but the mean for specific ranges of x is distinctly nonzero. A pattern such as this in observed data would serve as a useful indicator that the assumption of the linear regression should be questioned. In this particular case, the true conditional mean function (which the researcher would not know in advance) is actually $E[y|x] = 1 + \exp(1.5x)$. The sample data are suggesting that the linear model is not appropriate for these data. This possibility is pursued in an application in Example 5.2 6.6.

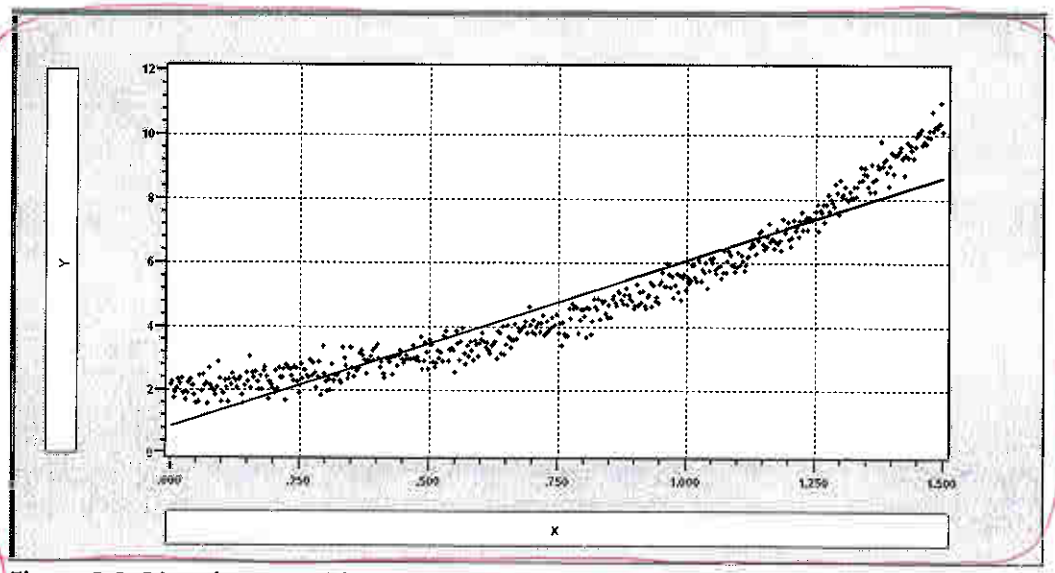

Figure 2.2 Disturbances with nonzero conditional mean and zero unconditional mean.

end of insert E

"inertia" is precisely what is meant by **autocorrelation,** and it is assumed away at this point. Methods of handling autocorrelation in economic data occupy a large proportion of the literature and will be treated at length in Chapter 19. Note that nonautocorrelation does not imply that observations y_i and y_j are uncorrelated. The assumption is that *deviations* of observations from their expected values are uncorrelated.

The two assumptions imply that

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] = \begin{bmatrix} E[\varepsilon_1\varepsilon_1 \mid \mathbf{X}] & E[\varepsilon_1\varepsilon_2 \mid \mathbf{X}] & \cdots & E[\varepsilon_1\varepsilon_n \mid \mathbf{X}] \\ E[\varepsilon_2\varepsilon_1 \mid \mathbf{X}] & E[\varepsilon_2\varepsilon_2 \mid \mathbf{X}] & \cdots & E[\varepsilon_2\varepsilon_n \mid \mathbf{X}] \\ \vdots & \vdots & \vdots & \vdots \\ E[\varepsilon_n\varepsilon_1 \mid \mathbf{X}] & E[\varepsilon_n\varepsilon_2 \mid \mathbf{X}] & \cdots & E[\varepsilon_n\varepsilon_n \mid \mathbf{X}] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix},$$

which we summarize in Assumption 4:

$$\boxed{\text{ASSUMPTION:} \quad E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] = \sigma^2\mathbf{I}.} \tag{2-9}$$

By using the variance decomposition formula in (B-69), we find

$$\text{Var}[\boldsymbol{\varepsilon}] = E[\text{Var}[\boldsymbol{\varepsilon} \mid \mathbf{X}]] + \text{Var}[E[\boldsymbol{\varepsilon} \mid \mathbf{X}]] = \sigma^2\mathbf{I}.$$

Once again, we should emphasize that this assumption describes the information about the variances and covariances among the disturbances that is provided by the independent variables. For the present, we assume that there is none. We will also drop this assumption later when we enrich the regression model. We are also assuming that the disturbances themselves provide no information about the variances and covariances. Although a minor issue at this point, it will become crucial in our treatment of time-series applications. Models such as $\text{Var}[\varepsilon_t \mid \varepsilon_{t-1}] = \sigma^2 + \alpha\varepsilon_{t-1}^2$, a "GARCH" model (see Chapter 19), do not violate our conditional variance assumption, but do assume that $\text{Var}[\varepsilon_t \mid \varepsilon_{t-1}] \neq \text{Var}[\varepsilon_t]$.

