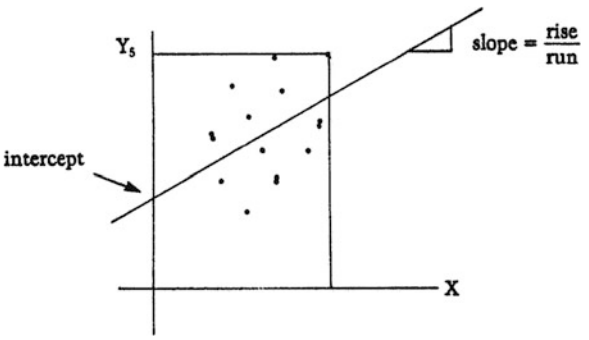
Chapter 2
Regression Analysis and Forecasting Models

A forecast is merely a prediction about the future values of data. However, most extrapolative model forecasts assume that the past is a proxy for the future. That is, the economic data for the 2012–2020 period will be driven by the same variables as was the case for the 2000–2011 period, or the 2007–2011 period. There are many traditional models for forecasting: exponential smoothing, regression, time series, and composite model forecasts, often involving expert forecasts. Regression analysis is a statistical technique to analyze quantitative data to estimate model parameters and make forecasts. We introduce the reader to regression analysis in this chapter.

The horizontal line is called the *X*-axis and the vertical line the *Y*-axis. Regression analysis looks for a relationship between the *X* variable (sometimes called the "independent" or "explanatory" variable) and the *Y* variable (the "dependent" variable).



For example, *X* might be the aggregate level of personal disposable income in the United States and *Y* would represent personal consumption expenditures in the United States, an example used in Guerard and Schwartz (2007). By looking up these numbers for a number of years in the past, we can plot points on the graph. More specifically, regression analysis seeks to find the "line of best fit" through the points. Basically, the regression line is drawn to best approximate the relationship

J.B. Guerard, Jr., *Introduction to Financial Forecasting in Investment Analysis*,
DOI 10.1007/978-1-4614-5239-3_2, © Springer Science+Business Media New York 2013

between the two variables. Techniques for estimating the regression line (i.e., its intercept on the Y-axis and its slope) are the subject of this chapter. Forecasts using the regression line assume that the relationship which existed in the past between the two variables will continue to exist in the future. There may be times when this assumption is inappropriate, such as the "Great Recession" of 2008 when the housing market bubble burst. The forecaster must be aware of this potential pitfall. Once the regression line has been estimated, the forecaster must provide an estimate of the future level of the independent variable. The reader clearly sees that the forecast of the independent variable is paramount to an accurate forecast of the dependent variable.

Regression analysis can be expanded to include more than one independent variable. Regressions involving more than one independent variable are referred to as multiple regression. For example, the forecaster might believe that the number of cars sold depends not only on personal disposable income but also on the level of interest rates. Historical data on these three variables must be obtained and a plane of best fit estimated. Given an estimate of the future level of personal disposable income and interest rates, one can make a forecast of car sales.

Regression capabilities are found in a wide variety of software packages and hence are available to anyone with a microcomputer. Microsoft Excel, a popular spreadsheet package, SAS, SCA, RATS, and EViews can do simple or multiple regressions. Many statistics packages can do not only regressions but also other quantitative techniques such as those discussed in Chapter 3 (Time Series Analysis and Forecasting). In simple regression analysis, one seeks to measure the statistical association between two variables, X and Y. Regression analysis is generally used to measure how changes in the independent variable, X, influence changes in the dependent variable, Y. Regression analysis shows a statistical association or correlation among variables, rather than a causal relationship among variables.

The case of simple, linear, least squares regression may be written in the form

$$Y = \alpha + \beta X + \varepsilon, \tag{2.1}$$

where Y, the dependent variable, is a linear function of X, the independent variable. The parameters α and β characterize the population regression line and ε is the randomly distributed error term. The regression estimates of α and β will be derived from the principle of least squares. In applying least squares, the sum of the squared regression errors will be minimized; our regression errors equal the actual dependent variable minus the estimated value from the regression line. If Y represents the actual value and Y the estimated value, then their difference is the error term, e. Least squares regression minimized the sum of the squared error terms. The simple regression line will yield an estimated value of Y, $\hat{Y}$, by the use of the sample regression:

$$\hat{Y} = a + \beta X. \tag{2.2}$$

In the estimation (2.2), a is the least squares estimate of α and b is the estimate of β. Thus, a and b are the regression constants that must be estimated. The least

squares regression constants (or statistics) α and β are unbiased and efficient (smallest variance) estimators of α and β. The error term, e_i, is the difference between the actual and estimated dependent variable value for any given independent variable values, X_i.

$$e_i = \hat{Y}_i - Y_i. \tag{2.3}$$

The regression error term, e_i, is the least squares estimate of ε_i, the actual error term.[1]

To minimize the error terms, the least squares technique minimizes the sum of the squares error terms of the N observations,

$$\sum_{i=1}^{N} e_i^2. \tag{2.4}$$

The error terms from the N observations will be minimized. Thus, least squares regression minimizes:

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left[Y_i - \hat{Y}_i\right]^2 = \sum_{i=1}^{N} \left[Y_i - (\alpha + b X_i)\right]^2. \tag{2.5}$$

To assure that a minimum is reached, the partial derivatives of the squared error terms function

$$\sum_{i=1}^{N} = \left[Y_i - (\alpha + b X_i)\right]^2$$

will be taken with respect to a and b.

$$\begin{aligned} \frac{\partial \sum_{i=1}^{N} e_i^2}{\partial a} &= 2 \sum_{i=1}^{N} (Y_i - a - bX_i)(-1) \\ &= -2\left(\sum_{i=1}^{N} Y_i - \sum_{i=1}^{N} a - b \sum_{i=1}^{N} X_i\right) \end{aligned}$$

[1] The reader is referred to an excellent statistical reference, S. Makridakis, S.C. Wheelwright, and R. J. Hyndman, *Forecasting: Methods and Applications*, Third Edition (New York; Wiley, 1998), Chapter 5.

$$\frac{\partial \sum_{i=1}^{N} e_i^2}{\partial b} = 2\sum_{i=1}^{N}(Y_i - a - bX_i)(-Xi)$$
$$= -2\left(\sum_{i=1}^{N} Y_i X_i - \sum_{i=1}^{N} X_i - b\sum_{i=1}^{N} X_1^2\right).$$

The partial derivatives will then be set equal to zero.

$$\begin{aligned}\frac{\partial \sum_{i=1}^{N} e_i^2}{\partial a} &= -2\left(\sum_{i=1}^{N} Y_i - \sum_{i=1}^{N} a - b\sum_{i=1}^{N} X_i\right) = 0\\ \frac{\partial \sum_{i=1}^{N} e_i^2}{\partial b} &= -2\left(\sum_{i=1}^{N} YX_i - \sum_{i=1}^{N} X_l - b\sum_{i=1}^{N} X_1^2\right) = 0.\end{aligned} \tag{2.6}$$

Rewriting these equations, one obtains the normal equations:

$$\begin{aligned}\sum_{i=1}^{N} Y_i &= \sum_{i=1}^{N} a + b\sum_{i=1}^{N} X_i\\ \sum_{i=1}^{N} Y_i X_i &= a\sum_{i=1}^{N} X_i + b\sum_{i=1}^{N} X_1^2.\end{aligned} \tag{2.7}$$

