APPENDIX B

Regression Analysis[1]

The purpose of this appendix is to provide a quick and informative review of ordinary least squares (OLS) regression. OLS regression is used in almost every field imaginable, from anthropology to zoology. In the field of finance, the most common application of OLS regression is estimating betas for individual stocks. In the finance subfield of derivatives risk management, OLS regression is frequently used in identifying risk-minimizing hedge ratios. In this appendix, we review OLS regression by discussing topics such as regression estimation, testing, and prediction using both simple and multiple regression models. To avoid unnecessary repetition, the content of Appendix A, "Elementary Statistics," is assumed to be background knowledge.

OBJECTIVES

After reviewing this appendix, you should be able to:

1. State and understand the four OLS regression assumptions.
2. Estimate a simple OLS regression model from summary statistics.
3. Interpret OLS regression and ANOVA results from a statistical software package.
4. Perform hypothesis tests and construct confidence intervals for individual regression coefficients.
5. Perform hypothesis tests on an entire model.
6. Calculate and interpret the *R*-squared and adjusted *R*-squared for a model.
7. Choose from among a collection of models based on explanatory power and parsimony.
8. Recognize when model assumptions are violated and understand the consequences.

[1] I am grateful to Jon Stroud for providing assistance in developing this appendix. Nick Bollen, Emma Rasiel, and Tom Smith also provided valuable comments and suggestions.

SIMPLE LINEAR REGRESSION

The goal of regression is to learn about a relation between variables. Pay attention to the adjectives used when describing the word regression. They provide important information regarding the structure of the model being investigated. In this section, we focus on simple linear regression. The term *simple* refers to the fact that we have only *two* variables, X and Y, and the term *linear* refers to the fact that the relation between the variables will be represented by a *line*. In contrast to simple linear regression, *multiple* regression involves more than two variables, and *nonlinear* regression involves a relation between X and Y that is not a straight line.

In simple linear regression, X appears on the right-hand side of the equation and is called the *independent variable*. Other names for it include *explanatory variable* and *predictor variable*. On the left-hand side is Y, the *dependent variable* or *response variable*. In regression, X is assumed to be nonrandom (taking on values that are fixed by the investigator). Y depends linearly on X but also has a random component, ε. Thus the relation between X and Y in a simple linear regression is written

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{B.1}$$

where β_0 represents the intercept of the regression line, β_1 represents the slope, and i represents the ith pair of observations of the variables X and Y. We have assumed only that the relation between X and Y is linear and that the values of X are nonrandom or fixed.

The remaining regression assumptions pertain to the error term, ε_i. First, the expected value of ε_i is 0 and the variance of ε_i is constant across observations, that is, $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$. Note that, if X is nonrandom, the error term will have constant variance if and only if the response variable Y has constant variance. The constant variance assumption is commonly referred to as *homoscedasticity* and is the basis for *ordinary* least squares regression estimation. "Ordinary" applies because every observation of Y_i has equal variance and is therefore given equal weight in the estimation of the model. In contrast, if the response variable Y_i and the error term ε_i have nonconstant variance (i.e., are *heteroscedastic*), a *weighted* least squares approach is appropriate. This allows observations with smaller variances to be given more weight than those with larger variances.

The second assumption governing the residual error term is that the errors are independent of one another, that is, $\text{Cov}(\varepsilon_i,\varepsilon_j) = 0$ for $i \neq j$. Violation of this assumption induces *autocorrelation* or *serial correlation,* a problem frequently encountered in time-series data. Finally, the residual errors are assumed to be normally distributed. Because the X's are nonrandom, this assumption implies that the response variable Y is also normally distributed.

The OLS regression assumptions are summarized in the following statement:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \tag{B.2}$$

The relation between X and Y is linear and the values of X are fixed. The expression, $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, means that the errors are independent and identically distributed (*iid*), where $N(0,\sigma^2)$ signifies the distribution is normal with mean 0 and variance σ^2.

OLS Regression Assumptions

1. The relation between X and Y is linear.
2. The error term ε is independent, identically (normally) distributed with mean 0 and constant variance σ^2.

Before moving on to model estimation, it is important to clarify one commonly misinterpreted point about linear regression. "Linear" refers to the fact that the regression equation is linear in the parameters, and not necessarily in the variables. Consider, for example, a nonlinear model such as

$$Y_i = e^{\beta_0 + \beta_1 X_i + \varepsilon_i}$$

On face appearance, linear regression seems inappropriate. Such a model, however, is *inherently linear* in the sense that it may be re-specified as the linear model,

$$\ln Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Nonlinear models that can be re-specified into a linear form using only a transformation of the X or Y variables are still considered to be linear.

Ordinary Least Squares (OLS) Estimation

Under the condition that our data satisfy the three assumptions of OLS regression, we can proceed by estimating the model in the following way. First, we denote the estimated regression line by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

where $\hat{\beta}_0$ represents our best guess for the true intercept β_0, $\hat{\beta}_1$ is our best guess for the population slope β_1, and $\hat{Y}$ is the predicted value of Y that falls along the regression line. In order to calculate this line, all we need to do is choose values of $\hat{\beta}_0$ and $\hat{\beta}_1$. This is done here using a method known as *ordinary least square* (OLS) estimation. As noted earlier, "ordinary" arises because every observation is assumed to have equal variance and is therefore given equal weight in the estimation of the model.[2] "Least squares" is used because we will choose the line that minimizes the squared distances between the observed and

[2] The application of regression techniques sometimes requires weighting observations unequally. For an explanation of *weighted least squares* regression, see Pindyck and Rubinfeld (1998).

the predicted response variables. Defining the sample residual e_i as $Y_i - \hat{Y}_i$, the OLS requirement is explained in the next section.

OLS Requirement

The sum of the squared residuals,

$$\sum_{i=1}^{n} e_i^2$$

is minimized.

Among other things, minimizing the sum of squares errors implies that the sum of the regression errors (and the average error, for that matter) will be equal to zero. This means that the regression can be re-expressed in deviations from the mean form. That is, if the mean in the regression model (B.2) is 0, the mean value of Y is

$$\overline{Y} = \beta_0 + \beta_1 \overline{X} \tag{B.3}$$

where $\overline{X}$ is the mean of X. Taking the difference between the expressions,

$$\begin{aligned} Y_i - \overline{Y} &= \beta_0 + \beta_1 X_i + e_i - \beta_0 - \beta_1 \overline{X} \\ &= \beta_1 (X_i - \overline{X}) + e_i \end{aligned}$$

Expressing the deviations from the mean as $y_i = Y_i - \overline{Y}$ and $x_i = X_i - \overline{X}$, the regression equation becomes

$$y_i = \beta_1 x_i + e_i \tag{B.4}$$

Next, the least squares *estimators* of β_0 and β_1 are identified. To do so, write the sum of squared errors,

$$\sum_{i=1}^{n} e_i^2$$

as

$$\begin{aligned} \sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2 \\ &= \sum_{i=1}^{n} (y_i^2 + \beta_1^2 x_i^2 - 2\beta_1 x_i y_i) \end{aligned} \tag{B.5}$$

Differentiating (B.5) with respect to β_1,

$$\frac{d\sum_{i=1}^{n} e_i^2}{d\beta_1} = \sum_{i=1}^{n} (2\beta_1 x_i^2 - 2x_i y_i) \tag{B.6}$$

Setting (B.6) equal to 0, simplifying, and rearranging provides the least squares estimator of the slope coefficient, that is,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{B.7}$$

Because the mean residual error is zero, the estimator for the intercept follows from (B.3), that is,

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \tag{B.8}$$

ILLUSTRATION B.1 Estimate beta for common stock.

A common application of simple linear regression in finance is estimating a stock's beta coefficient or relative systematic risk.[3] *A stock's beta is the slope coefficient in a regression of a stock's return on the return of the market portfolio. Suppose you are interested in the relation between General Electric's (ticker symbol: GE) stock return and the return of the S&P 500 portfolio. To learn more about the relation, you collect annual return data for both series over the period January 1985 through December 2004. The data are contained in the worksheet* ***B-1*** *in the Excel file* ***B Illustrations.xls.*** *Find the OLS regression line and interpret the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ where the dependent variable is GE's annual return and the independent variable X is the S&P 500 return.*

The first step in applying the regression coefficient estimators (B.7) and (B.8) is to compute the means of *X* and *Y*. Next, compute the deviations from the mean for the *X* and *Y* variables. Denote them as *x* and *y*. Finally, compute the products of the deviations, that is, *xx*, *yy*, and *xy*. The results are:

[3] The theoretical importance of beta is motivated by the capital asset pricing model discussed in Chapter 3.

	Annual Returns		Deviations from Mean		Products		
Year Ending	GE *Y*	S&P 500 *X*	*y*	*x*	*xy*	*yy*	*xx*
12/31/1985	0.33066	0.2633	0.1268	0.1470	0.0186	0.0161	0.0216
12/31/1986	0.21803	0.1462	0.0141	0.0298	0.0004	0.0002	0.0009
12/31/1987	0.05246	0.0203	–0.1514	–0.0961	0.0146	0.0229	0.0092
12/30/1988	0.04895	0.1240	–0.1549	0.0076	–0.0012	0.0240	0.0001
12/29/1989	0.48758	0.2725	0.2837	0.1561	0.0443	0.0805	0.0244
12/31/1990	–0.08170	–0.0656	–0.2856	–0.1820	0.0520	0.0816	0.0331
12/31/1991	0.37194	0.2631	0.1680	0.1467	0.0247	0.0282	0.0215
12/31/1992	0.15068	0.0446	–0.0532	–0.0717	0.0038	0.0028	0.0051
12/31/1993	0.26017	0.0705	0.0563	–0.0458	–0.0026	0.0032	0.0021
12/30/1994	0.00252	–0.0154	–0.2014	–0.1318	0.0265	0.0406	0.0174
12/29/1995	0.45125	0.3411	0.2474	0.2247	0.0556	0.0612	0.0505
12/31/1996	0.40347	0.2026	0.1996	0.0862	0.0172	0.0398	0.0074
12/31/1997	0.50937	0.3101	0.3055	0.1937	0.0592	0.0933	0.0375
12/31/1998	0.40974	0.2667	0.2058	0.1503	0.0309	0.0424	0.0226
12/31/1999	0.53542	0.1953	0.3315	0.0789	0.0261	0.1099	0.0062
12/29/2000	–0.06041	–0.1014	–0.2643	–0.2178	0.0576	0.0699	0.0474
12/31/2001	–0.15021	–0.1304	–0.3541	–0.2468	0.0874	0.1254	0.0609
12/31/2002	–0.37653	–0.2336	–0.5804	–0.3500	0.2032	0.3369	0.1225
12/31/2003	0.30742	0.2638	0.1035	0.1474	0.0153	0.0107	0.0217
12/31/2004	0.20708	0.0899	0.0032	–0.0265	–0.0001	0.0000	0.0007
Mean	0.20390	0.1164					
Total			0.0000	0.0000	0.7335	1.1895	0.5130

Based on these results, the estimate of the slope is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{0.7335}{0.5130} = 1.4298$$

