

CORRELATION AND REGRESSION

LOS 11a: Calculate and interpret a sample covariance and a sample correlation coefficient, and interpret a scatter plot. Vol 1, pg 280 - 285

Two of the most popular methods for examining how two sets of data are related are scatter plots and correlation analysis.

Scatter Plots

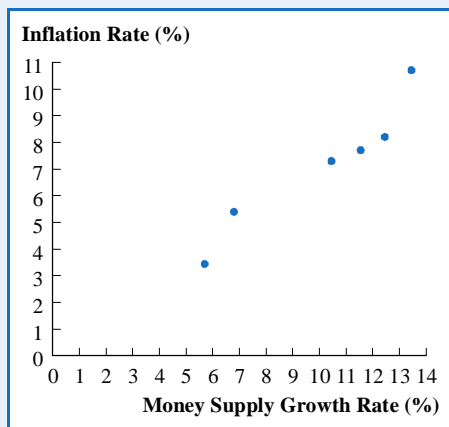
A **scatter plot** is a graph that illustrates the relationship between observations of two data series in two dimensions.

Example 1: Scatter Plot

The following table lists annual observations of money supply growth and inflation for 6 countries over the period 1990-2010. Illustrate the data on a scatter plot and comment on the relationship.

Country	Money Supply Growth Rate (X_i)	Inflation Rate (Y_i)
A	0.0685	0.0545
B	0.1160	0.0776
C	0.0575	0.0349
D	0.1050	0.0735
E	0.1250	0.0825
F	0.1350	0.1076

Figure 1: Scatter Plot



Note that each observation in the scatter plot is represented as a point, and the points are not connected. The scatter plot does not show which point relates to which country; it just plots the observations of both data series as pairs. The data plotted in Figure 1 suggests a fairly strong linear relationship with a positive slope.

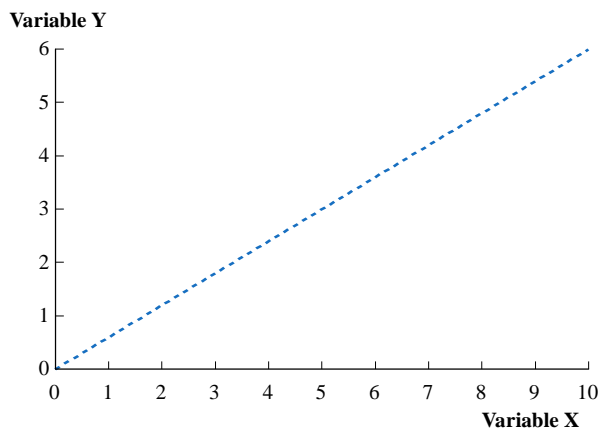
Correlation Analysis

Correlation analysis expresses the relationship between two data series in a single number. The **correlation coefficient** measures how closely two data series are related. More formally, it measures the strength and direction of the linear relationship between two random variables. The correlation coefficient can have a maximum value of +1 and a minimum value of -1.

- A correlation coefficient greater than 0 means that when one variable increases (decreases) the other tends to increase (decrease) as well.
- A correlation coefficient less than 0 means that when one variable increases (decreases) the other tends to decrease (increase).
- A correlation coefficient of 0 indicates that no linear relation exists between the two variables.

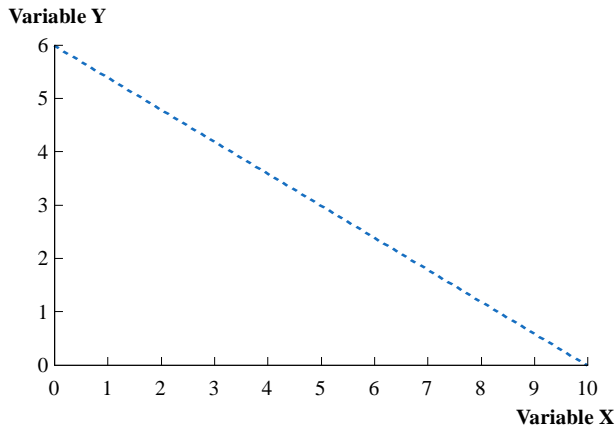
Figures 2, 3 and 4 illustrate the scatter plots for data sets with different correlations.