Solving the normal equations simultaneously for a and b yields the least squares regression estimates:

$$\begin{aligned}\hat{a} &= \frac{\left(\sum_{i=1}^{N} X_i^2\right)\left(\sum_{i=1}^{N} Y_i\right) - \left(\sum_{i=1}^{N} X_i Y_i\right)}{N\left(\sum_{i=1}^{N} X_i^2\right) - \left(\sum_{i=1}^{N} X_i\right)^2},\\ \hat{b} &= \frac{\left(\sum_{i=1}^{N} X_i Y_i\right) - \left(\sum_{i=1}^{N} X_i\right)\left(\sum_{i=1}^{N} Y_i\right)}{N\left(\sum_{i=1}^{N} X_i^2\right) - \left(\sum_{i=1}^{N} X_i\right)^2}.\end{aligned} \tag{2.8}$$

An estimation of the regression line's coefficients and goodness of fit also can be found in terms of expressing the dependent and independent variables in terms of deviations from their means, their sample moments. The sample moments will be denoted by M.

$$M_{XX} = \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} (x_i - \bar{x})^2$$
$$= N \sum_{i=1}^{N} X_i - \left(\sum_{i=1}^{N} X_i \right)^2$$

$$M_{XY} = \sum_{i=1}^{N} x_i y_i = \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})$$
$$= N \sum_{i=1}^{N} X_i Y_i - \left(\sum_{i=1}^{N} X_i \right) \left(\sum_{i=1}^{N} Y_i \right)$$

$$M_{YY} = \sum_{i=1}^{N} y_i^2 = \sum_{i=1}^{N} (Y - \bar{Y})^2$$
$$= N \left(\sum_{i=1}^{N} Y_i^2 \right) - \sum_{i=1}^{N} (Y_i)^2.$$

The slope of the regression line, b, can be found by

$$b = \frac{M_{XY}}{M_{XX}} \tag{2.9}$$

$$a = \frac{\sum_{i=1}^{N} Y_i}{N} - b \frac{\sum_{i=1}^{N} X_i}{N} = \bar{y} - b\bar{X}. \tag{2.10}$$

The standard error of the regression line can be found in terms of the sample moments.

$$S_e^2 = \frac{M_{XX}(M_{YY}) - (M_{XY})^2}{N(N-2)M_{XX}}$$
$$S_e = \sqrt{S_e^2}. \tag{2.11}$$

The major benefit in calculating the sample moments is that the correlation coefficient, r, and the coefficient of determination, r^2, can easily be found.

$$r = \frac{M_{XY}}{(M_{XX})(M_{YY})}$$
$$R^2 = (r)^2. \tag{2.12}$$

The coefficient of determination, R^2, is the percentage of the variance of the dependent variable explained by the independent variable. The coefficient of determination cannot exceed 1 nor be less than zero. In the case of $R^2 = 0$, the regression line's $Y = Y$ and no variation in the dependent variable are explained. If the dependent variable pattern continues as in the past, the model with time as the independent variable should be of good use in forecasting.

The firm can test whether the a and b coefficients are statistically different from zero, the generally accepted null hypothesis. A t-test is used to test the two null hypotheses:

H_{0_1}: $a = 0$
H_{A_1}: a ne 0
H_{0_2}: $\beta = 0$
H_{A_2}: β ne 0,
where ne denotes not equal.

The H_0 represents the null hypothesis while H_A represents the alternative hypothesis. To reject the null hypothesis, the calculated t-value must exceed the critical t-value given in the t-tables in the appendix. The calculated t-values for a and b are found by

$$t_a = \frac{a - \alpha}{S_e}\sqrt{\frac{N(M_{XX})}{M_{XX} + (N\bar{X})^2}}$$
$$t_b = \frac{b - \beta}{S_e}\sqrt{\frac{(M_{XX})}{N}}. \tag{2.13}$$

The critical t-value, t_c, for the 0.05 level of significance with $N - 2$ degrees of freedom can be found in a t-table in any statistical econometric text. One has a statistically significant regression model if one can reject the null hypothesis of the estimated slope coefficient.

We can create 95% confidence intervals for a and b, where the limits of a and b are

$$a + ta/_2 S_e{}^{+}\sqrt{\frac{(N\bar{X})^2 + M_{XX}}{N(M_{XX})}}$$
$$b + ta/_2 S_e\sqrt{\frac{N}{M_{XX}}}. \tag{2.14}$$

To test whether the model is a useful model, an F-test is performed where

$H_0 = \alpha = \beta = 0$
$H_A = \alpha$ ne β ne 0

$$F = \frac{\sum_{i=1}^{N} Y^2 \div 1 - \beta^2 \sum_{i=1}^{N} X_i^2}{\sum_{i=1}^{N} e^2 \div N - 2}. \tag{2.15}$$

As the calculated F-value exceeds the critical F-value with $(1, \mathrm{N} - 2)$ degrees of freedom of 5.99 at the 0.05 level of significance, the null hypothesis must be rejected. The 95% confidence level limit of prediction can be found in terms of the dependent variable value:

$$(a + bX_0) + ta/_2 S_e \sqrt{\frac{N(X_0 - \bar{X})^2}{1 + N + M_{XX}}}. \tag{2.16}$$

Examples of Financial Economic Data

The most important use of simple linear regression as developed in (2.9) and (2.10) is the estimation of a security beta. A security beta is estimated by running a regression of 60 months of security returns as a function of market returns. The market returns are generally the Standard & Poor's 500 (S&P500) index or a capitalization-weighted index, such as the value-weighted Index from the Center for Research in Security Prices (CRSP) at the University of Chicago. The data for beta estimations can be downloaded from the Wharton Research Data Services (WRDS) database. The beta estimation for IBM from January 2005 to December 2009, using monthly S&P 500 and the value-weighted CRSP Index, produces a beta of approximately 0.80. Thus, if the market is expected to increase 10% in the coming year, then one would expect IBM to return about 8%. The beta estimation of IBM as a function of the S&P 500 Index using the SAS system is shown in Table 2.1. The IBM beta is 0.80. The t-statistic of the beta coefficient, the slope of the regression line, is 5.53, which is highly statistically significant. The critical 5% t-value is with 30 degrees of freedom 1.96, whereas the critical level of the t-statistic at the 10% is 1.645. The IBM beta is statistically different from zero. The IBM beta is not statistically different from one; the normalized z-statistical is significantly less than 1. That is, 0.80 − 1.00 divided by the regression coefficient standard error of 0.144 produces a Z-statistic of −1.39, which is less than the critical level of −1.645 (at the 10% level) or −1.96 (at the 5% critical level). The IBM beta is 0.78 (the corresponding t-statistic is 5.87) when calculated versus the value-weighted CRSP Index.[2]

[2] See Fama, *Foundations of Finance*, 1976, Chapter 3, p. 101–2, for an IBM beta estimation with an equally weighted CRSP Index.