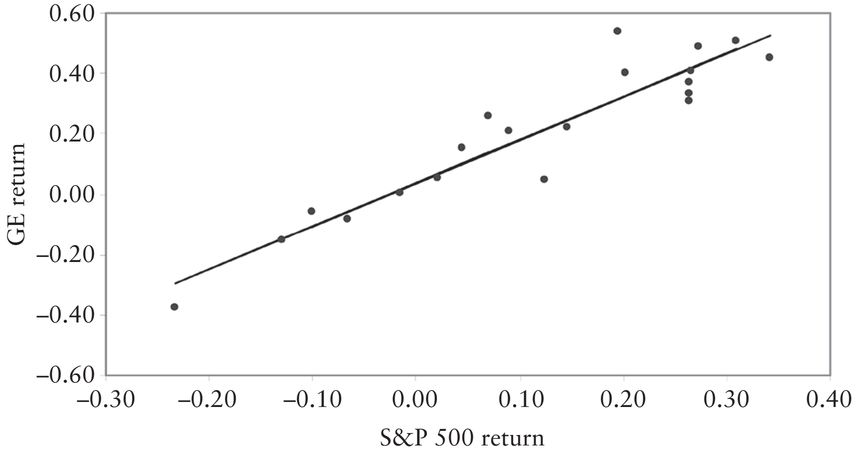
and the estimate of the intercept is

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 0.2039 - 1.4298(0.1164) = 0.0375$$

Combining results, the estimated OLS regression line is

$$\hat{Y}_i = 0.0375 + 1.4298 X_i$$

The actual returns of GE conditional on the returns of the S&P 500 and the returns predicted by the regression line are summarized in the figure below. The line segment represents the "best" (i.e., least squares) fit of a line drawn through the pairs of actual returns. Note that the line segment is drawn only through the range of observed values of the S&P 500 return. The reason is that regression is only valid over the range of the data, that is, S&P 500 returns in the range between the sample's minimum and maximum returns (–0.2336,0.3411). The slope coefficient $\hat{\beta}_1$ can be interpreted as follows: for an increase of 1% in S&P 500 return, the return on GE will change by about 1.4298%. Stocks like GE with $\beta_1 < 1$ are called *aggressive stocks* because they do better than the market when the market goes up, and worse than the market when the market goes down. Stocks with are called *defensive stocks* for the opposite reasons.



Hypothesis Tests for Individual Regression Parameters

After a model has been fitted, we may wish to test whether the independent variable has a significant effect on the response variable. This can be done by testing the hypothesis that the parameter β_1 equals zero. If the slope of the regression line is zero, this implies that there is no linear relation between X and Y; in this case, the "true" regression line is $E(Y|X) = \beta_0$, and we are just as well off using the sample mean, $\bar{Y}$, to predict future Y's. A population slope not equal to zero implies that the variables are somehow linearly correlated. A regression line with slope greater than zero means that X has a positive effect on Y, and one with a downward slope implies a negative relation between the two variables. We can test for any of these relations using the data collected in our sample.

Suppose we are interested in learning about a relation between two variables. The variables are believed to be linearly related, but it is not known whether the relation is negative or positive. A two-tailed hypothesis test can be used to check whether the population slope is equal to or unequal to zero. (If we had an idea that the slope was either positive or negative, then we would use a one-tailed test). But before any testing is done, we must preset a desired level of the test, denoted by α. In our case, α represents the probability of incorrectly concluding that the two variables are related in some linear manner ($\beta_1 \neq 0$), when in fact they are not ($\beta_1 = 0$). A researcher can formalize these possibilities by specifying two different hypotheses: a null hypothesis, denoted H_0, and an alternative hypothesis, denoted H_1. The alternative hypothesis usually represents what the researcher is trying to prove, for example, that the variables are related; and the null hypothesis usually refers to the status quo, or the accepted state of the world. So, in our example, we would choose the null hypothesis to be $H_0 : \beta_1 = 0$ and the alternative as $H_1 : \beta_1 \neq 0$. Thus once we have stated our null and alternative hypotheses, specified the level of the test, and collected the data, we can formally test our beliefs.

Based on our sample, the best guess for the "true" slope of the regression line is $\hat{\beta}_1$. There is, however, error associated this estimate. This inaccuracy can be quantified by the standard error of the estimate, $s_{\hat{\beta}_1}$, which is defined as

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_{i=1}^{n} x_i^2}} \tag{B.9}$$

where

$$s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} \tag{B.10}$$

is an estimate for σ, the standard deviation of the error. Using this measure of uncertainty, we can standardize our parameter estimate to get the test statistic,

$$t = \hat{\beta}_1 / s_{\hat{\beta}_1}$$

which follows a t-distribution with $n - 2$ degrees of freedom. If the absolute value of $\hat{\beta}_1$ is much larger than its standard error, then t will also grow large in absolute value, indicating that β_1 is different than zero. A large positive value of the test statistic is evidence of a positive relation, whereas a large negative value of the test statistic is a strong indication of a negative relation. Because the test statistic has a t-distribution, we can set cutoffs, or critical values, for rejecting the null hypothesis for any specified level of significance. The table that follows gives the rejection rules for three types of hypothesis tests of the regression line slope. Given a probability α that a t-distributed random variable (with $n - 2$ degrees of freedom) is greater than some *critical value* t_α, the following rules apply:

Common Hypothesis Tests for the Slope of a Regression Line

Null Hypothesis	Alternative Hypothesis	Rejection Rule
H_0: $\beta_1 = 0$	H_1: $\beta_1 \neq 0$	Reject H_0 if $\lvert t \rvert > t_{\alpha/2}$
H_0: $\beta_1 = 0$	H_1: $\beta_1 > 0$	Reject H_0 if $t > t_\alpha$
H_0: $\beta_1 = 0$	H_1: $\beta_1 < 0$	Reject H_0 if $t < -t_\alpha$

Recall the critical t-values for different levels of α were tabulated in Table A.5 of Appendix A. Note that the first row of the table represents a "two-tailed" hypothesis test. Since the alternative hypothesis, $\beta_1 \neq 0$, does not specify whether β_1 is greater than or less than 0, we compare the absolute value of the t-statistic with the critical t-value corresponding to a probability $\alpha/2$, that is, $\alpha/2$ in each tail of the two tails of the distribution.

We can use a similar procedure to perform hypothesis tests on the intercept. We specify the null and alternative hypotheses involving β_0, select the level of significance, and calculate $\hat{\beta}_0$ according to the formula given above, and its standard error using

$$s_{\hat{\beta}_0} = s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2}} \tag{B.11}$$

Then

$$t = (\hat{\beta}_0 - c)/s_{\hat{\beta}_0}$$

is the test statistic for the null hypothesis $H_0 : \beta_0 = c$ that the intercept is equal to some value c. (This is more general than the decision rules for the slope. For the slope, an investigator usually wants to ascertain whether it is equal to something other than zero. To avoid confusion, the tests of β_1 assume that c is always equal to zero.) It should also be noted that, in a simple regression setting, hypothesis tests for the slope are far more common than for the intercept.

Common Hypothesis Tests for the Intercept of a Regression Line

Null Hypothesis	Alternative Hypothesis	Rejection Rule
H_0: $\beta_0 = c$	H_1: $\beta_0 \neq c$	Reject H_0 if $\|t\| > t_{\alpha/2}$
H_0: $\beta_0 = c$	H_1: $\beta_0 > c$	Reject H_0 if $t > t_\alpha$
H_0: $\beta_0 = c$	H_1: $\beta_0 < c$	Reject H_0 if $t < -t_\alpha$

ILLUSTRATION B.2 Test hypothesis that slope and intercept are zero.

Using the return data from Illustration B.1, perform a hypothesis test of $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ at the $\alpha = 0.05$ level of significance. Also test the hypothesis $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 > 0$ at the $\alpha = 0.05$ level of significance.

The first step is to calculate the squared errors in the regression. Using the estimated intercept and the slope coefficients, the errors and squared errors are as follows:

	Annual Returns				
Year Ending	GE Y	S&P 500 X	Predicted Y	e	e^2
12/31/1985	0.3307	0.2633	0.4140	–0.0833	0.0069
12/31/1986	0.2180	0.1462	0.2465	–0.0285	0.0008
12/31/1987	0.0525	0.0203	0.0665	–0.0140	0.0002
12/30/1988	0.0490	0.1240	0.2148	–0.1659	0.0275
12/29/1989	0.4876	0.2725	0.4271	0.0604	0.0037
12/31/1990	–0.0817	–0.0656	–0.0563	–0.0254	0.0006
12/31/1991	0.3719	0.2631	0.4137	–0.0417	0.0017
12/31/1992	0.1507	0.0446	0.1013	0.0494	0.0024
12/31/1993	0.2602	0.0705	0.1384	0.1218	0.0148
12/30/1994	0.0025	–0.0154	0.0155	–0.0129	0.0002
12/29/1995	0.4512	0.3411	0.5252	–0.0740	0.0055
12/31/1996	0.4035	0.2026	0.3272	0.0763	0.0058
12/31/1997	0.5094	0.3101	0.4809	0.0285	0.0008
12/31/1998	0.4097	0.2667	0.4188	–0.0091	0.0001
12/31/1999	0.5354	0.1953	0.3167	0.2187	0.0479
12/29/2000	–0.0604	–0.1014	–0.1075	0.0471	0.0022
12/31/2001	–0.1502	–0.1304	–0.1490	–0.0012	0.0000
12/31/2002	–0.3765	–0.2336	–0.2966	–0.0799	0.0064
12/31/2003	0.3074	0.2638	0.4147	–0.1073	0.0115
12/31/2004	0.2071	0.0899	0.1661	0.0410	0.0017
Mean	0.2039	0.1164	0.2039	0.0000	
Total					0.1408

Note that the sum of the errors is zero, as expected. The sum of the squared errors is 0.1408.

Next we apply (B.10) to obtain the standard error of the estimate,

$$s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} = \sqrt{\frac{0.1408}{20-2}} = 0.0884$$

The standard error of the slope coefficient from (B.9) is

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_{i=1}^{n} x_i^2}} = \frac{0.0884}{\sqrt{0.5130}} = 0.1235$$

Under the null hypothesis that the slope is zero, the test statistic, t, is

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{1.4298}{0.1235} = 11.5799$$

Since we are performing a two-tailed test at the 5% probability level, we look up the critical t-value at a probability level of $\alpha/2$ or 2.5% and 18 degrees of freedom. From Table C.3 in Appendix C, we see that the critical value is $t_{\alpha/2} = 2.101$. Since our sample t-statistic, 11.5799, exceeds the critical value, we can reject the null hypothesis that $\beta_1 = 0$.

The procedure for testing the null hypothesis that $\beta_0 = 0$ follows a similar procedure. The standard error of the intercept from (B.11) is

$$s_{\hat{\beta}_0} = s\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} x_i^2}} = 0.0884\sqrt{\frac{1}{20} + \frac{0.1164^2}{0.5130}} = 0.0244$$

The t-statistic is therefore

$$t = \frac{\hat{\beta}_0}{s_{\hat{\beta}_0}} = \frac{0.0375}{0.0244} = 1.5335$$

Since we are testing the null against the alternative hypothesis that $\beta_0 > 0$, a one-tailed test is appropriate. From Table C.3, the critical t-value for a one-tailed test at the 5% significance level is $t_\alpha = 1.734$. Since the sample t-statistic, 1.5335, is not greater than 1.734, we cannot reject the null hypothesis that the intercept is zero.

