**Chapter 1 Summary**

Now that we have presented the basics of the CLRM, where do we go from here? The answer follows.

The wage regression given in Table 1.2 is based on the assumptions of the CLRM. The question that naturally arises is: how do we know that this model satisfies the assumptions of the CLRM? We need to know answers to the following questions:

**1** The wage model given in Table 1.2 is linear in variables as well as parameters. Could the wage variable, for instance, be in logarithmic form? Could the variables for education and experience be also in logarithmic form? Since wages are not expected to grow linearly with experience forever, could we include experience squared as an additional regressor? All these questions pertain to the **functional** **form** of the regression model, and there are several of them. We consider this topic in Chapter 2.

**2** Suppose some of the regressors are quantitative and some are qualitative or nominal scale variables, also called dummy variables. Are there special problems in dealing with dummy variables? How do we handle the interaction between quantitative and dummy variables in a given situation? In our wage regression we have three dummy variables, *female*, *nonwhite*, and *union*. Do female union workers earn more than non-union female workers? We will deal with this and other aspects of **qualitative regressors** in Chapter 3.

**3** If we have several regressors in a regression model, how do we find out that we do not have the problem of **multicollinearity**? If we have that problem, what are the consequences? And how do we deal with them? We discuss this topic in Chapter 4.

**4** In cross-sectional data the error variance may be **heteroscedastic** rather than homoscedastic. How do we find that out? And what are the consequences of heteroscedasticity? Are OLS estimators still BLUE? How do we correct for heteroscedasticity? We answer these questions in Chapter 5.

**5** In time series data the assumption of no **autocorrelation** in the error term is unlikely to be fulfilled. How do we find that out? What are the consequences of autocorrelation? How do we correct for autocorrelation? We will answer these questions in Chapter 6.

**6** One of the assumptions of the CLRM is that the model used in empirical analysis is “correctly specified” in the sense that all relevant variables are included in the model, no superfluous variables are included in the model, the probability distribution of the error term is correctly specified, and there are no errors of measurement in the regressors and regressand. Obviously, this is a tall order. But it is important that we find out the consequences of one or more of these situations if they are suspected in a concrete application. We discuss the **problem** **of model specification** in some detail in Chapter 7. We also discuss briefly in this chapter the case of stochastic regressors instead of fixed regressors, as assumed in the CLRM. This topic is discussed more extensively in Chapter 19. In this chapter we also discuss the topics of simultaneous equations models and dynamic regression models.

**7** Suppose the dependent variable is not a ratio or interval scale variable but is a nominal scale variable, taking values of 1 and 0. Can we still apply the usual OLS techniques to estimate such models? If not, what are the alternatives? The answer to these questions can be found in Chapter 8, where we discuss the **logit** and **probit** models, which can handle a nominal dependent variable.

**8** Chapter 9 extends the bivariate logit and probit models to multi-category nominal scale variables, where the regressand has more than two nominal values. For example, consider the means of transportation to work. Suppose we have three choices: private car, public bus, or train. How do we decide among these choices? Can we still use OLS? As we will show in this chapter, such problems require nonlinear estimation techniques. **Multinomial conditional logit or multinomial** **probit models** discussed in this chapter show how multi-category nominal scale variables can be modeled.

**9** Although nominal scale variables cannot be readily quantified, they can sometimes be ordered or ranked. **Ordered logit** and **ordered probit models**, discussed in Chapter 10, show how ordered or ranked models can be estimated.

**10** Sometimes the regressand is restricted in the values it takes because of the design of the problem under study. Suppose we want to study expenditure on housing by families making income under $50,000 a year. Obviously, this excludes families with income over this limit. The **censored sample** and **truncated sample** **modeling** discussed in Chapter 11 show how we can model phenomena such as this.

I

**11** Occasionally we come across data that is of the *count type*, such as the number of visits to a doctor, the number of patents received by a firm, the number customers passing through a check-out counter in a span of 15 minutes, and so on. To model such count data, the **Poisson probability distribution** (**PPD**) is often used. Because the assumption underlying the PPD may not always be fulfilled, we will discuss briefly an alternative model, knows as the **negative binomial** **distribution** (**NBD**). We discuss these topics in Chapter 12.