Table 2.1 WRDS IBM Beta 1/2005–12/2009

Dependent variable: ret					
Number of observations read: 60					
Number of observations used: 60					
Analysis of variance					
Source	DF	Sum of squares	Mean square	F-value	Pr > F
Model	1	0.08135	0.08135	30.60	<0.0001
Error	58	0.15419	0.00266		
Corrected total	59	0.23554			
Root MSE	0.05156	R^2	0.3454		
Dependent mean	0.00808	Adjusted R^2	0.3341		
Coeff var	638.12982				
Parameter estimates					
Variable	DF	Parameter estimate	Standard error	t-Value	Pr > $\|t\|$
Intercept	1	0.00817	0.00666	1.23	0.2244
Sprtrn	1	0.80063	0.14474	5.53	<0.0001

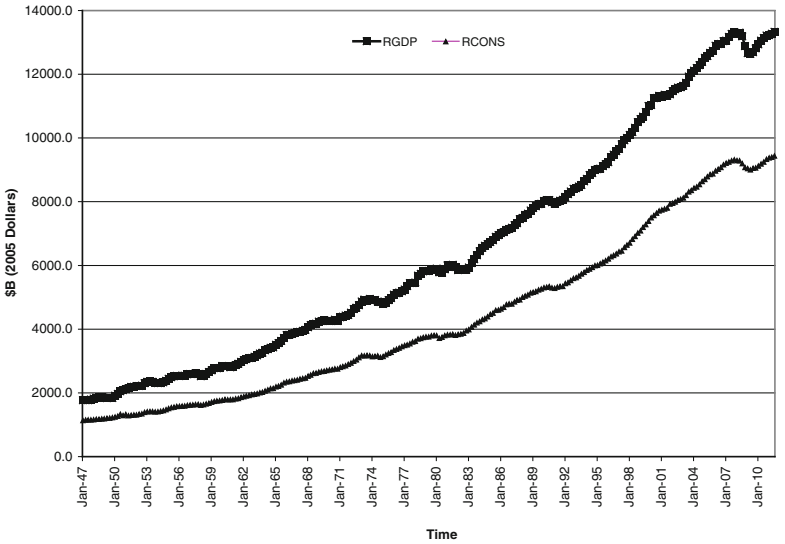
Table 2.2 An Estimated Consumption Function, 1947–2011

Dependent variable: RPCE				
Method: least squares				
Sample(adjusted): 1,259				
Included observations: 259 after adjusting endpoints				
Variable	Coefficient	Std. error	t-Statistic	Prob.
C	−120.0314	12.60258	−9.524349	0.0000
RPDI	0.933251	0.002290	407.5311	0.0000
R^2	0.998455	Mean dependent var		4,319.917
Adjusted R^2	0.998449	S.D. dependent var		2,588.624
S.E. of regression	101.9488	Akaike info criterion		12.09451
Sum squared resid	2,671,147	Schwarz criterion		12.12198
Log likelihood	−1,564.239	F-statistic		166,081.6
Durbin–Watson stat	0.197459	Prob(F-statistic)		0.000000

Let us examine another source of real-business economic and financial data. The St. Louis Federal Reserve Bank has an economic database, denoted FRED, containing some 41,000 economic series, available at no cost, via the Internet, at http://research.stlouisfed.org/fred2. Readers are well aware that consumption makes up the majority of real Gross Domestic Product, denoted GDP, the accepted measure of output in our economy. Consumption is the largest expenditure, relative to gross investment, government spending, and net exports in GDP data. If we download and graph real GDP and real consumption expenditures from FRED from 1947 to 2011, shown in Chart 2, one finds that real GDP and real consumption expenditures, in 2005 $, have risen substantially in the postwar period. Moreover, there is a highly statistical significant relationship between real GDP and consumption if one estimates an ordinary least squares (OLS) line of the form of (2.8) with real GDP as the dependent variable and real consumption as the independent variable. The reader is referred to Table 2.2.

Table 2.3 An estimated consumption function, with lagged income

Dependent variable: RPCE				
Method: least squares				
Sample(adjusted): 2,259				
Included observations: 258 after adjusting endpoints				
Variable	Coefficient	Std. error	*t*-Statistic	Prob.
C	−118.5360	12.73995	−9.304274	0.0000
RPDI	0.724752	0.126290	5.738800	0.0000
LRPDI	0.209610	0.126816	1.652864	0.0996
R^2	0.998470	Mean dependent var		4,332.278
Adjusted R^2	0.998458	S.D. dependent var		2,585.986
S.E. of regression	101.5529	Akaike info criterion		12.09060
Sum squared resid	2,629,810	Schwarz criterion		12.13191
Log likelihood	−1,556.687	*F*-statistic		83,196.72
Durbin–Watson stat	0.127677	Prob(*F*-statistic)		0.000000



Source: US Department of Commerce, Bureau of Economic Analysis, Series GDPC1 and PCECC96, 1947–2011, seasonally-adjusted, Chained 2005 Dollars

The slope of consumption function is 0.93, and is highly statistically significant.[3]

The introduction of current and lagged income variables in the consumption function regression produces statistically significant coefficients on both current and lagged income, although the lagged income variable is statistically significant at the 10% level. The estimated regression line, shown in Table 2.3, is highly statistically significant.

[3] In recent years the marginal propensity to consume has risen to the 0.90 to 0.97 range, see Joseph Stiglitz, *Economics*, 1993, p.745.

Table 2.4 An estimated consumption function, with twice-lagged consumption

Dependent variable: RPCE				
Method: least squares				
Included observations: 257 after adjusting endpoints				
Variable	Coefficient	Std. error	*t*-Statistic	Prob.
C	−120.9900	12.92168	−9.363331	0.0000
RPDI	0.736301	0.126477	5.821607	0.0000
LRPDI	0.229046	0.177743	1.288633	0.1987
L2RPDI	−0.030903	0.127930	−0.241557	0.8093
R^2	0.998474	Mean dependent var		4,344.661
Adjusted R^2	0.998456	S.D. dependent var		2,583.356
S.E. of regression	101.5049	Akaike info criterion		12.09353
Sum squared resid	2,606,723	Schwarz criterion		12.14877
Log likelihood	−1,550.019	*F*-statistic		55,188.63
Durbin–Watson stat	0.130988	Prob(*F*-statistic)		0.000000

The introduction of current and once- and twice-lagged income variables in the consumption function regression produces statistically significant coefficients on both current and lagged income, although the lagged income variable is statistically significant at the 20% level. The twice-lagged income variable is not statistically significant. The estimated regression line, shown in Table 2.4, is highly statistically significant.

Autocorrelation

An estimated regression equation is plagued by the first-order correlation of residuals. That is, the regression error terms are not white noise (random) as is assumed in the general linear model, but are serially correlated where

$$\varepsilon_t = \rho\varepsilon_{t=1} + U_t, \quad t = 1, 2, \ldots, N \tag{2.17}$$

ε_t = regression error term at time t, ρ = first-order correlation coefficient, and U_t = normally and independently distributed random variable.

The serial correlation of error terms, known as autocorrelation, is a violation of a regression assumption and may be corrected by the application of the Cochrane–Orcutt (CORC) procedure.[4] Autocorrelation produces unbiased, the expected value of parameter is the population parameter, but inefficient parameters. The variances of the parameters are biased (too low) among the set of linear unbiased estimators and the sample *t*- and *F*-statistics are too large. The CORC

[4] D. Cochrane and G.H. Orcutt, "Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms," *Journal of the American Statistical Association*, 1949, 44: 32–61.

procedure was developed to produce the best linear unbiased estimators (BLUE) given the autocorrelation of regression residuals. The CORC procedure uses the information implicit in the first-order correlative of residuals to produce unbiased and efficient estimators:

$$Y_t = \alpha + \beta X_t + \varepsilon_t$$
$$\hat{\rho} = \frac{\sum e_t, e_t - 1}{\sum e_t^2 - 1}.$$

The dependent and independent variables are transformed by the estimated rho, $\hat{\rho}$, to obtain more efficient OLS estimates:

$$Y_t - \rho Y_{t-1} = \alpha(l - \rho) + \beta(X_t - \rho X_{t-1}) + ut. \tag{2.19}$$

The CORC procedure is an iterative procedure that can be repeated until the coefficients converge. One immediately recognizes that as ρ approaches unity the regression model approaches a first-difference model.