Confidence Intervals

The idea of constructing confidence intervals is very much related to that of hypothesis testing. Again we are usually interested in finding out whether the independent variable in the regression has a significant effect on the dependent variable. But this time, instead of using a test statistic and critical value, we construct intervals and attempt to "pin down" the true value of a parameter and base our inferences on that. The tools used in constructing confidence intervals are identical to those used in hypothesis testing as previously shown.

To calculate a confidence interval for the slope of the regression line, we need only three things: the point estimate for the parameter, the standard error of the estimate, and the confidence coefficient that is taken from a t-distribution with $n - 2$ degrees of freedom. The interval itself is just the estimate of the parameter plus or minus some margin, which is related to the standard error of the estimate and the selected level of confidence.

Confidence Intervals for the Intercept and Slope of a Regression Line

Parameter	Interval Size	Confidence Interval
Intercept	$(1 - \alpha)\%$	$\hat{\beta}_0 \pm t_{\alpha/2} s_{\hat{\beta}_0}$
Slope	$(1 - \alpha)\%$	$\hat{\beta}_1 \pm t_{\alpha/2} s_{\hat{\beta}_1}$

The confidence coefficient, α, identical to the critical value used in hypothesis testing, is based on a t-distribution with $n - 2$ degrees of freedom and is chosen by the modeler to give a specified level of confidence. Clearly, a larger interval will yield a higher level of confidence and vice versa. The most common confidence widths are 90%, 95%, and 99%.

After the formulas above are used to construct confidence intervals for a population parameter, one can check to see if a certain value of interest falls in the interval. If we are testing that the independent variable has an effect on Y, for example, we should construct a confidence interval for the slope. If zero is contained in the interval, then the data does not give sufficient evidence that X has an effect on Y. Conversely, if the interval does not contain zero, then there is evidence that the two variables are related at the α level of confidence. The conclusions obtained from building confidence intervals will give the exact same results as using hypothesis tests.

ILLUSTRATION B.3 Compute confidence interval for slope.

Using the return data from Illustration B.1, compute a 95% confidence interval for β_1. Does the interval contain zero?

A 95% confidence interval for β_1 is given by $\hat{\beta}_1 \pm t_{\alpha/2}s(\hat{\beta}_1)$. The *t*-statistic again comes from a *t*-distribution with 18 degrees of freedom. The critical t-value is 2.101. The estimate and the standard error of β_1 are as calculated above. Therefore a 95% interval is

$$1.4298 \pm 2.101(0.1235) = (1.1704,1.6892)$$

Note that the 95% confidence interval for β_1 ranges from 1.1704 to 1.6892 and does not include 0. We should not be surprised by this result since we had already concluded that β_1 is not equal to zero by virtue of a *t*-test. By the same logic, we also know that β_1 is not equal to 1 at the 5% probability level.

Prediction

Another reason for using ordinary least squares regression is to predict future observations. Forecasting sales as a function of marketing expenses is an example. Prediction can be summarized in one paragraph in the following way. Once a linear model has been fitted using previous data, our best guess of Y (call it $\hat{Y}_p$) for a specified value of X (call it X_p) is

$$\hat{Y}_p = \hat{\beta}_0 - \hat{\beta}_1 X_p$$

We can also quantify our uncertainty about the estimate with a prediction interval. A $(1 - \alpha)$% confidence interval for a new observation Y_p is $\hat{Y}_p \pm t_{\alpha/2}s_p$ where the standard error of the prediction is

$$s_p = s\sqrt{1 + \frac{1}{n} + \frac{(X_p - \overline{X})^2}{\sum_{i=1}^{n} x_i^2}} \tag{B.12}$$

A warning about prediction, however. The estimates and intervals for new values of Y are only valid if X_p falls within the range of the X values used in the regression. Extrapolation must be treated with extreme caution.

ILLUSTRATION B.4 Develop stock return prediction based on market return.

Using the return data from Illustration B.1, compute the predicted annualized rate of return for GE assuming the return on the S&P 500 is 30%. Compute the return assuming the S&P 500 return is 40%. Which predicted is more reliable, and why? Show that your intuition is consistent with the confidence interval of each of the predictions.

For a 30% S&P 500 return, GE's predicted return is

$$\hat{Y}_p = 0.0375 + 1.4298(0.30) = 46.64\%$$

For a 40% S&P 500 return, GE's predicted return is

$$\hat{Y}_p = 0.0375 + 1.4298(0.40) = 60.94\%$$

The predicted return for the 30% S&P 500 return is more believable. This is because the 40% S&P 500 return falls well outside the range of our data. The minimum and maximum S&P 500 returns in the sample are –23.36% and 34.11%, respectively.

The 95% confidence interval for each of our predictions confirms our intuition. For the 30% S&P 500 return, the standard error of the prediction is

$$s_p = s\sqrt{1 + \frac{1}{n} + \frac{(X_p - \overline{X})^2}{\sum_{i=1}^{n} x_i^2}} = 0.0884\sqrt{1 + \frac{1}{20} + \frac{(0.30 - 0.1164)^2}{0.5130}} = 0.0934$$

and the 95% confidence interval is

$$0.4664 \pm 2.101(0.0934) = (0.2703, 0.6626)$$

For the 40% S&P 500 return, the confidence interval is

$$0.6094 \pm 2.101(0.0971) = (0.4054, 0.8134)$$

Note that we are more confident in our prediction of GE stock return (i.e., the range of the confidence interval is 39.23%) when the S&P 500 return is 30% than we are when the S&P 500 return is 40% and the range is 40.80%.

Goodness of Fit

Another important aspect of regression analysis is model testing. This can be used when trying to assess a model's predictive power or when choosing between two or more models. A common way to measure goodness of fit is by decomposing the sum of squares of the data into the amount explained by the model and the amount left unexplained. The higher the amount explained by the model, the better the model. The decomposition is done as follows. For a single variable Y with n observations, the *total sum of squares* is given by

$$SST = \sum_{i=1}^{n} y_i^2$$

Once we have fitted a model, we obtain the fitted values, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, for each observation of X. The squared distances between these predictions and the overall mean $\overline{Y}$ give the *regression* or *explained sum of squares,*

$$SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} \hat{y}_i^2$$

which is the amount of the *SST* explained by the model. Finally, the part left unexplained by the model, called the *error* or *residual sum of squares* is just

$$SSE = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} \hat{y}_i^2$$

By Pythagorus' Theorem, the regression and error sum of squares must add up to the total sum of squares. This type of decomposition is closely related to ANOVA, or *analysis of variance*, and is frequently used to determine how well an estimated model fits the data.

To illustrate, consider now a model that perfectly predicts all the data points, i.e., $\hat{Y}_i = Y_i$ for all *i*. In this case, the regression sum of squares equals the total sum of squares and the error sum of squares equals zero—a perfect fit. On the other hand, a model with an estimated slope $\hat{\beta}_1 = 0$ will have no predictive power at all (because $\hat{Y}_i = \hat{\beta}_0$ for all *i*), and therefore the total sum of squares will equal the error sum of squares, leaving the variation explained by the model as zero.

R-Squared A commonly used indicator of regression goodness of fit is the R^2 statistic. It is also referred to as the *coefficient of determination* and represents the proportion of the total variation that is explained by the model. In the case of simple linear regression, R^2 is the square of the correlation between X and Y. The R^2 is simply the ratio of the regression sum of squares (*SSR*) to the total sum of squares (*SST*):

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} \hat{y}_i^2}$$

Since the range of *SSR* is 0 to *SST*, the range of R^2 is from 0 to 1. A perfect model fit will yield an R^2 of 1, and a model with no explanatory power whatsoever gives $R^2 = 0$. In general, a model with a high R^2 is preferred to one with a low one.

ILLUSTRATION B.5 Compute sums of squares and *R*-square.

Using the return data from Illustration B.1, compute the total sum of squares (SST), the regression sum of squares (SSR), and the error sum of squares (SSE) for our estimated return prediction model. Also, compute the regression R*-squared.*

Y_i	X_i	y_i	y_i^2	$\hat{Y}_i$	e_i	e_i^2	$\hat{y}_i$	$\hat{y}_i^2$
0.3307	0.2633	0.1268	0.0161	0.4140	–0.0833	0.0069	0.2101	0.0441
0.2180	0.1462	0.0141	0.0002	0.2465	–0.0285	0.0008	0.0426	0.0018
0.0525	0.0203	–0.1514	0.0229	0.0665	–0.0140	0.0002	–0.1374	0.0189
0.0490	0.1240	–0.1549	0.0240	0.2148	–0.1659	0.0275	0.0109	0.0001
0.4876	0.2725	0.2837	0.0805	0.4271	0.0604	0.0037	0.2233	0.0498
–0.0817	–0.0656	–0.2856	0.0816	–0.0563	–0.0254	0.0006	–0.2602	0.0677
0.3719	0.2631	0.1680	0.0282	0.4137	–0.0417	0.0017	0.2098	0.0440
0.1507	0.0446	–0.0532	0.0028	0.1013	0.0494	0.0024	–0.1026	0.0105
0.2602	0.0705	0.0563	0.0032	0.1384	0.1218	0.0148	–0.0655	0.0043
0.0025	–0.0154	–0.2014	0.0406	0.0155	–0.0129	0.0002	–0.1884	0.0355
0.4512	0.3411	0.2474	0.0612	0.5252	–0.0740	0.0055	0.3213	0.1032
0.4035	0.2026	0.1996	0.0398	0.3272	0.0763	0.0058	0.1233	0.0152
0.5094	0.3101	0.3055	0.0933	0.4809	0.0285	0.0008	0.2770	0.0767
0.4097	0.2667	0.2058	0.0424	0.4188	–0.0091	0.0001	0.2149	0.0462
0.5354	0.1953	0.3315	0.1099	0.3167	0.2187	0.0479	0.1128	0.0127
–0.0604	–0.1014	–0.2643	0.0699	–0.1075	0.0471	0.0022	–0.3114	0.0970
–0.1502	–0.1304	–0.3541	0.1254	–0.1490	–0.0012	0.0000	–0.3529	0.1245
–0.3765	–0.2336	–0.5804	0.3369	–0.2966	–0.0799	0.0064	–0.5005	0.2505
0.3074	0.2638	0.1035	0.0107	0.4147	–0.1073	0.0115	0.2108	0.0444
0.2071	0.0899	0.0032	0.0000	0.1661	0.0410	0.0017	–0.0378	0.0014
Total		0.0000	1.1895		0.0000	0.1408	0.0000	1.0487

The table above shows the raw computations of the sums of squared errors. The sum of the squared deviations of Y_i about its mean (*SST*) is 1.1895, the sum of squared errors (*SSE*) is 0.1408, and the regression sum of squares (*SSR*) is 1.0487. The regression R^2 is therefore

$$R^2 = \frac{SSR}{SST} = \frac{1.0487}{1.1895} = 0.8817 \text{ or } 88.17\%.$$

Applying a Regression Program

The purpose of reviewing simple linear regression in such detail is to remove the mystery of regression analysis. A summary of all of the simple linear regression estimation formulas is contained in Table B.1. In practice, all of these formulas are preprogrammed in a number of different regression software packages. Microsoft Excel 2003, for example, has linear regression as one of its data analysis functions. Below we illustrate the application of the regression analysis function.