Figure 2: Scatter Plot of Variables with Correlation of +1

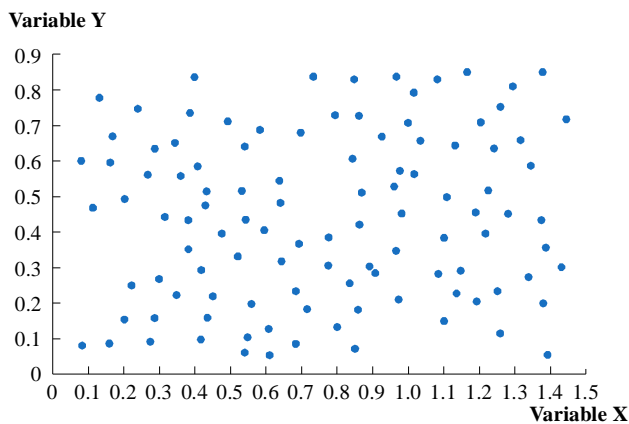


Analysis:

- Note that all the points on the scatter plot illustrating the relationship between the two variables lie along a straight line.
- The slope (gradient) of the line equals +0.6, which means that whenever the independent variable (X) increases by 1 unit, the dependent variable (Y) *increases* by 0.6 units.
- If the slope of the line (on which all the data points lie) were different (from +0.6), but positive, the correlation between the two variables would equal +1 as long as the points lie on a straight line.

Figure 3: Scatter Plot of Variables with Correlation of -1**Analysis:**

- Note that all the points on the scatter plot illustrating the relationship between the two variables lie along a straight line.
- The slope (gradient) of the line equals -0.6 , which means that whenever the independent variable (X) increases by 1 unit, the dependent variable (Y) *decreases* by 0.6 units.
- If the slope of the line (on which all the data points lie) were different (from -0.6) but negative, the correlation between the two variables would equal -1 as long as all the points lie on a straight line.

Figure 4: Scatter Plot of Variables with Correlation of 0**Analysis:**

- Note that the two variables exhibit no linear relation.
- The value of the independent variable (X) tells us nothing about the value of the dependent variable (Y).

Calculating and Interpreting the Correlation Coefficient

In order to calculate the correlation coefficient, we first need to calculate **sample covariance**. Covariance is a similar concept to variance. The difference lies in the fact that variance measures how a random variable varies with itself, while covariance measures how a random variable varies with another random variable.

Properties of Covariance

- Covariance is symmetric i.e., $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.
- The covariance of X with itself, $\text{Cov}(X, X)$, equals the variance of X , $\text{Var}(X)$.

Interpreting the Covariance

- Basically, covariance measures the nature of the relationship between two variables.
- When the covariance between two variables is *negative*, it means that they tend to move in opposite directions.
- When the covariance between two variables is *positive*, it means that they tend to move in the same direction.
- The covariance between two variables equals zero if they are not related.

Sample covariance is calculated as:

$$\text{Sample covariance} = \text{Cov}(X, Y) = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) / (n - 1)$$

where:

n = sample size

X_i = i th observation of Variable X

\bar{X} = mean observation of Variable X

Y_i = i th observation of Variable Y

\bar{Y} = mean observation of Variable Y

The numerical value of sample covariance is not very meaningful as it is presented in terms of units squared, and can range from negative infinity to positive infinity. To circumvent these problems, the covariance is standardized by dividing it by the product of the standard deviations of the two variables. This standardized measure is known as the sample **correlation coefficient** (denoted by r) and is easy to interpret as it always lies between -1 and $+1$, and has no unit of measurement attached.

$$\text{Sample correlation coefficient} = r = \frac{\text{Cov}(X, Y)}{s_X s_Y}$$

$$\text{Sample variance} = s_X^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$$

$$\text{Sample standard deviation} = s_X = \sqrt{s_X^2}$$

Example 2: Calculating the Correlation Coefficient

Using the money supply growth and inflation data from 1990-2010 for the 6 countries in Example 1, calculate the covariance and the correlation coefficient.

Solution:

Country	Money Supply Growth Rate (X_i)	Inflation Rate (Y_i)	Cross Product ($X_i - \bar{X})(Y_i - \bar{Y})$)	Squared Deviations ($X_i - \bar{X})^2$)	Squared Deviations ($Y_i - \bar{Y})^2$)
A	0.0685	0.0545	0.000564	0.001067	0.000298
B	0.116	0.0776	0.000087	0.00022	0.000034
C	0.0575	0.0349	0.00161	0.001907	0.001359
D	0.105	0.0735	0.000007	0.000015	0.000003
E	0.125	0.0825	0.000256	0.000568	0.000115
F	0.135	0.1076	0.001212	0.001145	0.001284
Sum	0.607	0.4306	0.003735	0.004921	0.003094
Average	0.1012	0.0718			
Covariance			0.000747		
Variance				0.000984	0.000619
Std. Dev (s)				0.031373	0.024874

Illustrations of Calculations

Covariance = Sum of cross products / $n - 1 = 0.003735/5 = 0.000747$

Var (X) = Sum of squared deviations from the sample mean / $n - 1 = 0.004921/5 = 0.000984$

Var (Y) = Sum of squared deviations from the sample mean / $n - 1 = 0.003094/5 = 0.000619$

Correlation coefficient = $r = \frac{\text{Cov}(X,Y)}{s_X s_Y} = \frac{0.000747}{(0.031373)(0.024874)} = 0.9573$ or 95.73%

The correlation coefficient of 0.9573 suggests that over the period, a strong linear relationship exists between the money supply growth rate and the inflation rate for the countries in the sample.