The Durbin–Watson, *D–W*, statistic was developed to test for the absence of autocorrelation:

H_0: $\rho = 0$.

One generally tests for the presence of autocorrelation ($\rho = 0$) using the Durbin–Watson statistic:

$$D - W = d = \frac{\sum_{t=2}^{N} (e_t = e_{t-1})^2}{\sum_{t=2}^{N} e_t^2}. \tag{2.20}$$

The *e*s represent the OLS regression residuals and a two-tailed tail is employed to examine the randomness of residuals. One rejects the null hypothesis of no statistically significant autocorrelation if

$$d < d_L \text{ or } d > 4 - d_U,$$

where d_L is the "lower" Durbin–Watson level and d_U is the "upper" Durbin–Watson level.

The upper and lower level Durbin–Watson statistic levels are given in Johnston (1972). The Durbin–Watson statistic is used to test only for first-order correlation among residuals.

$$D = 2(1 - \rho). \tag{2.21}$$

If the first-order correlation of model residuals is zero, the Durbin–Watson statistic is 2. A very low value of the Durbin–Watson statistic, $d < d_L$, indicates

Table 2.5 An estimated consumption function, 1947–2011

Dependent variable: D(RPCE)				
Method: least squares				
Included observations: 258 after adjusting endpoints				
Variable	Coefficient	Std. error	*t*-Statistic	Prob.
C	22.50864	2.290291	9.827849	0.0000
D(RPDI)	0.280269	0.037064	7.561802	0.0000
R^2	0.182581	Mean dependent var		32.18062
Adjusted R^2	0.179388	S.D. dependent var		33.68691
S.E. of regression	30.51618	Akaike info criterion		9.682113
Sum squared resid	238,396.7	Schwarz criterion		9.709655
Log likelihood	−1,246.993	*F*-statistic		57.18084
Durbin-Watson stat	1.544444	Prob(*F*-statistic)		0.000000

positive autocorrelation between residuals and produces a regression model that is not statistically plagued by autocorrelation.

The inconclusive range for the estimated Durbin–Watson statistic is

$$d_{\mathrm{L}}<d<d_{\mathrm{U}} \text{ or } 4-d_{\mathrm{U}}<4-d_{\mathrm{U}}.$$

One does not reject the null hypothesis of no autocorrelation of residuals if $d_{\mathrm{U}} < d < 4 - d_{\mathrm{U}}$.

One of the weaknesses of the Durbin–Watson test for serial correlation is that only first-order autocorrelation of residuals is examined; one should plot the correlation of residual with various time lags

$$\text{corr } (e_t, e_{t-k})$$

to identify higher-order correlations among residuals.

The reader may immediately remember that the regressions shown in Tables 2.1–2.3 had very low Durbin–Watson statistics and were plagued by autocorrelation. We first-difference the consumption function variables and rerun the regressions, producing Tables 2.5–2.7. The R^2 values are lower, but the regressions are not plagued by autocorrelation. In financial economic modeling, one generally first-differences the data to achieve stationarity, or a series with a constant standard deviation.

The introduction of current and lagged income variables in the consumption function regression produces statistically significant coefficients on both current and lagged income, although the lagged income variable is statistically significant at the 10% level. The estimated regression line, shown in Table 2.6, is highly statistically significant, and is not plagued by autocorrelation.

The introduction of current and lagged income variables in the consumption function regression produces statistically significant coefficients on both current and lagged income, statistically significant at the 1% level. The estimated regression line, shown in Table 2.5, is highly statistically significant, and is not plagued by autocorrelation.

Table 2.6 An estimated consumption function, with lagged income

Dependent variable: D(RPCE)				
Method: least squares				
Included observations: 257 after adjusting endpoints				
Variable	Coefficient	Std. error	*t*-Statistic	Prob.
C	14.20155	2.399895	5.917570	0.0000
D(RPDI)	0.273239	0.034027	8.030014	0.0000
D(LRPDI)	0.245108	0.034108	7.186307	0.0000
R^2	0.320314	Mean dependent var		32.23268
Adjusted R^2	0.314962	S.D. dependent var		33.74224
S.E. of regression	27.92744	Akaike info criterion		9.508701
Sum squared resid	198,105.2	Schwarz criterion		9.550130
Log likelihood	−1,218.868	*F*-statistic		59.85104
Durbin-Watson stat	1.527716	Prob(*F*-statistic)		0.000000

Table 2.7 An estimated consumption function, with twice-lagged consumption

Dependent variable: D(RPCE)				
Method: least squares				
Included observations: 256 after adjusting endpoints				
Variable	Coefficient	Std. error	*t*-Statistic	Prob.
C	12.78746	2.589765	4.937692	0.0000
D(RPDI)	0.262664	0.034644	7.581744	0.0000
D(LRPDI)	0.242900	0.034162	7.110134	0.0000
D(L2RPDI)	0.054552	0.034781	1.568428	0.1180
R^2	0.325587	Mean dependent var		32.34414
Adjusted R^2	0.317558	S.D. dependent var		33.76090
S.E. of regression	27.88990	Akaike info criterion		9.509908
Sum squared resid	196,017.3	Schwarz criterion		9.565301
Log likelihood	−1,213.268	*F*-statistic		40.55269
Durbin–Watson stat	1.535845	Prob(*F*-statistic)		0.000000

The introduction of current and once- and twice-lagged income variables in the consumption function regression produces statistically significant coefficients on both current and lagged income, although the twice-lagged income variable is statistically significant at the 15% level. The estimated regression line, shown in Table 2.7, is highly statistically significant, and is not plagued by autocorrelation.

Many economic time series variables increase as a function of time. In such cases, a nonlinear least squares (NLLS) model may be appropriate; one seeks to estimate an equation in which the dependent variable increases by a constant growth rate rather than a constant amount.[5] The nonlinear regression equation is

[5] The reader is referred to C.T. Clark and L.L. Schkade, Statistical Analysis for Administrative Decisions (Cincinnati: South-Western Publishing Company, 1979) and Makridakis, Wheelwright, and Hyndman, *Op. Cit.*, 1998, pages 221–225, for excellent treatments of this topic.

$$\begin{aligned} Y &= ab^x \\ \text{or } \log\ Y &= \log a + \log BX. \end{aligned} \tag{2.22}$$

The normal equations are derived from minimizing the sum of the squared error terms (as in OLS) and may be written as

$$\begin{aligned} \sum (\log\ Y) &= N(\log\ a) + (\log\ b) \sum X \\ \sum (X \log\ Y) &= (\log\ a) \sum X + (\log\ b) \sum X^2. \end{aligned} \tag{2.23}$$

The solutions to the simplified NLLS estimation equation are

$$\log\ a = \frac{\sum (\log Y)}{N} \tag{2.24}$$

$$\log\ b = \frac{\sum (X \log Y)}{\sum X^2}. \tag{2.25}$$

Multiple Regression

It may well be that several economic variables influence the variable that one is interested in forecasting. For example, the levels of the Gross National Product (GNP), personal disposable income, or price indices can assert influences on the firm. Multiple regression is an extremely easy statistical tool for researchers and management to employ due to the great proliferation of computer software. The general form of the two-independent variable multiple regression is

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \varepsilon_t, \quad t = 1, \ldots, N. \tag{2.26}$$

In matrix notation multiple regression can be written:

$$Y = X\beta + \varepsilon. \tag{2.27}$$

Multiple regression requires unbiasedness, the expected value of the error term is zero, and the X's are fixed and independent of the error term. The error term is an identically and independently distributed normal variable. Least squares estimation of the coefficients yields

$$\begin{aligned} \hat{\beta} &= (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) \\ Y &= X\hat{\beta} + e. \end{aligned} \tag{2.28}$$