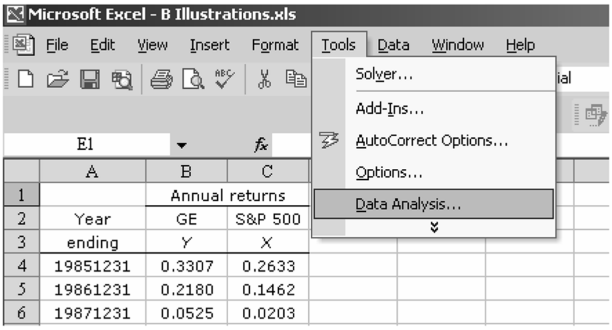
TABLE B.1 Summary of simple linear regression estimation formulas.

Estimator for slope:	$\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$	Standard error:	$s(\hat{\beta}_1) = \dfrac{s}{\sqrt{\sum_{i=1}^{n} x_i^2}}$
Estimator for intercept:	$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$	Standard error:	$s(\hat{\beta}_0) = s\sqrt{\dfrac{1}{n} + \dfrac{\overline{X}^2}{\sum_{i=1}^{n} x_i^2}}$
Prediction:	$\hat{Y}_p = \hat{\beta}_0 - \hat{\beta}_1 X_p$	Standard error:	$s_p = s\sqrt{1 + \dfrac{1}{n} + \dfrac{(X_p - \overline{X})^2}{\sum_{i=1}^{n} x_i^2}}$
Standard error of the estimate:	$s = \sqrt{\dfrac{\sum_{i=1}^{n} e_i^2}{n-2}}$		
R-squared:	$R^2 = \dfrac{\text{Explained variation}}{\text{Total variation}} = \dfrac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} y_i^2}$		

ILLUSTRATION B.6 Estimate simple linear regression using Excel regression routine.

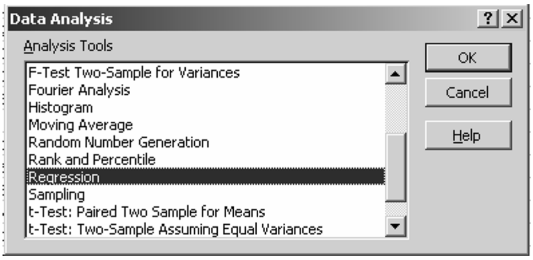
Estimate the ordinary least squares regression of GE's stock returns on the returns of the S&P 500 using the Microsoft Excel regression function. The return data are included in worksheet ***B6*** *of the Excel file,* ***B Illustrations.xls****.*

The first step is to click Data Analysis in the Tools menu.

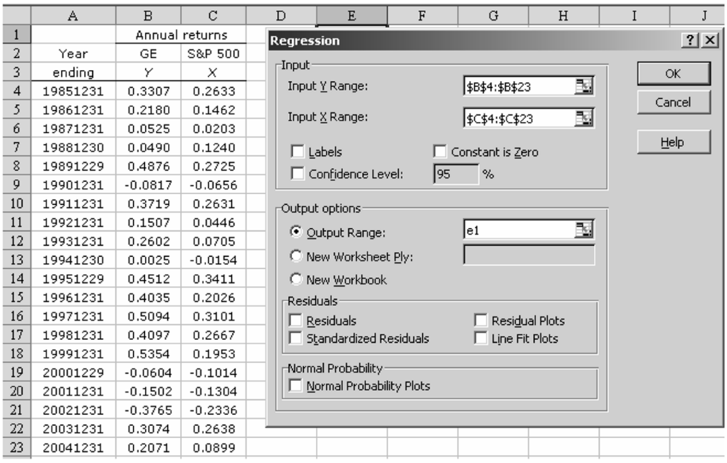


	A	B	C
1		Annual returns	
2	Year	GE	S&P 500
3	ending	Y	X
4	19851231	0.3307	0.2633
5	19861231	0.2180	0.1462
6	19871231	0.0525	0.0203

From the Data Analysis dialog box, AnalysisTools list, click Regression.



When the Regression dialog box appears, enter the vector of *Y* observations in the Input Y Range, and the vector *X* observations in the Input X Range. Click the Output Range option button under the Output options, and then enter the location of the cell which will be the upper left-hand corner of the output panel of results. Click OK.



	A	B	C
1		Annual returns	
2	Year	GE	S&P 500
3	ending	Y	X
4	19851231	0.3307	0.2633
5	19861231	0.2180	0.1462
6	19871231	0.0525	0.0203
7	19881230	0.0490	0.1240
8	19891229	0.4876	0.2725
9	19901231	-0.0817	-0.0656
10	19911231	0.3719	0.2631
11	19921231	0.1507	0.0446
12	19931231	0.2602	0.0705
13	19941230	0.0025	-0.0154
14	19951229	0.4512	0.3411
15	19961231	0.4035	0.2026
16	19971231	0.5094	0.3101
17	19981231	0.4097	0.2667
18	19991231	0.5354	0.1953
19	20001229	-0.0604	-0.1014
20	20011231	-0.1502	-0.1304
21	20021231	-0.3765	-0.2336
22	20031231	0.3074	0.2638
23	20041231	0.2071	0.0899

Regression

Input

Input Y Range: $B$4:$B$23

Input X Range: $C$4:$C$23

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

● Output Range: e1

○ New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

The following results will appear.

Summary Output	
Regression Statistics	
Multiple R	0.9390
R-Square	0.8817
Adjusted R-Square	0.8751
Standard Error	0.0884
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	Significance *F*
Regression	1	1.0487	1.0487	134.10	0.0000
Residual	18	0.1408	0.0078		
Total	19	1.1895			

	Coefficients	Std. Error	*t* Stat	P-value	Lower 95%	Upper 95%
Intercept	0.0375	0.0244	1.5335	0.1425	–0.0139	0.0888
X Variable 1	1.4298	0.1235	11.5799	0.0000	1.1704	1.6892

In the first table of results, note that our computations are consistent with the reported values R^2 and the standard error of the regression. The "multiple R" is simply the square root of R^2. Finally, we have not yet discussed the adjusted R^2. We will do so in the multiple regression analysis section that follows. It is an important statistic for deciding between competing model specifications.

The second table is labeled ANOVA, short for analysis of variance. Like the adjusted R^2, the ANOVA results are most typically used in a multiple regression context. At this juncture, it is sufficient to recognize only that ANOVA results are based on the sums of squares computations that we discussed earlier. Note that, in our computations in earlier illustrations, we identified the *SSR*, *SSE*, and *SST* values reported in the column with the heading *SS*.

The third table contains the parameter estimates, standard errors, and 95% confidence intervals. The reported values are, again, consistent with our computations. The t-ratios correspond to the null hypothesis that the coefficient is 0. The Excel regression also reports the p-value of each coefficient under the null hypothesis that the coefficient equals 0. The p-value is the probability that the sample t-statistic was observed by chance. The p-value for the intercept term, for example, is 0.1425. This means that the probability of the sample t-statistic, 1.5335, was observed by chance is 14.25%. Since conventional hypothesis testing usually involves 5% or 1% cutoff levels, we cannot reject the hypothesis that the intercept is different from 0.

OLS REGRESSION THROUGH ORIGIN

On occasion, it is necessary to consider a simple regression whose intercept term, for economic reasons, equals zero, that is,

$$Y_i = \beta_1 X_i + \varepsilon_i \tag{B.13}$$

As before, the error term ε_i is assumed to be independent, identically (normally) distributed with mean zero and constant variance. The least squares estimator of β_1 is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \tag{B.14}$$

which is similar to (B.7) except the levels of X_i and Y_i are used rather than their deviations from their respective means. The standard error of the estimate, $s(\hat{\beta}_1)$, is defined as

$$s(\hat{\beta}_1) = \frac{s}{\sqrt{\sum_{i=1}^{n} X_i^2}} \tag{B.15}$$

where

$$s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-1}} \tag{B.16}$$

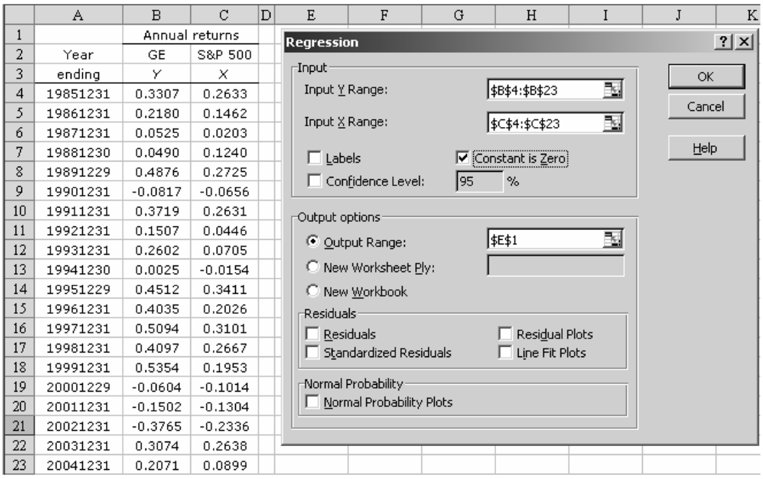
The standard error of the prediction is

$$s_p = s\sqrt{1 + \frac{X_p^2}{\sum_{i=1}^{n} X_i^2}} \tag{B.17}$$

ILLUSTRATION B.7 Estimate simple linear regression through the origin using Excel regression routine.

Estimate the ordinary least squares regression through the origin of GE's stock returns on the returns of the S&P 500 using the Microsoft Excel regression function. The return data are included in worksheet ***B7*** *of the Excel file,* ***B Illustrations.xls****.*

The steps in applying the Excel regression function are the same as in Illustration B.6, except that when the regression dialog box appears, Constant is Zero option.



	A	B	C
1		Annual returns	
2	Year	GE	S&P 500
3	ending	Y	X
4	19851231	0.3307	0.2633
5	19861231	0.2180	0.1462
6	19871231	0.0525	0.0203
7	19881230	0.0490	0.1240
8	19891229	0.4876	0.2725
9	19901231	-0.0817	-0.0656
10	19911231	0.3719	0.2631
11	19921231	0.1507	0.0446
12	19931231	0.2602	0.0705
13	19941230	0.0025	-0.0154
14	19951229	0.4512	0.3411
15	19961231	0.4035	0.2026
16	19971231	0.5094	0.3101
17	19981231	0.4097	0.2667
18	19991231	0.5354	0.1953
19	20001229	-0.0604	-0.1014
20	20011231	-0.1502	-0.1304
21	20021231	-0.3765	-0.2336
22	20031231	0.3074	0.2638
23	20041231	0.2071	0.0899

Regression

Input

Input Y Range: $B$4:$B$23

Input X Range: $C$4:$C$23

☐ Labels ☑ Constant is Zero

☐ Confidence Level: 95 %

Output options

● Output Range: $E$1

○ New Worksheet Ply:

○ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

The regressions results are as follows:

Summary Output

Regression Statistics	
Multiple *R*	0.9307
R Square	0.8662
Adjusted *R* Square	0.8136
Standard Error	0.0915
Observations	20

ANOVA

	df	SS	MS	F	Significance *F*
Regression	1	1.0304	1.0304	122.9939	0.0000
Residual	19	0.1592	0.0084		
Total	20	1.1895			

	Coefficients	Std. Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
X Variable 1	1.5411	0.1034	14.9079	0.0000	1.3248	1.7575

As the results show, the intercept term no longer appears. The estimated slope coefficient, 1.5411, is slightly greater than the slope estimated in Illustration B-6, 1.4298, since we have forced the regression line through 0, as shown in the following figure:

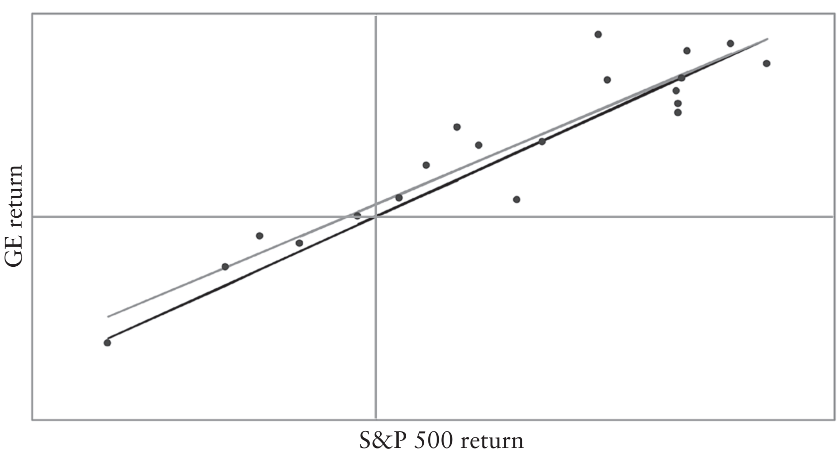


TABLE B.2 Summary of simple linear regression through the origin estimation formulas.