Disturbances that meet the twin assumptions of homoscedasticity and nonautocorrelation are sometimes called **spherical disturbances.**[4]

2.3.5 DATA GENERATING PROCESS FOR THE REGRESSORS

It is common to assume that $\mathbf{x}_i$ is nonstochastic, as it would be in an experimental situation. Here the analyst chooses the values of the regressors and then observes y_i. This process might apply, for example, in an agricultural experiment in which y_i is yield and $\mathbf{x}_i$ is fertilizer concentration and water applied. The assumption of **nonstochastic regressors** at this point would be a mathematical convenience. With it, we could use

[4] The term will describe the multivariate normal distribution; see (B-95). If $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ in the multivariate normal density, then the equation $f(\mathbf{x}) = c$ is the formula for a "ball" centered at $\boldsymbol{\mu}$ with radius σ in n-dimensional space. The name *spherical* is used whether or not the normal distribution is assumed; sometimes the "spherical normal" distribution is assumed explicitly.

2-17

the results of elementary statistics to obtain our results by treating the vector $\mathbf{x}_i$ simply as a known constant in the probability distribution of y_i. With this simplification, Assumptions A3 and A4 would be made unconditional and the counterparts would now simply state that the probability distribution of ε_i involves none of the constants in $\mathbf{X}$.

Social scientists are almost never able to analyze experimental data, and relatively few of their models are built around nonrandom regressors. Clearly, for example, in any model of the macroeconomy, it would be difficult to defend such an asymmetric treatment of aggregate data. Realistically, we have to allow the data on $\mathbf{x}_i$ to be random the same as y_i, so an alternative formulation is to assume that $\mathbf{x}_i$ is a random vector and our formal assumption concerns the nature of the random process that produces $\mathbf{x}_i$. If $\mathbf{x}_i$ is taken to be a random vector, then Assumptions 1 through 4 become a statement about the joint distribution of y_i and $\mathbf{x}_i$. The precise nature of the regressor and how we view the sampling process will be a major determinant of our derivation of the statistical properties of our estimators and test statistics. In the end, the crucial assumption is Assumption 3, the uncorrelatedness of $\mathbf{X}$ and $\boldsymbol{\varepsilon}$. Now, we do note that this alternative is not completely satisfactory either, since $\mathbf{X}$ may well contain nonstochastic elements, including a constant, a time trend, and dummy variables that mark specific episodes in time. This makes for an ambiguous conclusion, but there is a straightforward and economically useful way out of it. We will assume that $\mathbf{X}$ can be a mixture of constants and random variables, and the mean and variance of ε_i are both independent of all elements of $\mathbf{X}$.

ASSUMPTION: $\mathbf{X}$ may be fixed or random. **(2-10)**

2.3.6 NORMALITY

KT

It is convenient to assume that the disturbances are **normally distributed,** with zero mean and constant variance. That is, we add normality of the distribution to Assumptions 3 and 4.

ASSUMPTION: $\boldsymbol{\varepsilon} \mid \mathbf{X} \sim N[\mathbf{0}, \sigma^2\mathbf{I}]$. **(2-11)**

In view of our description of the source of $\boldsymbol{\varepsilon}$, the conditions of the central limit theorem will generally apply, at least approximately, and the normality assumption will be reasonable in most settings. A useful implication of Assumption 6 is that it implies that observations on ε_i are statistically independent as well as uncorrelated. [See the third point in Section B.9, (B-97) and (B-99).] **Normality** is often viewed as an unnecessary and possibly inappropriate addition to the regression model. Except in those cases in which some alternative distribution is explicitly assumed, as in the stochastic frontier model discussed in ~~Section 16.9.5.a~~, the normality assumption is probably quite reasonable.

KT

Chapter 18,

Normality is not necessary to obtain many of the results we use in multiple regression analysis, although it will enable us to obtain several exact statistical results. It does prove useful in constructing test statistics, as shown in Section 4.7. Later, it will be possible to relax this assumption and retain most of the statistical results we obtain here. (See Sections ~~4.9~~ 4.4 and ~~5.4~~ 5.6.) confidence intervals and 4.5 and chapter 5.