Multiple regression, using the least squared principle, minimizes the sum of the squared error terms:

$$\sum_{i=1}^{N} e_1^2 = e'e$$
$$(Y - X\hat{\beta})'(Y - X\hat{\beta}). \tag{2.29}$$

To minimize the sum of the squared error terms, one takes the partial derivative of the squared errors with respect to $\hat{\beta}$ and the partial derivative set equal to zero.

$$\partial \frac{(e'e)}{\partial \beta} = -2X'Y + 2X'X\hat{\beta} = 0 \tag{2.30}$$

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

Alternatively, one could solve the normal equations for the two-variable to determine the regression coefficients.

$$\begin{aligned} \sum Y &= \beta_1 N + \hat{\beta}_2 \sum X_2 + \hat{\beta}_3 \sum X_3 \\ \sum X_2 Y &= \hat{\beta}_1 \sum X_2 + \hat{\beta}_2 {X_2}^2 + \hat{\beta}_3 \sum X_3^2 \\ \sum X_3 Y &= \hat{\beta}_1 \sum X_3 + \hat{\beta}_2 \sum X_2 X_3 + \hat{\beta}_3 \sum X_3^2. \end{aligned} \tag{2.31}$$

When we solved the normal equation, (2.7), to find the a and b that minimized the sum of our squared error terms in simple liner regression, and when we solved the two-variable normal equation, equation (2.31), to find the multiple regression estimated parameters, we made several assumptions. First, we assumed that the error term is independently and identically distributed, i.e., a random variable with an expected value, or mean of zero, and a finite, and constant, standard deviation. The error term should not be a function of time, as we discussed with the Durbin–Watson statistic, equation (2.21), nor should the error term be a function of the size of the independent variable(s), a condition known as heteroscedasticity. One may plot the residuals as a function of the independent variable(s) to be certain that the residuals are independent of the independent variables. The error term should be a normally distributed variable. That is, the error terms should have an expected value of zero and 67.6% of the observed error terms should fall within the mean value plus and minus one standard deviation of the error terms (the so-called Bell Curve or normal distribution). Ninety-five percent of the observations should fall within the plus or minus two standard deviation levels, the so-called 95% confidence interval. The presence of extreme, or influential, observations may distort estimated regression lines and the corresponding estimated residuals. Another problem in regression analysis is the assumed independence of the

independent variables in equation (2.31). Significant correlations may produce estimated regression coefficients that are "unstable" and have the "incorrect" signs, conditions that we will observe in later chapters. Let us spend some time discussing two problems discussed in this section, the problems of influential observations, commonly known as outliers, and the correlation among independent variables, known as multicollinearity.

There are several methods that one can use to identify influential observations or outliers. First, we can plot the residuals and 95% confidence intervals and examine how many observations have residuals falling outside these limits. One should expect no more than 5% of the observations to fall outside of these intervals. One may find that one or two observations may distort a regression estimate even if there are 100 observations in the database. The estimated residuals should be normally distributed, and the ratio of the residuals divided by their standard deviation, known as standardized residuals, should be a normal variable. We showed, in equation (2.31), that in multiple regression

$$\hat{\beta} = (X'X)X'Y.$$

The residuals of the multiple regression line are given by

$$e = Y' - \hat{\beta}X.$$

The standardized residual concept can be modified such that the reader can calculate a variation on that term to identify influential observations. If we delete observation i in a regression, we can measure the change in estimated regression coefficients and residuals. Belsley et al. (1980) showed that the estimated regression coefficients change by an amount, DFBETA, where

$$\text{DFBETA}_i = \frac{(X'X)^{-1}X'e_i}{1 - h_i}, \tag{2.32}$$

where $h_i = X_i(X'X)^{-1}X_i'$.

The h_i or "hat" term is calculated by deleting observation i. The corresponding residual is known as the studentized residual, sr, and defined as

$$\text{sr}_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_i}}, \tag{2.33}$$

where $\hat{\sigma}$ is the estimated standard deviation of the residuals. A studentized residual that exceeds 2.0 indicates a potential influential observation (Belsley et al. 1980). Another distance measure has been suggested by Cook (1977), which modifies the studentized residual, to calculate a scaled residual known as the Cook distance measure, CookD. As the researcher or modeler deletes observations, one needs to

compare the original matrix of the estimated residual's variance matrix. The COVRATIO calculation performs this calculation, where

$$\text{COVRATIO} = \frac{1}{\left[\frac{n-p-1}{n-p} + \frac{e_i^*}{(n-p)}\right]^p (1-h_i)}, \tag{2.34}$$

where n = number of observations, p = number of independent variables, and e_i^* = deleted observations.

If the absolute value of the deleted observation >2, then the COVRATIO calculation approaches

$$1 - \frac{3p}{n}. \tag{2.35}$$

A calculated COVRATIO that is larger than $3p/n$ indicates an influential observation. The DFBETA, studentized residual, CookD, and COVRATIO calculations may be performed within SAS. The identification of influential data is an important component of regression analysis. One may create variables for use in multiple regression that make use of the influential data, or outliers, to which they are commonly referred.

The modeler can identify outliers, or influential data, and rerun the OLS regressions on the re-weighted data, a process referred to as robust (ROB) regression. In OLS all data is equally weighted. The weights are 1.0. In ROB regression one weights the data universally with its OLS residual; i.e., the larger the residual, the smaller the weight of the observation in the ROB regression. In ROB regression, several weights may be used. We will see the Huber (1973) and Beaton-Tukey (1974) weighting schemes in our analysis. In the Huber robust regression procedure, one uses the following calculation to weigh the data:

$$w_i = \left(1 - \left(\frac{|e_i|}{\sigma_i}\right)^2\right)^2, \tag{2.36}$$

where e_i = residual i, σ_i = standard deviation of residual, and w_i = weight of observation i.

The intuition is that the larger the estimated residual, the smaller the weight. A second robust re-weighting scheme is calculated from the Beaton-Tukey biweight criteria where

$$w_i = \begin{cases} \left(1 - \left(\dfrac{\frac{|e_i|}{\sigma_e}}{4.685}\right)^2\right)^2, & \text{if } \dfrac{|e_i|}{\sigma_e} > 4.685; \\ 1, & \text{if } \dfrac{|e_i|}{\sigma_e} < 4.685. \end{cases} \tag{2.37}$$

A second major problem is one of multicollinearity, the condition of correlations among the independent variables. If the independent variables are perfectly correlated in multiple regression, then the $(X'X)$ matrix of (2.31) cannot be inverted and the multiple regression coefficients have multiple solutions. In reality, highly correlated independent variables can produce unstable regression coefficients due to an unstable $(X'X)^{-1}$ matrix. Belsley et al. advocate the calculation of a condition number, which is the ratio of the largest latent root of the correlation matrix relative to the smallest latent root of the correlation matrix. A condition number exceeding 30.0 indicates severe multicollinearity.