Estimator for slope:	$\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$	Standard error:	$s(\hat{\beta}_1) = \dfrac{s}{\sqrt{\sum_{i=1}^{n} X_i^2}}$
Prediction:	$\hat{Y}_p = \hat{\beta}_1 X_p$	Standard error:	$s_p = s\sqrt{1 + \dfrac{X_p^2}{\sum_{i=1}^{n} X_i^2}}$
Standard error of the estimate:	$s = \sqrt{\dfrac{\sum_{i=1}^{n} e_i^2}{n-1}}$		

MULTIPLE LINEAR REGRESSION

Analogous to simple linear regression models, we can fit a model using two or more explanatory variables of the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i \tag{B.18}$$

where k denotes the number of independent variables in the model, Y is the dependent variable, $X_1, \cdots, X_k$ are the independent variables, and $\beta_0, \beta_1, \cdots, \beta_k$ are the $k+1$ regression coefficients. This is referred to as a *multiple linear regression model* because the equation is linear in the parameters. Note that $X_2, \cdots, X_k$ could all be functions of X_1 such as X_1^2 or ln X_1 and this would still be considered a linear model. The idea of multiple regression is that one independent variable may not be enough to predict Y effectively, so additional variables are added to give more explanatory power to the model.

Model Assumptions

The assumptions for multiple linear regression are essentially the same as the five we used in simple regression model. First, the relation between X and Y is assumed to be linear except that now there are multiple Xs, as shown in (B.18). Second, the values of the Xs are nonrandom. In addition, we require that there is no exact linear relation among any of the independent variables. Finally, the error term ε is assumed to be independent and identically (normally) distributed with mean 0 and constant variance σ^2.

Estimation

Multiple linear regression models are fitted using ordinary least squares in a similar manner to their simple regression analogues. Again, the requirement is that the sum of squared errors (*SSE*) is minimized. Estimation of the parameters in the case $k > 1$ involves using linear algebra and matrix inversion, so the calculations are nearly impossible to perform by hand. Luckily, most statistical software packages have built-in estimation routines so these models can be fitted quite easily.

Hypothesis Tests for Individual Parameters

Once we have fit a multiple regression model, we may wish to find out whether a particular independent variable has a significant effect on Y. We can do this by testing the hypothesis that the corresponding parameter value is equal to zero. The procedure for these tests is almost identical to the case of simple regression models, except that, in a more general sense, the test statistic,

$$t = \hat{\beta}_i / s_{\hat{\beta}_i}$$

has a t-distribution with $n - k - 1$ degrees of freedom. Most regression software packages automatically provide the parameter estimates, standard errors, t-ratios, and p-values when a linear regression is performed.

ILLUSTRATION B.8 Test difference between two means using dummy variable.

*The worksheet **B8** of the Excel file, **B Illustrations.xls**, contains 60 months of returns for IBM during the period January 2000 through December 2004. In Illustration A.8 of Appendix A, we used this data to test the null hypothesis that the mean during the first 30 months is no different than the mean return in the second 60 months. Assuming the variances of the two samples are equal, usea dummy variable regression to perform the same statistical test.*

The test the difference between means using regression analysis, we must create a dummy (or binary) variable to use as an independent variable. The construction of a dummy variable is simple—it is set equal to either 0 or 1 depending on a particular criterion. In the current illustration, we set the dummy equal to 0 during the first 30 months and 1 during the second. We then regress the 60 months of IBM stock returns on the 60 dummy variable observations and find following results:

Summary Output

Regression Statistics	
Multiple *R*	0.10499
R-Square	0.01102
Adjusted *R*-Square	–0.00603
Standard Error	0.10410
Observations	60

ANOVA

	df	*SS*	*MS*	F	Significance *F*
Regression	1	0.00701	0.00701	0.64647	0.42466
Residual	58	0.62849	0.01084		
Total	59	0.63549			

	Coefficients	Std. Error	*t* Stat	*P*-value	Lower 95%	Upper 95%
Intercept	–0.00673	0.01901	–0.35385	0.72473	–0.04477	0.03132
X Variable 1	0.02161	0.02688	0.80403	0.42466	–0.03219	0.07541

To interpret these results, recognize that the dummy equals 0 in the first half of the sample. This means that the average value of the dummy variable in the first half of the sample must also equal 0, and that the average monthly return equals the intercept term, –0.00673, that is,

$$\begin{aligned}\overline{Y} &= \hat{\beta}_0 + \hat{\beta}_1\overline{X}\\ &= -0.00673 + 0.02161(0)\\ &= -0.00673\end{aligned}$$

During the second half of the sample, the average value of the dummy variable is 1. Consequently, the average monthly return during the second half is

$$\begin{aligned}\overline{Y} &= \hat{\beta}_0 + \hat{\beta}_1\overline{X}\\ &= -0.00673 + 0.02161(1)\\ &= 0.01488\end{aligned}$$

Note that these values conform exactly to the results shown in Illustration A.8.

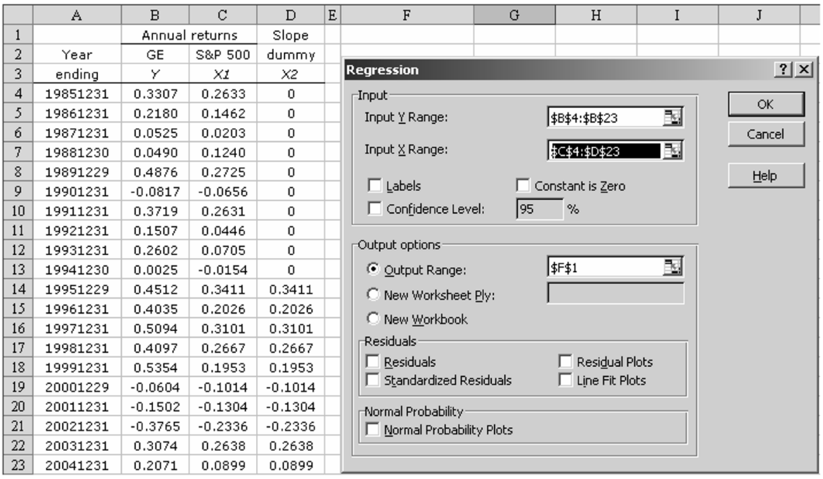
To test if there is a significant difference in the means of the two samples, we simply use the estimated slope coefficient and its standard error. To see this, recognize that the intercept applies to the returns over the entire 60-month sample and that the slope applies to only the second half of the sample. This means that the slope can be interpreted as the incremental mean return in the second half of the sample. Hence, the reported t-ratio for the slope tests the null hypothesis that there is a difference in the means of the two sub-periods. At $t = 0.804$, the p-value is 0.425 so we do not reject the null. Note that these values are, again, identical to those we computed in Illustration A.8.

ILLUSTRATION B.9 Test stationarity of return relation using dummy variable slope-shifter.

In the regression of GE's returns on the returns of the S&P 500, it is implicitly assumed that the intercept and slope coefficients are constant through time. Often in such time series, there is reason to believe that the relation has changed in some way during the sample period, and you want to test whether it has. Test the null hypothesis that the coefficient β_1 is the same during the first half of the sample than the second half of the sample. The return data are included in worksheet, ***B9****, of the Excel file,* ***B Illustrations.xls****.*

The worksheet contains the returns of GE and the S&P 500. It also contains an additional variable called a *dummy variable slope-shifter.* Note that this variable is 0 during the first half of the sample period, and is equal to the market return during the second half. When we run the regression of GE's returns on the two independent variables, the "beta" of GE during the first half of the sample is β_1 and the beta during the second half of the sample is $\beta_1 + \beta_2$. Thus, to test the hypotheses that the slope coefficient has changed, we perform a t-test on the slope coefficient β_2. If the coefficient is not different from 0 in the statistical sense, the null hypothesis that the relative systematic risk of GE has not changed from the first half of the period to the second cannot be rejected.

To perform the regression in Microsoft Excel, we follow the same steps as before(see the Regression dialog box illustrated below). The only distinction is that the in the Input X Range, we highlight both columns C and D, which contain the two independent variables in our regression. If we click OK, the multiple regression will be performed. Before turning to the results, however, the fact that the independent variables must be in adjacent columns is a limitation of the regression function in Excel. In multiple regression problems, it is often the case that the researcher has many more variables than is necessary. Having the flexibility to use, say, the second, fourth, and ninth columns of data as independent variables without editing the file would be useful. Excel demands that you rearrange the data so that the independent variables are in adjacent columns. For this reason, Excel is not frequently used in academic research or other large-scale applications. For our purposes, however, it is more than adequate.



	A	B	C	D
1		Annual returns		Slope
2	Year	GE	S&P 500	dummy
3	ending	*Y*	*X1*	*X2*
4	19851231	0.3307	0.2633	0
5	19861231	0.2180	0.1462	0
6	19871231	0.0525	0.0203	0
7	19881230	0.0490	0.1240	0
8	19891229	0.4876	0.2725	0
9	19901231	-0.0817	-0.0656	0
10	19911231	0.3719	0.2631	0
11	19921231	0.1507	0.0446	0
12	19931231	0.2602	0.0705	0
13	19941230	0.0025	-0.0154	0
14	19951229	0.4512	0.3411	0.3411
15	19961231	0.4035	0.2026	0.2026
16	19971231	0.5094	0.3101	0.3101
17	19981231	0.4097	0.2667	0.2667
18	19991231	0.5354	0.1953	0.1953
19	20001229	-0.0604	-0.1014	-0.1014
20	20011231	-0.1502	-0.1304	-0.1304
21	20021231	-0.3765	-0.2336	-0.2336
22	20031231	0.3074	0.2638	0.2638
23	20041231	0.2071	0.0899	0.0899

Regression

Input

Input Y Range: $B$4:$B$23

Input X Range: $C$4:$D$23

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

● Output Range: $F$1

○ New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

The regression results are reported below. Turning immediately to the slope coefficient for X_2, we see that the coefficient estimate is 0.1745, the t-ratio is 0.7935, and the p-value is 0.4384. Using a 5% significance level ($\alpha = 0.05$), the null hypothesis that $\beta_2 = 0$ cannot be rejected. In other words, there is no reason to believe the relation between GE's returns and the returns of the S&P 500 has changed through time, at least with respect to the two halves of the sample.