**12** In cases of time series data, an underlying assumption of the CLRM is that the time series are **stationary**. If that is not the case, is the usual OLS methodology still applicable? What are the alternatives? We discuss this topic in Chapter 13.

**13** If you regress a nonstationary time series on one or more nonstationary time series, it might lead to the so-called **spurious** or **nonsense regression phenomenon**. However, if there is a stable long-term relationship between variables, that is if the variables are **cointegrated**, there need not be spurious regression. In Chapter 14 we show how we find this out and what happens if the variables are not cointegrated.

**14** Although heteroscedasticity is generally associated with cross-sectional data, it can also arise in time series data in the so-called **volatility** clustering phenomenon observed in financial time series. The **ARCH** and **GARCH** models discussed in Chapter 15 will show how we model volatility clustering.

**15** Forecasting is a specialized field in time series econometrics. In Chapter 16 we discuss the topic of economic forecasting using the LRM as well as two prominently used methods of forecasting, namely, **ARIMA** (autoregressive integrated moving average) and **VAR** (vector autoregression). With examples, we show how these models work.

**16** The models discussed in the preceding chapters dealt with cross-sectional or time series data. Chapter 17 deals with models that combine cross-sectional and time series data. Th ese models are known as **panel data regression models**. We show in this chapter how such models are estimated and interpreted.

**17** In Chapter 18 we discuss the topic of **duration or survival analysis**. Duration of a marriage, duration of a strike, duration of an illness, and duration of unemployment are some examples of duration data.

**18** In Chapter 19, we discuss a topic that has received considerable attention in the literature, namely, the method of **Instrumental Variables** (**IV**). The bulk of this book has been devoted to the case of nonstochastic or fixed regressors, but there are situations where we have to consider stochastic, or random, regressors. If the stochastic regressors are correlated with the error term, the OLS estimators are not only biased but are also inconsistent – that is, the bias does not diminish no matter how large the sample. The basic principle of IV is that it replaces the stochastic regressors with another set of regressors, called **instrumental variables** (or simply **instruments**), that are correlated with the stochastic regressors but are uncorrelated with the error term. As a result, we can obtain consistent estimates of the regression parameters. In this chapter we show how this can be accomplished.

**19** Chapter 20, on quantile regression models (QRM), is new to this edition. QRMs examine a more complete description of the conditional distribution of a variable than just the conditional mean, as in the traditional linear regression models. In some situations, it may be better to estimate a median regression model, which is less sensitive to outlier observations.

We can go further and divide our data into several groups, such as quartiles (division into four groups), deciles (division into ten groups), or centiles (division into 100 groups). By examining the various quantiles, we may be able to find if the mean value of the regressand differs from quantile to quantile. If such is the case, running just one regression for all the observations in the sample may distort the true nature of the relationship between the regressand and the regressors.

**20** Chapter 21, also new to this edition, discusses the topic of **Multivariate Regression Models** (MRMs), which have found applications in several disciplines.In MRM, we have several dependent variables, or regressands, but theyall have the same regressors. For example, the Scholastic Aptitude Test (SAT)in the USA measures the performance of a student on verbal and quantitativequestions in relation to several socio-economic variables. In principle, we canestimate separate OLS regressions for the verbal and quantitative parts of theSAT. But this neglects the possibility that the verbal and quantitative scores ofstudents may be correlated. The MRM takes into account such a correlation andestimates the two scores jointly.

The MRMS are a special case of **Seemingly Unrelated Regressions**. For example, the capital expenditure decisions of General Motors (GM), Ford Motor, and Chrysler, while made independently, may not in fact be independent because all these companies operate in the same regulatory and capital market conditions. Therefore, it may be necessary to take these factors into account in developing the expenditure (regression) functions of the three companies. In this chapter we show how the MRMs handle such problems.