The latent roots of the correlation matrix of independent variables can be used to estimate regression parameters in the presence of multicollinearity. The latent roots, $l_1, l_2, \ldots, l_p$ and the latent vectors $\gamma_1, \gamma_2, \ldots, \gamma_p$ of the P independent variables can describe the inverse of the independent variable matrix of (2.29).

$$(X'X)^{-1} = \sum_{j=1}^{p} l_j^{-1} \gamma_j \gamma_j'.$$

Multicollinearity is present when one observes one or more small latent vectors. If one eliminates latent vectors with small latent roots ($l < 0.30$) and latent vectors ($\gamma < 0.10$), the "principal component" or latent root regression estimator may be written as

$$\hat{\beta}_{\text{LRR}} = \sum_{j=0}^{P} f_j \delta_j,$$

where $f_j = \frac{-\eta \gamma_0 \lambda_j^{-1}}{\sum_q \gamma_0^2 \lambda_q^{-1}}$,

where $n^2 = \Sigma (y - \bar{y})^2$

and λ are the "nonzero" latent vectors. One eliminates the latent vectors with non-predictive multicollinearity. We use latent root regression on the Beaton-Tukey weighted data in Chapter 4.

The Conference Board Composite Index of Leading Economic Indicators and Real US GDP Growth: A Regression Example

The composite indexes of leading (leading economic indicators, LEI), coincident, and lagging indicators produced by The Conference Board are summary statistics for the US economy. Wesley Clair Mitchell of Columbia University constructed the indicators in 1913 to serve as a barometer of economic activity. The leading indicator series was developed to turn upward before aggregate economic activity increased, and decrease before aggregate economic activity diminished.

Historically, the cyclical turning points in the leading index have occurred before those in aggregate economic activity, cyclical turning points in the coincident index have occurred at about the same time as those in aggregate economic activity, and cyclical turning points in the lagging index generally have occurred after those in aggregate economic activity.

The Conference Board's components of the composite leading index for the year 2002 reflects the work and variables shown in Zarnowitz (1992) list, which continued work of the Mitchell (1913 and 1951), Burns and Mitchell (1946), and Moore (1961). The Conference Board index of leading indicators is composed of

1. Average weekly hours (mfg.)
2. Average weekly initial claims for unemployment insurance
3. Manufacturers' new orders for consumer goods and materials
4. Vendor performance
5. Manufacturers' new orders of nondefense capital goods
6. Building permits of new private housing units
7. Index of stock prices
8. Money supply
9. Interest rate spread
10. Index of consumer expectations

The Conference Board composite index of LEI is an equally weighted index in which its components are standardized to produce constant variances. Details of the LEI can be found on The Conference Board Web site, www.conference-board.org, and the reader is referred to Zarnowitz (1992) for his seminal development of underlying economic assumption and theory of the LEI and business cycles (see Table 2.8).

Let us illustrate a regression of real US GDP as a function of current and lagged LEI. The regression coefficient on the LEI variable, 0.232, in Table 2.9, is highly statistically significant because the calculated *t*-value of 6.84 exceeds 1.96, the 5% critical level. One can reject the null hypothesis of no association between the growth rate of US GDP and the growth rate of the LEI. The reader notes, however, that we estimated the regression line with current, or contemporaneous, values of the LEI series.

The LEI series was developed to "forecast" future economic activity such that current growth of the LEI series should be associated with future US GDP growth rates. Alternatively, one can examine the regression association of the current values of real US GDP growth and previous or lagged values, of the LEI series. How many lags might be appropriate? Let us estimate regression lines using up to four lags of the US LEI series. If one estimates multiple regression lines using the EViews software, as shown in Table 2.10, the first lag of the LEI series is statistically significant, having an estimated *t*-value of 5.73, and the second lag is also statistically significant, having an estimated *t*-value of 4.48. In the regression analysis using three lags of the LEI series, the first and second lagged variables are highly statistically significant, and the third lag is not statistically significant because third LEI lag variable has an estimated *t*-value of only 0.12. The critical

Table 2.8 The conference board leading, coincident, and lagging indicator components

Leading index			Standardization factor
1	BCI-01	Average weekly hours, manufacturing	0.1946
2	BCI-05	Average weekly initial claims for unemployment insurance	0.0268
3	BCI-06	Manufacturers' new orders, consumer goods and materials	0.0504
4	BCI-32	Vendor performance, slower deliveries diffusion index	0.0296
5	BCI-27	Manufacturers' new orders, nondefense capital goods	0.0139
6	BCI-29	Building permits, new private housing units	0.0205
7	BCI019	Stock prices, 500 common stocks	0.0309
8	BCI-106	Money supply, M2	0.2775
9	BCI-129	Interest rate spread, 10-year Treasury bonds less federal funds	0.3364
10	BCI-83	Index of consumer expectations	0.0193
Coincident index			
1	BCI-41	Employees on nonagricultural payrolls	0.5186
2	BCI-51	Personal income less transfer payments	0.2173
3	BCI-47	Industrial production	0.1470
4	BCI-57	Manufacturing and trade sales	0.1170
Lagging index			
1	BCI-91	Average duration of unemployment	0.0368
2	BCI-77	Inventories-to-sales ratio, manufacturing and trade	0.1206
3	BCI-62	Labor cost per unit of output, manufacturing	0.0693
4	BCI-109	Average prime rate	0.2692
5	BCI-101	Commercial and industrial loans	0.1204
6	BCI-95	Consumer installment credit-to-personal income ratio	0.1951
7	BCI-120	Consumer price index for services	0.1886

Table 2.9 Real US GDP and the leading indicators: A contemporaneous examination

Dependent variable: DLOG(RGDP)
Sample(adjusted): 2,210
Included observations: 209 after adjusting endpoints

Variable	Coefficient	Std. error	*t*-Statistic	Prob.
C	0.006170	0.000593	10.40361	0.0000
DLOG(LEI)	0.232606	0.033974	6.846529	0.0000
R^2	0.184638	Mean dependent var		0.007605
Adjusted R^2	0.180699	S.D. dependent var		0.008860
S.E. of regression	0.008020	Akaike info criterion		−6.804257
Sum squared resid	0.013314	Schwarz criterion		−6.772273
Log likelihood	713.0449	*F*-statistic		46.87497l
Durbin–Watson stat	1.594358	Prob(*F*-statistic)		0.000000

t-level at the 10% level is 1.645, for 30 observations, and statistical studies often use the 10% level as a minimum acceptable critical level. The third lag is not statistically significant in the three quarter multiple regression analysis. In the four quarter lags analysis of the LEI series, we report that the lag one variable has a *t*-statistic of

Table 2.10 Real GDP and the conference board leading economic indicators

1959 Q1–2011 Q2								
			Lags (LEI)					
Model	Constant	LEI	One	Two	Three	Four	R^2	F-statistic
RGDP	0.006	0.232					0.181	46.875
(t)	10.400	6.850						
RGDP	0.056	0.104	0.218				0.285	42.267
	9.910	2.750	5.730					
RGDP	0.005	0.095	0.136	0.162			0.353	38.45
	9.520	2.600	3.260	4.480				
RGDP	0.005	0.093	0.135	0.164	0.005		0.351	28.679
	9.340	2.530	3.220	3.900	0.120			
RGDP	0.005	0.098	0.140	0.167	−0.041	0.061	0.369	24.862
	8.850	2.680	3.360	4.050	−0.990	1.670		

Table 2.11 The REG procedure

Dependent variable: DLUSGDP				
Sample(adjusted): 6,210				
Included observations: 205 after adjusting endpoints				
Variable	Coefficient	Std. error	t-Statistic	Prob.
C	0.004915	0.000555	8.849450	0.0000
DLOG(LEI)	0.098557	0.036779	2.679711	0.0080
DLOG(L1LEI)	0.139846	0.041538	3.366687	0.0009
DLOG(L2LEI)	0.167168	0.041235	4.054052	0.0001
DLOG(L3LEI)	−0.041170	0.041305	−0.996733	0.3201
DLOG(L4LEI)	0.060672	0.036401	1.666786	0.0971
R^2	0.384488	Mean dependent var		0.007512
Adjusted R^2	0.369023	S.D. dependent var		0.008778
S.E. of regression	0.006973	Akaike info criterion		−7.064787
Sum squared resid	0.009675	Schwarz criterion		−6.967528
Log likelihood	730.1406	F-statistic		24.86158
Durbin–Watson stat	1.784540	Prob(F-statistic)		0.000000

3.36, highly significant; the second lag has a t-statistic of 4.05, which is statistically significant; the third LEI lag variable has a t-statistic of −0.99, not statistically significant at the 10% level; and the fourth LEI lag variable has an estimated t-statistic of 1.67, which is statistically significant at the 10% level. The estimation of multiple regression lines would lead the reader to expect a one, two, and four variable lag structure to illustrate the relationship between real US GDP growth and The Conference Board LEI series. The next chapter develops the relationship using time series and forecasting techniques. This chapter used regression analysis to illustrate the association between real US GDP growth and the LEI series.