Summary Output

Regression Statistics	
Multiple *R*	0.9412
R Square	0.8859
Adjusted *R* Square	0.8725
Standard Error	0.0894
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	**Significance *F***
Regression	2	1.0538	0.5269	65.9832	0.0000
Residual	17	0.1357	0.0080		
Total	19	1.1895			

	Coefficients	**Std. Error**	***t* Stat**	***P*-value**	**Lower 95%**	**Upper 95%**
Intercept	0.0421	0.0254	1.6594	0.1154	–0.0114	0.0957
X Variable 1	1.2998	0.2060	6.3102	0.0000	0.8652	1.7344
X Variable 2	0.1745	0.2199	0.7935	0.4384	–0.2895	0.6386

Confidence Intervals

In a similar vein to hypothesis testing, we can also learn about the effects of individual explanatory variables by constructing confidence intervals. For each of the $k + 1$ parameters in the model, we can obtain interval estimates using the regression output from software packages in the following manner:

$$\text{A } (1 - \alpha)\% \text{ confidence interval for } \beta_i \text{ is: } \hat{\beta}_i \pm t_{\alpha/2} s_{\hat{\beta}_i}$$

To see whether the ith explanatory variable has an effect on Y, we check for the presence of zero in the confidence interval for β_i.

Hypothesis Tests for Significant Overall Regression

In addition to testing whether individual explanatory variables have an effect on Y, we may also want to check whether the model as a whole has significant predictive power. We can do this using an F-test, where the null hypothesis under question is that all coefficients are equal to zero: H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$. This is compared with the alternative hypothesis that at least one of the coefficients is nonzero. The test statistic here is denoted by F, which can be regarded as a *signal-to-noise* ratio. The signal refers to the portion of the variation explained by the model, and the noise relates to the part left unexplained. The F-test can also be understood by examining the following generic analysis of variance (ANOVA) table resulting from a linear regression procedure:

ANOVA Table

Source	Sum of Squares	Degrees of Freedom	Mean-Squared Error	F-ratio
Regression	SSR	k	SSR/k	$\dfrac{SSR/k}{SSE/(n-k-1)}$
Error	SSE	$n - k - 1$	$SSE/(n - k - 1)$	
Total	SST	$n - 1$	$SST/(n - 1)$	

An analysis of variance table like the one above is standard output from most software packages when a regression procedure is performed. The first column shows how the total variation of Y is decomposed: *Regression*, the part explained by the model; *Error*, the part unexplained by the model; and, *Total*, both parts put together. In the *Sum of Squares* column, SSR, SSE, and SST stand for regression, error, and total sum of squares, respectively. The *Degrees of Freedom* (*df*) are divided up as follows: total degrees of freedom are the number of observations minus one (which is lost because the overall mean, $\bar{Y}$, is estimated); degrees of freedom for the regression are equal to the number of explanatory variables in the model; and the difference between the two gives the degrees pf freedom of the error. The error *df* is also the degrees of freedom on which t-test for individual parameter significance is based. The *Mean-Squared Error* column gives essentially the average sum of squares for each source. Note that the mean-squared total is the unbiased estimate for the variance of Y. The F-statistic in column five can be thought of as the signal-to-noise ratio: the regression mean squared, SSR/k, being the signal, and the error mean squared, $SSE/(n - k - 1)$, as the noise. If the F-statistic gets very high, this means that the regression is explaining a large proportion of the variance, and, if it gets quite low, this indicates that a great deal is left unexplained by the model. Therefore we reject the null hypothesis of a useless model if F is large enough. And for any given level of significance, we can find the critical value of F using an F-distribution with k and $n - k - 1$ degrees of freedom. Critical values of the F-distribution are reported in Table C.4 of Appendix C.

Prediction

For any combination of independent variables lying in the range of the observed X's, we can obtain point estimates and predictions intervals by using a formula similar to the one given for the simple regression case. The only difference is that in the standard error, which depends on the distance from the prediction point, X_p and the mean, $\bar{X}$, we need to incorporate the fact that this should be measured in k-dimensional space. However, most statistical software packages provide this output, so again no matrix algebra is necessary.

Model Selection and Goodness of Fit

R-Squared and Adjusted R-Squared Just as for a simple regression, goodness-of-fit analysis for multiple regression is commonly based on the sum of squares decomposition. The coefficient of determination, R^2, is widely used because of its ease of interpretation. R^2 can be quite misleading, however, because it increases monotonically with k. In other words, as each additional independent variable is added to the model, the coefficient of determination must either increase or remain the same. This becomes a problem when extraneous variables are included in the model solely to boost the value of R^2, ignoring scientific or statistical indications that they do not belong. Parsimony, or economy of explanation, is of great value to an investigator because it greatly eases the burden of understanding and interpreting the model. Therefore, we commonly use a related statistic, the adjusted-R^2, to take into account the size of the model when assessing goodness of fit in a multiple regression.

The adjusted-R^2, denoted by $\bar{R}^2$, is given by the formula,

$$\bar{R}^2 = 1 - \frac{Var(e)}{Var(Y)} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1-(1-R^2)\left(\frac{n-1}{n-k-1}\right) \tag{B.19}$$

where k is the number of independent variables in the model. We can see that the adjusted-R^2 is always less than or equal to R^2. (It only equals R^2 when the model has only an intercept.) It is also possible for $\bar{R}^2$ to be negative. The advantage of the adjusted-R^2 is that it penalizes the model for including variables that do not provide information about Y. Note that, for two models with the same R^2 but different k, the $\bar{R}^2$ will be larger for the smaller model. This is intended to ensure that the issue of parsimony is addressed in the model selection process.

ILLUSTRATION B.10 Test purchasing power parity.

Purchasing power parity (PPP) is a simple arbitrage relation that says the price of a commodity or security in one country equals the price of the same commodity or security in another after adjusting for the rate of exchange in the currency. Suppose we are considering the price of a stock index in euros, S_{EURO}, and the price of the same index in USD, S_{USD}. The PPP relation is

$$S_{EUR,t} = S_{USD,t} \times S_{EURO/USD,t} \tag{a}$$

where $S_{EURO/USD}$ is the price in Euros for one USD. Since the equation is nonlinear, we cannot test it directly using OLS regression. PPP is, however, intrinsically linear. Taking the natural logarithm of both sides, we get

$$\ln(S_{EUR,t}) = \ln(S_{USD,t}) + \ln(S_{EUR/USD,t}) \qquad (b)$$

To avoid issues of nonstationarity (see Chapter 5), the logged PPP relation is usually tested in differenced form. Since the differenced in logged prices is a continuous return, for example,

$$R_{EUR,t} \equiv \ln(S_{EUR,t}/S_{EUR,t-1}) \equiv \ln(S_{EUR,t}) - \ln(S_{EUR,t-1}) \qquad (c)$$

we can test PPP by running the OLS regression,

$$R_{EUR,t} = \beta_0 + \beta_1 R_{USD,t} + \beta_2 R_{EUR/USD,t} + \varepsilon_t \qquad (d)$$

*The worksheet **B10** in the Excel file **B Illustrations.xls** contains the monthly returns of the DAX 30, the S&P 500, and the EUR/USD exchange rate over the period January 2000 through January 2006. The DAX 30 is a diversified portfolio of German stocks, and the S&P 500 is a diversified portfolio of U.S. stocks. Use the returns on these portfolios to proxy for the returns on equities in each country. Test the PPP relation.*

The regression results are as follows:

Summary Output	
Regression Statistics	
Multiple *R*	0.7855
R-Square	0.6170
Adjusted *R*-Square	0.6059
Standard Error	0.0452
Observations	72

ANOVA

	df	*SS*	*MS*	*F*	Significance *F*
Regression	2	0.22752	0.11376	55.57159	0.00000
Residual	69	0.14125	0.00205		
Total	71	0.36879			

	Coefficients	Std. Error	*t* Stat	*P*-value	Lower 95%	Upper 95%
Intercept	–0.00087	0.00535	–0.16196	0.87181	–0.01154	0.00981
X Variable 1	1.29608	0.12626	10.26480	0.00000	1.04419	1.54797
X Variable 2	0.53488	0.18234	2.93332	0.00455	0.17111	0.89864

These results are interesting in a number of respects. First, note that the adjusted *R*-square is 60.59%. This means that the S&P 500 and exchange rate returns explain 60.59% of the variance in the DAX 30 returns. Second, note that both slope coefficients are significantly different from 0 at the 5% level. Indeed, both coefficients are significant at the .5%. This means both regressors have a significant effect on the dependent variable. Third, we are reject the joint hypothesis that $\beta_1 = 0$ *and* $\beta_2 = 0$. The *F*-statistic is

55.57, and its p-value is less than 0.000005. Fourth, the intercept term is not significantly different from 0. This is not unexpected. Comparing the regression equation (d) with the logged PPP relation (b), it should be obvious that expected values of the coefficients are: $\beta_0 = 0$, and $\beta_1 = \beta_2 = 1$.

This last observation is somewhat discomforting. Although we have shown that the slope coefficients are different from 0, a test of PPP requires that we test the null hypothesis that the coefficients β_1 and β_2 are equal to 1. Performing t-tests on these hypothesis also rejects these hypotheses, so, technically, the PPP relation is rejected.

The most likely reason that we reject PPP is that the DAX 30 and S&P 500 stock indexes are not perfect substitutes. The DAX 30 consists of only 30 high market capitalization stocks and is not particularly well-diversified. The S&P 500, on the other hand, is well-diversified and accounts for more than 70% of the total market value of all stocks traded in the U.S. What the regression results do show, however, is that the rate of return on the DAX does systematically covary with U.S. stock return (its beta is 1.66) and with change in the EUR/USD exchange rate.

Specification Errors

Implicit in the specification of the multiple regression model (B.13) are the assumptions that we know the identity of all of the k relevant explanatory variables, and that their relation with the dependent variable is linear. But, since regression is by its nature exploratory, we need to be concerned about the "correctness" of our specification. What impact does failing to include a relevant explanatory variable have on estimation? Along the same line, what is the effect of including an explanatory variable that does not belong in the regression model? Finally, what is the effect of estimating a linear relation when the actual relation is nonlinear? We address each of these issues in turn.

Omitting Relevant Explanatory Variables Failing to include a relevant explanatory variable can have serious implications. To see this, assume that the "true" model is

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \tag{B.20}$$

where the variables are expressed as deviations from their means. Now suppose that, instead of estimating (B.20), we estimate

$$y_i = \beta_1^* x_{1i} + \varepsilon_i^* \tag{B.21}$$

What are the implications?

