PART II

LINEAR REGRESSION

This part of the book weds statistical methodology to the OLS technique. The fundamental difference between what has gone before and what is to come is that we build our analysis on probabilistic assumptions about the way the data are generated.

All of our previous analysis of OLS focused on geometric properties of the procedure. Such properties describe the nature of the fit and help us understand how OLS summarizes an entire data set. But these properties do not answer another set of questions encountered by those who collect and analyze data: what do the data "say" about the process that generated them? What can be inferred about the world in general from particular observations? Under what conditions is OLS useful for such inference? The rest of this book describes some of the ways statisticians and econometricians have narrowed these questions so that answers could be obtained. This part of the book focuses on generalizing the OLS analysis to such questions.

We assume that the reader is familiar with such concepts from probability as mean and variance and such basic statistical theory as estimation of a population mean and hypothesis testing for equality of the population means of two sampling experiments.

Let us summarize the simple location model in which the average is the central statistic. We intend this summary to be a brief review of material that is already familiar to the reader and to establish a common point of departure for the rest of this part of the book. Many of the concepts and results that one meets in this model have counterparts in the linear regression model and an increased understanding of linear regression will result from keeping the analysis of the location model in mind.

In the simple location model, interest focuses on the marginal mean of a random variable Y. Inference follows from a random sample of observations of Y, denoted $\{y_1, \ldots, y_N\}$. All statistical inference rests on assumptions or beliefs about the process that generates the sample data set. The classical assumptions are listed in Table II.1. The entries of the table follow the order in which the assumptions are often considered. The consequences in the second column follow from all the assumptions listed in the corresponding row in the table and the rows above.

Table II.1
Summary of Assumptions and Results for the Location Model

Assumptions	Results
$\mathrm{E}[y_n] = \beta_0$	• $\mathrm{E}[\hat{\beta}] = \beta_0$ for $\hat{\beta} = N^{-1} \sum_{n=1}^{N} y_n \equiv \bar{y}$
$\mathrm{Var}[y_n] = \sigma_0^2$, $\mathrm{Cov}[y_n, y_m] = 0$, n $\neq$ m	• $\mathrm{Var}[\hat{\beta}] = \sigma_0^2/N$ • $\mathrm{E}[s^2] = \sigma_0^2$, where $s^2 = \sum_{n=1}^{N} (y_n - \bar{y})^2/(N-1)$ • $\hat{\beta}$ is a minimum variance linear unbiased estimator
$y_n \sim \mathfrak{N}(\beta_0, \sigma_0^2)$	• $\sqrt{N}(\hat{\beta} - \beta_0)/\sigma_0 \xrightarrow{d} \mathfrak{N}(0, 1)$ • $\hat{\beta} \sim \mathfrak{N}(\beta_0, \sigma_0^2/N)$ • $s^2 \sim \chi^2_{N-1}\sigma_0^2/(N-1)$ • $\hat{\beta}$ and s^2 are independent • $[\hat{\beta}, (N-1)s^2/N]$ is the maximum likelihood estimator

Note especially that this analysis rests largely on the simple mathematical structure of the statistic $\hat{\beta}$, which is the average of the $\{y_n\}$. In mathematical terms, $\bar{y}$ is a *linear* function of the $\{y_n\}$. Because $\hat{\beta}$ is a sum of random variables, its mean and its variance are relatively easy to derive. This linearity is also fundamental to the normality of $\hat{\beta}$ under the assumption of normally distributed $\{y_n\}$: sums of normal random variables are also normally distributed.

The results involving s^2 are somewhat paradoxical. This statistic is the sum of squared, normally distributed, random variables—just as one would expect for a chi-square random variable. However, there are N, not $N-1$, elements in the sum; one might expect the degrees of freedom to be N instead of $N-1$. Furthermore, the normal random variables $\{y_n - \hat{\beta}\}$ are not independently distributed, as the standard motivation of a chi-square distribution requires. In fact, the resolution of the paradox lies in accounting for this dependence. The independence of s^2 and $\hat{\beta}$ is a second surprise. One might casually predict that these statistics are dependently distributed because they both depend on $\{y_n; n = 1, \ldots, N\}$. But, of course, this turns out to be incorrect.

It may be convenient to remember the connections between assumptions and consequences in terms of the nature of each assumption. The first is an assumption about the *first moment* of the data, and from it follow first-moment consequences: we have an unbiased estimator of the first moment. The second assumption is about the second moments of the data, and from it (and the first assumption) follow second-moment consequences: we obtain the second moment of our estimator, a second-moment optimality result, and an estimator of a second moment. Finally, the third assumption is about the distribution of the data, and from it (and the previous assumptions) follow distributional consequences: we obtain the actual distributions of our statistics.

Now let us compare the simple location model with ordinary least squares and the linear regression model in matrix notation. Compare the first column of Table II.2 with the entries in Table II.1 and see that the entries below are simply restatements in a new notation. Then compare the two columns of Table II.2 and see how similar the entries are. Matrix products replace scalar sums and a matrix inverse replaces a scalar reciprocal. In this table, we have not emphasized the relationships between assumptions and consequences, nor have we given a complete list of consequences. Our purpose is simply to introduce the linear regression model as a multivariate generalization of the location model and to provide an indication of coming results.

Table II.2
Analogues in the Location and Regression Models

Location Model	Linear Regression
Model Assumptions	
$\mathrm{E}\left[\mathbf{y}\right] = \boldsymbol{\iota}\beta_0$	$\mathrm{E}\left[\mathbf{y} \mid \mathbf{X}\right] = X\boldsymbol{\beta}_0$
$\mathrm{Var}\left[\mathbf{y}\right] = \sigma_0^2 \cdot \mathbf{I}$	$\mathrm{Var}\left[\mathbf{y} \mid \mathbf{X}\right] = \sigma_0^2 \cdot \mathbf{I}$
$\mathbf{y} \sim \mathfrak{N}(\beta_0, \sigma_0^2 \cdot \mathbf{I})$	$\mathbf{y} \mid \mathbf{X} \sim \mathfrak{N}(\mathbf{X}\boldsymbol{\beta}_0, \sigma_0^2 \cdot \mathbf{I})$
Analysis	
$\hat{\beta} = \dfrac{\boldsymbol{\iota}' y}{\boldsymbol{\iota}'\boldsymbol{\iota}} = \bar{y}$	$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
$\mathrm{E}\left[\hat{\beta}\right] = \beta_0$	$\mathrm{E}\left[\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right] = \boldsymbol{\beta}_0$
$\mathrm{Var}\left[\hat{\beta}\right] = \dfrac{\sigma_0^2}{\boldsymbol{\iota}'\boldsymbol{\iota}}$	$\mathrm{Var}\left[\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right] = \sigma_0^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}$
$\hat{\beta} \sim \mathfrak{N}[\beta_0, \sigma_0^2/(\boldsymbol{\iota}'\boldsymbol{\iota})]$	$\hat{\boldsymbol{\beta}} \mid \mathbf{X} \sim \mathfrak{N}[\boldsymbol{\beta}_0,\ \sigma_0^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}]$
$s^2 = \dfrac{\mathbf{y}'\left[\mathbf{I} - (\iota\iota'/\iota'\iota)\right] y}{N-1}$	$s^2 = \dfrac{\mathbf{y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}}{N-K}$
$s^2 \sim \dfrac{\chi_{N-1}^2 \sigma_0^2}{N-1}$	$s^2 \sim \dfrac{\chi_{N-K}^2 \sigma_0^2}{N-K}$