The reader is referred to Table 2.11 for EViews output for the multiple regression of the US real GDP and four quarterly lags in LEI.

Table 2.12 The REG procedure model: MODEL1

Dependent variable: dlRGDP
Number of observations read: 209
Number of observations used: 205
Number of observations with missing values: 4

Analysis of variance

Source	DF	Sum of squares	Mean square	F-value	Pr > F
Model	5	0.00604	0.00121	24.85	<0.0001
Error	199	0.00968	0.00004864		
Corrected total	204	0.01572			
	Root MSE	0.00697	R^2	0.3844	
	Dependent mean	0.00751	Adjusted R^2	0.3689	
	Coeff. var	92.82825			

Parameter estimates

Variable	DF	Parameter estimate	Standard error	t-Value	Pr > |t|	Variance inflation
Intercept	1	0.00492	0.00055545	8.85	<0.0001	0
dlLEI	1	0.09871	0.03678	2.68	0.0079	1.52694
dlLEI_1	1	0.13946	0.04155	3.36	0.0009	1.94696
dlLEI_2	1	0.16756	0.04125	4.06	<0.0001	1.92945
dlLEI_3	1	−0.04121	0.04132	−1.00	0.3198	1.93166
dlLEI_4	1	0.06037	0.03641	1.66	0.0989	1.50421

Collinearity diagnostics

Number	Eigenvalue	Condition index
1	3.08688	1.00000
2	1.09066	1.68235
3	0.74197	2.03970
4	0.44752	2.62635
5	0.37267	2.87805
6	0.26030	3.44367

Proportion of variation

Number	Intercept	dlLEI	dlLEI_1	dlLEI_2	dlLEI_3	dlLEI_4
1	0.02994	0.02527	0.02909	0.03220	0.02903	0.02481
2	0.00016369	0.18258	0.05762	0.00000149	0.06282	0.19532
3	0.83022	0.00047128	0.02564	0.06795	0.02642	0.00225
4	0.12881	0.32579	0.00165	0.38460	0.00156	0.38094
5	0.00005545	0.25381	0.41734	0.00321	0.44388	0.19691
6	0.01081	0.21208	0.46866	0.51203	0.43629	0.19977

We run the real GDP regression with four lags of LEI data in SAS. We report the SAS output in Table 2.12. The Belsley et al. (1980) condition index of 3.4 reveals little evidence of multicollinearity and the collinearity diagnostics reveal no two variables in a row exceeding 0.50. Thus, SAS allows the researcher to specifically address the issue of multicollinearity. We will return to this issue in Chap. 4.

Table 2.13 Modeling dlRGDP by OLS

	Coefficient	Std. error	*t*-Value	*t*-Prob	Part. R^2
Constant	0.00491456	0.0005554	8.85	0.0000	0.2824
dlLEI	0.0985574	0.03678	2.68	0.0080	0.0348
dlLEI_1	0.139846	0.04154	3.37	0.0009	0.0539
dlLEI_2	0.167168	0.04123	4.05	0.0001	0.0763
dlLEI_3	−0.0411702	0.04131	−0.997	0.3201	0.0050
dlLEI_4	0.0606721	0.03640	1.67	0.0971	0.0138
Sigma	0.00697274	RSS	0.00967519164		
R^2	0.384488; $F(5,199) = 24.86$ [0.000]				
Adjusted R^2	0.369023	Log-likelihood	730.141		
No. of observations	205	No. of parameters	6		
Mean(dlRGDP)	0.00751206	S.E.(dlRGDP)	0.00877802		
AR 1–2 test:	$F(2,197) = 3.6873$ [0.0268]*				
ARCH 1–1 test:	$F(1,203) = 1.6556$ [0.1997]				
Normality test:	Chi-squared(2) = 17.824 [0.0001]				
Hetero test:	$F(10,194) = 0.86780$ [0.5644]				
Hetero-X test:	$F(20,184) = 0.84768$ [0.6531]				
RESET23 test:	$F(2,197) = 2.9659$ [0.0538]				

The SAS estimates of the regression model reported in Table 2.12 would lead the reader to believe that the change in real GDP is associated with current, lagged, and twice-lagged LEI.

Alternatively, one could use Oxmetrics, an econometric suite of products for data analysis and forecasting, to reproduce the regression analysis shown in Table 2.13.[6]

An advantage to Oxmetrics is its Automatic Model selection procedure that addresses the issue of outliers. One can use the Oxmetrics Automatic Model selection procedure and find two statistically significant lags on LEI and three outliers: the economically volatile periods of 1971, 1978, and (the great recession of) 2008 (see Table 2.14).

The reader clearly sees the advantage of the Oxmetrics Automatic Model selection procedure.

[6] Ox Professional version 6.00 (Windows/U) (C) J.A. Doornik, 1994–2009, PcGive 13.0.See Doornik and Hendry (2009a, b).

Table 2.14 Modeling dlRGDP by OLS

	Coefficient	Std. error	*t*-Value	*t*-Prob	Part. R^2
Constant	0.00519258	0.0004846	10.7	0.0000	0.3659
dlLEI_1	0.192161	0.03312	5.80	0.0000	0.1447
dlLEI_2	0.164185	0.03281	5.00	0.0000	0.1118
I:1971-01-01	0.0208987	0.006358	3.29	0.0012	0.0515
I:1978-04-01	0.0331323	0.006352	5.22	0.0000	0.1203
I:2008-10-01	−0.0243503	0.006391	−3.81	0.0002	0.0680
Sigma	0.00633157	RSS	0.00797767502		
R^2	0.49248	$F(5,199) = 38.62$ [0.000]			
Adjusted R^2	0.479728	Log-likelihood	749.915		
No. of observations	205	No. of parameters	6		
Mean(dlRGDP)	0.00751206	se(dlRGDP)	0.00877802		
AR 1–2 test:	$F(2,197) = 3.2141$ [0.0423]				
ARCH 1–1 test:	$F(1,203) = 2.3367$ [0.1279]				
Normality test:	Chi-squared (2) = 0.053943 [0.9734]				
Hetero test:	$F(4,197) = 3.2294$ [0.0136]				
Hetero-X test:	$F(5,196) = 2.5732$ [0.0279]				
RESET23 test:	$F(2,197) = 1.2705$ [0.2830]				

Summary

In this chapter, we introduced the reader to regression analysis and various estimation procedures. We have illustrated regression estimations by modeling consumption functions and the relationship between real GDP and The Conference Board LEI. We estimated regressions using EViews, SAS, and Oxmetrics. There are many advantages with the various regression software with regard to ease of use, outlier estimations, collinearity diagnostics, and automatic modeling procedures. We will use the regression techniques in Chap. 4.