We know from our discussion of simple linear regression that the estimated slope in (B.21) is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{B.22}$$

Substituting (B.20) for y_i, we obtain

$$\begin{aligned}\hat{\beta}_1^* &= \frac{\beta_1 \sum\limits_{i=1}^{n} x_{1i}^2 + \beta_2 \sum\limits_{i=1}^{n} x_{1i}x_{2i} + \sum\limits_{i=1}^{n} x_{1i}\varepsilon_i}{\sum\limits_{i=1}^{n} x_{1i}^2} \\ &= \beta_1 + \frac{\beta_2 \sum\limits_{i=1}^{n} x_{1i}x_{2i}}{\sum\limits_{i=1}^{n} x_{1i}^2} + \frac{\sum\limits_{i=1}^{n} x_{1i}\varepsilon_i}{\sum\limits_{i=1}^{n} x_{1i}^2}\end{aligned} \tag{B.23}$$

Since X_1 is fixed and the expected value of the error is 0, the expected value of the estimated slope is

$$E\left(\hat{\beta}_1^*\right) = \beta_1 + \beta_2 \frac{\sum\limits_{i=1}^{n} x_{1i}x_{2i}}{\sum\limits_{i=1}^{n} x_{1i}^2} \tag{B.24}$$

This means that, in general, the estimated slope parameter in (B.21) will be biased. The direction of the bias depends on the product of β_2 and the covariance between the independent variables. If β_2 and the covariance are both positive, the estimated slope will be upward biased. The intuition for this is that the estimated slope picks up not only the co-variation of y_i with x_{1i} but also some of the co-variation of y_i with x_{2i}. Only in the event that the correlation between the independent variables is zero will the estimated slope be unbiased.

The standard error of the estimate will also be biased. In the case of estimating (B.21) when (B.20) is the correct model, the standard error of $\hat{\beta}_1^*$ will be less than the standard error of $\hat{\beta}_1$. We thereby run the risk of rejecting the null hypothesis that the slope parameter is 0 when in reality it is.

ILLUSTRATION B.11 Examine effect of omitted variable in regression specification.

In Illustration B.10, we estimated a multiple regression model with the return on the DAX 30 as the dependent variable and the returns of the S&P 500 and the EUR/USD exchange rate as the independent variables. Estimate the simple linear regression of DAX 30 returns on S&P 500 returns, and comment on the difference in results.

The regression results are as follows:

Summary Output	
Regression Statistics	
Multiple *R*	0.7545
R Square	0.5692
Adjusted *R* Square	0.5631
Standard Error	0.0476
Observations	72

ANOVA

	df	SS	MS	F	Significance *F*
Regression	1	0.20990	0.20990	92.49113	0.00000
Residual	70	0.15886	0.00227		
Total	71	0.36876			

	Coefficients	Std. Error	*t* Stat	*P*-value	Lower 95%	Upper 95%
Intercept	–0.00195	0.00562	–0.34756	0.72921	–0.01317	0.00926
X Variable 1	1.27685	0.13277	9.61723	0.00000	1.01205	1.54164

Comparing these results with those of Illustration B.10 shows at least two interesting facts. First, the adjusted *R*-square value falls from 60.59% to 56.31%. As noted earlier, the adjusted *R*-square value is often used as a criterion for choosing among competing models with the same dependent variable. Based on the results, the model in Illustration B.10 is preferred. Second, the estimated slope coefficient is now 1.27685 versus 1.29608 in Illustration B.10. This is the omitted variable bias just discussed. Equation (B.24) shows the nature of the bias. Since we know β_2 is positive, the bias in (B.24) depends on the sign of the covariance term in the numerator of the last term on the right hand side. Since the estimated coefficient falls in value from 1.29608 to 1.27685 (i.e., the bias is negative), we can deduce that the correlation between the returns of the S&P 500 and the EUR/USD exchange rate is negative. Indeed, if we compute the correlation matrix for the three return series, we find that:

	DAX30	S&P 500	EUR/USD
DAX30	1		
S&P 500	0.7545	1	
EUR/USD	0.1791	–0.0519	1

The correlation between the returns of the S&P 500 and the EUR/USD is negative, as expected. Its level is only –0.0519, thus the degree of bias is modest. The higher the correlation, the greater the bias. The only circumstance in which no bias occurs is if the correlation is 0—a highly unlikely event.

Including Irrelevant Explanatory Variables The effects of including an irrelevant variable are much less serious. The estimated parameters of the relevant explanatory variables remain unbiased. The only cost, so to speak, is that the standard errors of the estimates will be larger than they should be, making it more difficult to reject the null hypothesis of a zero slope parameter. Thus, if you reject the null in the presence of irrelevant variables, you can be quite confident of your decision.

Nonlinearities A separate discussion of the effects of nonlinearity is unwarranted. Fitting a linear regression to a nonlinear relation is a special case of the omitted variables discussed above. We can expect bias in the estimated coefficients, and the standard errors to be smaller than appropriate.

Multicollinearity

Multicollinearity arises when two or more variables are highly correlated with each other. While it is possible to obtain least squares estimates of the regression coefficients, the interpretation of the coefficients is difficult. Recall that the interpretation of a regression coefficient is the change in Y with respect to a change in X_1, *holding other factors constant*. The presence of multicollinearity means that other factors are not being held constant. If X_1 and X_2 are highly correlated, a change in X_1 implies a change in X_2, and vice versa.

A rule of thumb states that multicollinearity is likely to be a problem if a simple correlation between the two variables is larger than the correlation of either or both variables with the dependent variable. Such a rule may be reasonable when there are only two independent variables in the regression. With more than two independent variables, however, simple correlations will not detect a more complicated linear relation among variables. Perhaps the simplest way to detect if multicollinearity is a problem is to examine the standard errors of the coefficients. If several coefficients have high standard errors, and dropping one or more variables from the equation lowers the standard errors of the remaining variables, multicollinearity will usually be the source of the problem.

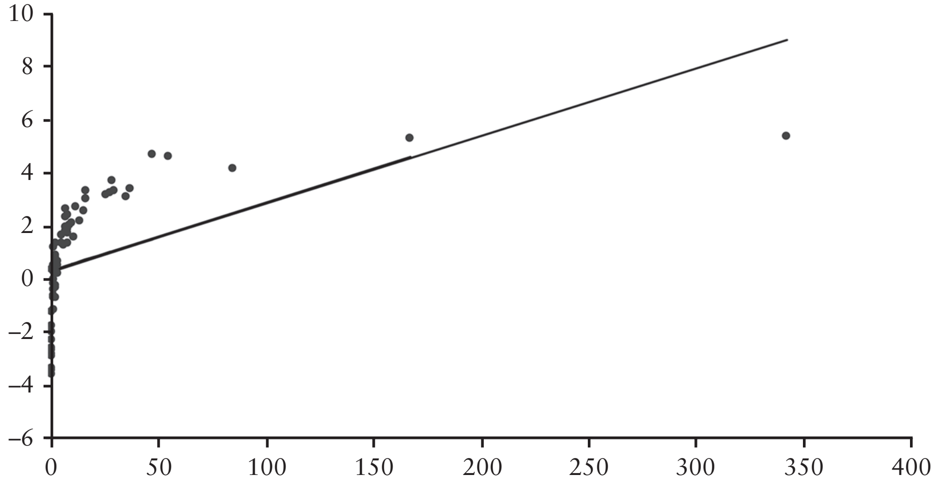
Violations of Disturbance Assumption

OLS regression assumes that the relation between dependent and independent variables is linear and that the error term ε is independent and identically (normally) distributed with mean 0 and constant variance σ^2. These assumptions can be violated in four ways: (1) the relation is nonlinear, (2) the error variance may not be constant, (3) the residual errors may be correlated, and (4) the errors may not be normally distributed. Below we discuss how to detect such violations, explain the consequences of each violation, and suggest remedies to fix or, at least mitigate, the effects of the violation.

Nonlinearity Plotting the relation between the Y and X variables is a useful first step in regression analysis. Among other things, it allows us to uncover potential nonlinearities in the data. To illustrate, consider the (X, Y) points plotted in the figure that follows. The solid line in the figure is the "best fit" obtained from the simple regression of Y on X, that is,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Obviously, the relation between Y and X is not linear. Where the level of X is near 0, the level of Y tends to be below its predicted level. For levels of X between 0 and 150, the levels of Y are above predicted, and, for levels of X above 200, the levels are below predicted. If the model was "correct," the (X,Y) points should be symmetrically distributed around the fitted line. Nonlinearity is usually revealed through a "bowed" pattern of residuals such as seen below (i.e., the model makes systematic errors whenever it is making unusually large or small predictions).

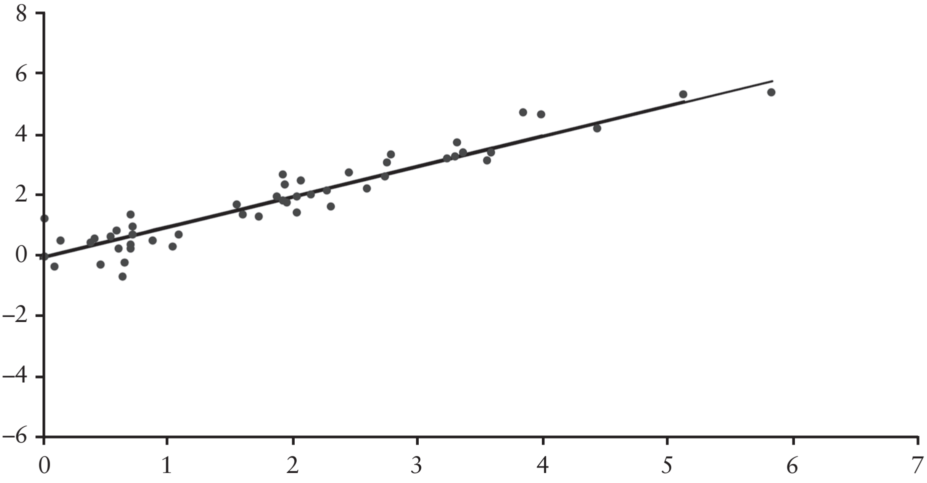


To remedy the problem, we may want to consider applying a *nonlinear transformation* to the dependent and/or independent variables. In the figure above, note that all of the values of X are positive and that the pattern of points looks like a log transformation. We may therefore want to consider applying a log transformation to the X variable. Another possibility to consider is adding another regressor which is a nonlinear function of one of the other variables. Since we have regressed Y on X, we may want to regress Y on both X and X^2. Note that, unlike the log transformation, this transformation can be applied even when X and/or Y have negative values.