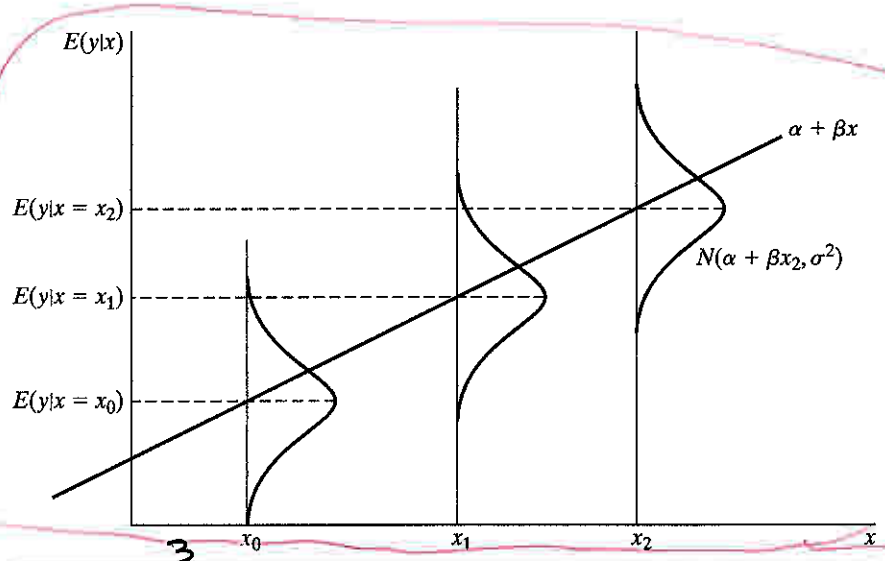

FIGURE 2.2 The Classical Regression Model

insert next page → (F) msp 2-19

2.4 SUMMARY AND CONCLUSIONS

This chapter has framed the linear regression model, the basic platform for model building in econometrics. The assumptions of the classical regression model are summarized in Figure 2.2, which shows the two-variable case.

FIG 2.3

Key Terms and Concepts

- Autocorrelation
- Constant elasticity
- Covariate
- Dependent variable
- Deterministic relationship
- Disturbance
- Exogeneity
- Explained variable
- Explanatory variable
- Flexible functional form
- Full rank
- Heteroscedasticity
- Homoscedasticity
- Identification condition
- Independent variable
- Linear regression model
- Loglinear model
- Multiple linear regression model
- Nonautocorrelation
- Nonstochastic regressors
- Normality
- Normally distributed
- Population regression equation
- Regressand
- Regression
- Regressor
- Second-order effects
- Semilog
- Spherical disturbances
- Translog model

AU: Terms "exogeneity" & "second-order effects" are not bold KTs in text. Delete them?

- Conditional median
- Conditional variation
- Counterfactual
- Impact of treatment on the treated
- Linear independence
- Mean independence
- Path diagram

(F)

Insert on msp 248 where indicated

2-19 / End 2

2.3.7 INDEPENDENCE

The term "independent" has been used several ways in this chapter.

In Section 2.2, the right hand side variables in the model are denoted the independent variables. Here, the notion of independence refers to the sources of variation. In the context of the model, the variation in the independent variables arises from sources that are outside of the process being described. Thus, in our health services vs. income example in the introduction, we have suggested a theory for how variation in demand for services is associated with variation in income. But, we have not suggested an explanation of the sample variation in incomes; income is assumed to vary for reasons that are outside the scope of the model.

The assumption in equation (2-6), $E[\varepsilon_i|\mathbf{X}] = 0$, is "**mean independence.**" Its implication is that variation in the disturbances in our data is not explained by variation in the independent variables. We have also assumed in Section 2.3.4 that the disturbances are uncorrelated with each other (Assumption A4 in Table 2.1). This implies that $E[\varepsilon_i|\varepsilon_j] = 0$ when $i \neq j$ – the disturbances are also mean independent of each other. Conditional normality of the disturbances assumed in Section 2.3.6 (Assumption A6) implies that they are **statistically independent** of each other, which is a stronger result than mean independence.

AV: term "statistically independent not in KT list. Addit.

Finally, Section 2.3.2 discusses the **linear independence** of the columns of the data matrix, **X**. The notion of independence here is an algebraic one relating to the column rank of **X**. In this instance, the underlying interpretation is that it must be possible for the variables in the model to vary linearly independently of each other. Thus, in example 2.6, we find that it is not possible for the logs of surface area, aspect ratio and height of a painting all to vary independently of one another. The modeling implication is that if the variables cannot vary independently of each other, then it is not possible to analyze them in a linear regression model that assumes the variables can each vary while holding the others constant. There is an ambiguity in this discussion of independence of the variables. We have both age and age squared in a model in Example 2.2. These cannot vary independently, but there is no obstacle to formulating a regression model containing both age and age squared. The resolution is that age and age squared, though not *functionally* independent, are *linearly* independent. That is the crucial assumption in the linear regression model.

end of insert F