Appendix

Let us follow The Conference Board definitions of the US LEI series and its components:

Leading Index Components

BCI-01 Average weekly hours, manufacturing. The average hours worked per week by production workers in manufacturing industries tend to lead the business cycle because employers usually adjust work hours before increasing or decreasing their workforce.

BCI-05 Average weekly initial claims for unemployment insurance. The number of new claims filed for unemployment insurance is typically more sensitive than either total employment or unemployment to overall business conditions, and this series tends to lead the business cycle. It is inverted when included in the leading index; the signs of the month-to-month changes are reversed, because initial claims increase when employment conditions worsen (i.e., layoffs rise and new hirings fall).

BCI-06 Manufacturers' new orders, consumer goods and materials (in 1996 $). These goods are primarily used by consumers. The inflation-adjusted value of new orders leads actual production because new orders directly affect the level of both unfilled orders and inventories that firms monitor when making production decisions. The Conference Board deflates the current dollar orders data using price indexes constructed from various sources at the industry level and a chain-weighted aggregate price index formula.

BCI-32 Vendor performance, slower deliveries diffusion index. This index measures the relative speed at which industrial companies receive deliveries from their suppliers. Slowdowns in deliveries increase this series and are most often associated with increases in demand for manufacturing supplies (as opposed to a negative shock to supplies) and, therefore, tend to lead the business cycle. Vendor performance is based on a monthly survey conducted by the National Association of Purchasing Management (NAPM) that asks purchasing managers whether their suppliers' deliveries have been faster, slower, or the same as the previous month. The slower-deliveries diffusion index counts the proportion of respondents reporting slower deliveries, plus one-half of the proportion reporting no change in delivery speed.

BCI-27 Manufacturers' new orders, nondefense capital goods (in 1996 $). New orders received by manufacturers in nondefense capital goods industries (in inflation-adjusted dollars) are the producers' counterpart to BCI-06.

BCI-29 Building permits, new private housing units. The number of residential building permits issued is an indicator of construction activity, which typically leads most other types of economic production.

BCI-19 Stock prices, 500 common stocks. The Standard & Poor's 500 stock index reflects the price movements of a broad selection of common stocks traded on the New York Stock Exchange. Increases (decreases) of the stock index can reflect both

the general sentiments of investors and the movements of interest rates, which is usually another good indicator for future economic activity.

BCI-106 Money supply (in 1996 $). In inflation-adjusted dollars, this is the M2 version of the money supply. When the money supply does not keep pace with inflation, bank lending may fall in real terms, making it more difficult for the economy to expand. M2 includes currency, demand deposits, other checkable deposits, travelers checks, savings deposits, small denomination time deposits, and balances in money market mutual funds. The inflation adjustment is based on the implicit deflator for personal consumption expenditures.

BCI-129 Interest rate spread, 10-year Treasury bonds less federal funds. The spread or difference between long and short rates is often called the yield curve. This series is constructed using the 10-year Treasury bond rate and the federal funds rate, an overnight interbank borrowing rate. It is felt to be an indicator of the stance of monetary policy and general financial conditions because it rises (falls) when short rates are relatively low (high). When it becomes negative (i.e., short rates are higher than long rates and the yield curve inverts) its record as an indicator of recessions is particularly strong.

BCI-83 Index of consumer expectations. This index reflects changes in consumer attitudes concerning future economic conditions and, therefore, is the only indicator in the leading index that is completely expectations-based. Data are collected in a monthly survey conducted by the University of Michigan's Survey Research Center. Responses to the questions concerning various economic conditions are classified as positive, negative, or unchanged. The expectations series is derived from the responses to three questions relating to (1) economic prospects for the respondent's family over the next 12 months; (2) economic prospects for the Nation over the next 12 months; and (3) economic prospects for the Nation over the next 5 years.

References

Beaton, A.E. and J.W. Tukey, 1974. "The Fitting of Power Series, Meaning Polynomials, Illustrated on Bank-Spectroscopic Data," *Technometrics* 16, 147-185.

Belsley, D.A., E. Kuh, and R.E. Welsch, 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons, Chapter 2.

Burns, A. F. and W.C. Mitchell. 1946. *Measuring Business Cycles*. New York, NBER.

Clements, M. P. and D. F. Hendry. 1998. *Forecasting Economic Time Series*. Cambridge: Cambridge University Press.

Cochrane, D. and G.H. Orcutt. 1949. Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms. *Journal of the American Statistical Association*. 44:32-61.

Cook, R.D. 1977. "Detection of Influential Observations in Linear Regression." *Technometrics* 19, 15–18.

Doornik, J. A. and D. F. Hendry. 2009a. *Empirical Econometric Modelling*. Timberlake Consultants, Ltd.

Doornik, J. A. and D. F. Hendry. 2009b. *Modelling Dynamic Systems*. Timberlake Consultants, Ltd.

Fama, E.F. 1976. *Foundations of Finance*. New York: Basic Books, Inc., Chapter 3.

Guerard, J.B., Jr. and E. Schwartz, 2007. *Quantitative Corporate Finance*. New York: Springer.

Gunst, R.F. and R.L. Mason, 1980. *Regression Analysis and its Application*, New York: Marcel Dekker, Inc.

Huber, P.J., 1973. "Robust Regression: Asymptotics, Conjectures, and Monte Carlo," *Annals of Statistics* 1, 799-82.

Johnston, J. 1972. *Econometric Methods*. 2nd ed. New York: McGraw-Hill.

Makridakis, S., S.C. Wheelwright, and R. J. Hyndman, 1998. *Forecasting: Methods and Applications*. John Wiley & Sons, 3rd edition. Chapters 5, 6.

Mansfield, E. 1994. *Statistics for Business and Economics*. 5th ed. New York: W.W. Norton &. Company.

Miller, Irwin, and John E. Freund. 1965. *Probability and Statistics for Engineers*. Englewood Cliffs, N.J.: Prentice-Hall.

Mitchell, W.C. 1913. *Business Cycles*, New York, Burt Franklin reprint.

Mitchell, W.C. 1951. *What Happens During Business Cycles: A Progress Report*, New York, NBER.

Moore, G.H. 1961. *Business Cycle Indicators*, 2 volumes, Princeton, Princeton University Press.

Murphy, James L. 1973. *Introductory Econometrics*. Homewood, IL: Richard D. Irwin, Inc.

Nelson, C.R. and C.I. Plosser. 1982. "Trends and Random Walks in Macroeconomic Time Series," *Journal of Monetary Economics*, vol. 10, pp. 139-162.

Stiglitz, Joseph. 1993. *Economics*. New York: W. W. Norton & Company.

Sharpe, W. F. 2007. *Investors and Markets*. Princeton; Princeton University Press, Chapter 4.

Zarnowitz, V. 2004. "The Autonomy of Recent US Growth and Business Cycles". In P. Dua, Ed., *Business Cycles and Economic Growth: An Analysis Using Leading Indicators*. New York: Oxford University Press, 44-82.

Zarnowitz, V. 2001. "The Old and the New in the U.S. Economic Expansion." The Conference Board. EPWP #01-01.

Zarnowitz, V. 1992. *Business Cycles: Theory, History, Indicators, and Forecasting*. Chicago, University of Chicago Press.

http://www.springer.com/978-1-4614-5238-6

Introduction to Financial Forecasting in Investment Analysis

Guerard Jr., J.B.

2013, XI, 236 p., Hardcover
ISBN: 978-1-4614-5238-6