As it turns out, the relation in the above figure is intrinsically linear. The figure below shows () points as well as the regression line,

$$Y_i = \beta_0 + \beta_1 \ln X_i + \varepsilon_i$$

fitted through the pairs of coordinates. Note that the residuals are now symmetrically distributed around the fitted line. This gives us comfort that we have uncovered the "correct" specification.[4]



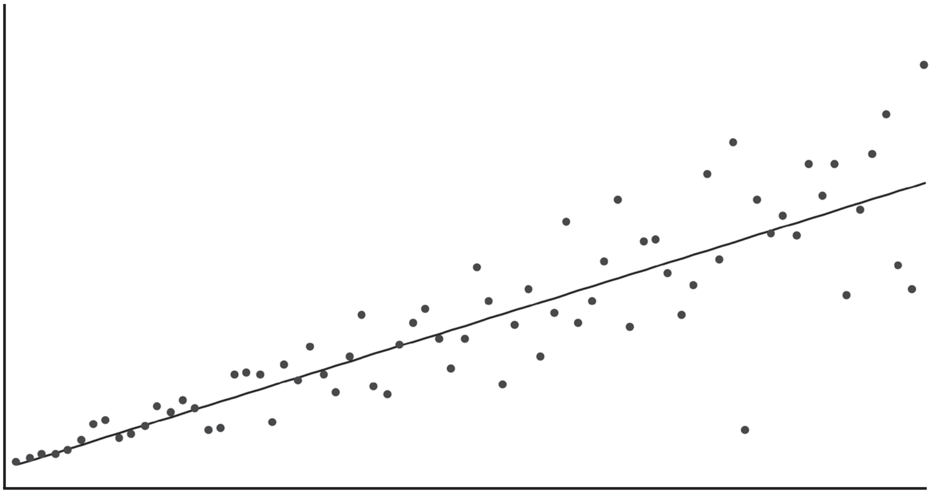
Heteroscedasticity Detecting a violation of the constant variance (or homoscedastic) error assumption is also facilitated by a plot of the residuals around the fitted values of Y. If the residuals have a constant variance at different levels of prediction, the error term is homoscedastic. If the residuals appear fan-shaped, as shown in the figure on the next page, heteroscedasticity may be a problem. A popular test for the presence of heteroscedasticity is the Goldfeld-Quandt (1965) test. The steps are as follows:

1. Order the data by the X_i observations.
2. Omit the c central observations.[5]
3. Fit separate regressions to the first $(n - c)/2$ and the last $(n - c)/2$ observations. Naturally, $(n - c)/2$ must exceed the number of parameters to be estimated.
4. Compute the ratio $R = SSE_2/SSE_1$, where SSE_1 and SSE_2 are the sum of squared errors from the first and second regressions, respectively.[6] Under the assumption of homoscedasticity, R has an F distribution with $(n - c)/2$ and $(n - c)/2$ degrees of freedom in the numerator and the denominator, respectively. If R exceeds the critical value reported in Table C.4 of Appendix C, we reject the null hypothesis that the error variance is the same in both subsamples.

[4] In financial economics, model specification is usually driven by theoretical considerations.

[5] The power of the test depends on the choice of c. The greater the value of c, the lower the power of the test. On the other hand, the lower the value of c, the greater the power, however, the more likely the residual variances will move closer together.

[6] As discussed in Appendix A, we place the largest SSE in the numerator. That is, the notation implicitly assumes $SSE_2 > SSE_1$.



With a heteroscedastic error term, ordinary least squares estimation places greater weight on observations with large error variances than on those with small error variances. This implicit weighting occurs because the sum of the squared residuals associated with the large variance error terms is likely to be substantially greater than the sum of the squared residuals associated with small error variances. This means that, while the parameter estimates remain unbiased, the standard errors of the parameter estimates will be biased.

Correcting for heteroscedasticity is possible using *weighted least squares* estimation. To illustrate, suppose that the error variances in the regression,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i \tag{B.25}$$

vary directly with one of the explanatory variables, say, $X_{i,1}$, that is,

$$Var(\varepsilon_i) = CX_{1i}^2$$

where C is a nonzero constant. To correct for heteroscedasticity, we multiply the terms of the regression by the inverse of X_{1i}, and run the regression,

$$\frac{Y_i}{X_{1i}} = \beta_0 \frac{1}{X_{1i}} + \beta_1 + \cdots + \beta_k \frac{X_{ki}}{X_{1i}} + \frac{\varepsilon_i}{X_{1i}} \tag{B.26}$$

The error term in the transformed regression model,

$$\varepsilon_i^* = \frac{\varepsilon_i}{X_{1i}}$$

now has constant variance, that is,

$$Var(\varepsilon_i^*) = Var\left(\frac{\varepsilon_i}{X_{1i}}\right) = \frac{1}{X_{1i}}Var(\varepsilon_i) = C \tag{B.27}$$

Uncovering the form of the heteroscedasticity is sometimes difficult since the error variance may be a nonlinear function of one of the independent variables, or it may be a function of some other variables, Z, not included in the regression model. Standard econometric textbooks offer guidance on appropriate correction procedures.[7] After the structure of the error variance is determined, however, variable transformations in a weighted least squares framework rectifies the problem.

Serial Correlation Violations of the error independence assumption are most often found in time series data.[8] A common method for their detection is the Durbin-Watson (1951) test. The DW-statistic looks at the sum of squared differences between subsequent residuals to see if they are, on average, too close together or too far apart. The DW-statistic tests the null hypothesis of no autocorrelation among the dependent variable against the alternative of autocorrelated data. The test statistic is

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{B.28}$$

where e_i is the ith time-ordered residual from the model. Note that the numerator will be small when subsequent errors are similar (positively correlated) and will be large when they tend to be far apart (negatively correlated). The level of the DW-statistic is approximately $2(1 - \rho)$, where ρ is the first-order autocorrelation of the residuals. Thus, a DW-statistic close to 2 indicates the residuals are uncorrelated.

In using the Durbin-Watson test of the null hypothesis that there is no serial correlation in the residuals, we must use a table of critical values such as those reported in Table C.5 of Appendix C for the five percent significance level. The rules for applying these values are as follows. If we are checking for positive autocorrelation, the null hypothesis is rejected if $DW < d_l$ and is accepted if $DW > d_u$. Between d_l and d_u the results are inconclusive. For a simple linear regression ($k = 2$) using 60 observations ($n = 60$), the critical values are $d_l = 1.51$ and $d_u = 1.65$. Thus, if $DW < 1.51$, we reject the null hypothesis of no serial correla-

[7] See, for example, Pindyck and Rubinfeld (1998, pp. 148–159).

[8] Any serial correlation in the errors of a cross-sectional regression can be eliminated by shuffling the order of the data.

tion in favor of the alternative hypothesis that there exists positive autocorrelation, and, if $DW > 1.65$, we accept the null. If $1.51 < DW < 1.65$, we cannot say one way or the other. If we are checking for negative autocorrelation, we view matters from an endpoint of 4 rather than an endpoint of 0. That is, the null hypothesis is rejected if $DW > 4 - d_l$ and is accepted if $DW < 4 - d_u$. Between $4 - d_u$ and $4 - d_l$ the results are inconclusive. For a simple linear regression ($k = 2$) using 60 observations ($n = 60$), the critical values are 2.35 and 2.49. If $DW > 2.49$, we reject the null hypothesis of no serial correlation in favor of the alternative hypothesis that there exists negative autocorrelation, and, if $DW < 2.35$, we accept the null. The results are inconclusive where $2.35 < DW < 2.49$.

The presence of serial correlation does not bias the parameter estimates. It does, however, affect the standard errors of the estimates. In the presence of positive serial correlation, the standard errors will be smaller than they should be, potentially causing us to reject the null when we should not. To understand possible correction procedures, consider the nature of the problem. Under the assumption the error term is serially correlated, the regression model is

$$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_k X_{kt} + \varepsilon_t \tag{B.29}$$

where

$$\varepsilon_t = \rho\varepsilon_t + \upsilon_t \tag{B.30}$$

ρ is the first-order serial-correlation, and υ_t is normally distributed with zero mean and constant variance and is independent of other errors through time. Since equation (B.29) holds for all time periods, we can write

$$Y_{t-1} = \beta_0 + \beta_1 X_{1t-1} + \cdots + \beta_k X_{kt-1} + \varepsilon_{t-1} \tag{B.31}$$

Multiplying (B.31) by ρ and subtracting it from (B.29), we get

$$Y_t^* = \beta_0(1-\rho) + \beta_1 X_{1t}^* + \cdots + \beta_k X_{kt}^* + \nu_t \tag{B.32}$$

where the asterisks denote *generalized differences*. The variable X_{1t}^*, for example, is defined as $X_{1t}^* = X_{1t} - \rho X_{1t-1}$. Since the error term in (B.32) is independent through time, the standard errors of the regression model (B.32) will be unbiased.

To implement the correction procedure, we need to estimate ρ. One simple procedure, called the Hildreth-Lu (1960) procedure, is to set ρ to a grid of different values between 0 and 1 (e.g., 0, 0.1, 0.2, ..., 1) and estimate (B.32) for each assumed value. Based upon the regression results, we choose the value of ρ' that produces the lowest sum of squared errors, and then set up a new, more refined grid that searches in the neighborhood of ρ' to find a new value that minimizes the sum of squared errors. The procedure is repeated until the desired degree of accuracy is attained. Another approach, called the Cochrane-Orcutt (1949), is to estimate ρ from the residuals of (B.29), that is,

$$e_t = \rho e_{t-1} + \upsilon_t \tag{B.33}$$

and then use the estimated serial correlation in the estimation of the generalized difference model (B.32). Using the estimated parameters from (B.32), we generate a new set of residuals from the original regression equation (B.29), re-estimate (B.33) to obtain a new estimate of ρ, and then reestimate (B.32). The procedure is repeated iteratively until the new estimates of ρ differ from the old ones by, say, 0.005 or less, or after 10 to 20 iterations.

Nonnormality A violation of the normality assumption is particularly serious. The reason is simple. Since parameter estimation is based on the minimization of the sum of *squared errors*, a few extreme observations can exert a disproportionate influence on the parameter estimates and their standard errors. One way to test for normally distributed errors is to use a *normal probability plot* of the residuals. A normal probability plot is a plot of the fractiles of error distribution versus the fractiles of a normal distribution having the same mean and variance. If the distribution is normal, the points on this plot should fall close to the diagonal line. A "bow-shaped" pattern of deviations from the diagonal indicates that the residuals have excessive skewness (i.e., they are not symmetrically distributed, with too many large errors in the same direction). An "S-shaped" pattern of deviations indicates that the residuals have excessive kurtosis (i.e., there are either two many or two few large errors in both directions).

Violations of normality often arise either because (1) the distributions of the dependent and/or independent variables are nonnormal, and/or (2) the linearity assumption is violated. In such cases, a nonlinear transformation of variables might cure both problems. In some cases, the problem with the residual distribution is mainly due to one or two very large errors called *outliers*. Outliers should be scrutinized closely. If they are merely errors or if they can be explained as unique events not likely to be repeated, you may have cause to remove them.

REFERENCES AND SUGGESTED READINGS

Cochrane, D., and G. H. Orcutt. 1949. Application of least-squares regressions to relationships containing autocorrelated error terms. *Journal of teh American Statistical Association* 44: 32–61.

Durbin, J., and G. S. Watson. 1951. Testing for serial correlation in least squares regression. *Biometrika* 38: 159–177.

Goldfeld, S. M., and R. E. Quandt. 1965. Some tests for homoscedasticity. *Journal of the American Statistical Society* 60: 539–547.

Hildreth, G., and J. Y. Lu. 1960. Demand relations with autocorrelated disturbances. *Michigan State University Agricultural Experiment Station Technical Bulletin* 276.

Intriligator, Michael D. 1978. *Econometric Techniques, and Applications*. Englewood Cliffs, NJ: Prenctice-Hall.

Jarque, C. M., and A. K. Bera. 1987. A test of normality of observations and regression residuals. *International Statistical Review* 11: 351–360.

Jarque, C. M., and A. K. Bera. 1980. Efficient tests of normality, homoscedasticity and serial dependence of regression residuals. *Economic Letters* 6: 255–259.

Kennedy, Peter. 1992. *A Guide to Econometrics*, 3rd ed. Cambridge, MA: MIT Press.

Kmenta, Jan. 1971. *Elements of Econometrics*. New York: Macmillan Publishing Co.

Pindyck, Robert S., and Daniel L. Rubinfeld. 1998. *Econometric Models and Economic Forecasts*, 4th ed. Boston: Irwin/McGraw-Hill.