Note that computed correlation coefficients are only valid if the means and variances of X and Y, as well as the covariance of X and Y, are finite and constant.

LOS 11b: Explain the limitations to correlation analysis, including outliers and spurious correlation. Vol 1, pg 285 - 288

Limitations of Correlation Analysis

- It is important to remember that the correlation is a measure of **linear** association. Two variables can be connected through a very strong non-linear relation and still exhibit low correlation. For example, the equation $Y = 10 + 3X$ represents a linear relationship. However, two variables may be perfectly linked by a nonlinear equation, for example, $Y = (5+X)^2$ but their correlation coefficient may still be close to 0.
- Correlation may be an unreliable measure when there are **outliers** in the data. Outliers are a small number of observations that are markedly numerically different from the rest of the observations in the sample. Analysts must evaluate whether outliers represent relevant information about the association between the variables (news) and therefore, should be included in the analysis, or whether they do not contain information relevant to the analysis (noise) and should be excluded.
- Correlation does not imply causation. Even if two variables exhibit high correlation, it does not mean that certain values of one variable bring about the occurrence of certain values of the other.
- Correlations may be spurious in that they may highlight relationships that are misleading. For example, a study may highlight a statistically significant relationship between the number of snowy days in December and stock market performance. This relationship obviously has no economic explanation. The term **spurious correlation** is used to refer to relationships where:
 - Correlation reflects chance relationships in a data set.
 - Correlation is induced by a calculation that mixes the two variables with a third.
 - Correlation between two variables arises from both the variables being directly related to a third variable.

LOS 11c: Formulate a test of the hypothesis that the population correlation coefficient equals zero, and determine whether the hypothesis is rejected at a given level of significance. Vol 1, pg 296 - 300

Testing the Significance of the Correlation Coefficient

Hypothesis tests allow us to evaluate whether apparent relationships between variables are caused by chance. If the relationship is not the result of chance, the parameters of the relationship can be used to make predictions about one variable based on the other. Let's go back to Example 2, where we calculated that the correlation coefficient between the money supply growth rate and inflation rate was 0.9573. This number seems pretty high, but is it statistically different from zero?

To test whether the correlation between two variables is significantly different from zero the hypotheses are structured as follows:

$$\begin{aligned} H_0: \rho &= 0 \\ H_a: \rho &\neq 0 \end{aligned}$$

Note: This would be a two-tailed t-test with $n-2$ degrees of freedom.

The test statistic is calculated as:

$$\text{Test-stat} = t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Where:

n = Number of observations

r = Sample correlation

The decision rule for the test is that we reject H_0 if $t\text{-stat} > +t_{\text{crit}}$ or if $t\text{-stat} < -t_{\text{crit}}$

From the expression for the test-statistic above, notice that the value of sample correlation, r , required to reject the null hypothesis, decreases as sample size, n , increases:

- As n increases, the degrees of freedom also increase, which results in the absolute critical value for the test (t_{crit}) falling and the rejection region for the hypothesis test increasing in size.
- The absolute value of the numerator (in calculating the test statistic) increases with higher values of n , which results in higher t-values. This increases the likelihood of the test statistic exceeding the absolute value of t_{crit} and therefore, increases the chances of rejecting the null hypothesis.

In order to use the t-test, we assume that the two populations are normally distributed.

ρ represents the population correlation.

Example 3: Testing the Correlation between Money Supply Growth and Inflation

Based on the data provided in Example 1, we determined that the correlation coefficient between money supply growth and inflation during the period 1990-2010 for the six countries studied was 0.9573. Test the null hypothesis that the true population correlation coefficient equals 0 at the 5% significant level.

Solution:

$$\text{Test statistic} = \frac{0.9573 \times \sqrt{6-2}}{\sqrt{1-0.9573^2}} = 6.623$$

Degrees of freedom = $6 - 2 = 4$

The critical t-values for a two-tailed test at the 5% significance level (2.5% in each tail) and 4 degrees of freedom are -2.776 and +2.776.

Since the test statistic (6.623) is greater than the upper critical value (+2.776) we can reject the null hypothesis of no correlation at the 5% significance level.

From the additional examples in the CFA Program Curriculum (Examples 8 and 9 on Page 298) you should understand the takeaways listed below. If you understand the math behind the computation of the test statistic, and the determination of the rejection region for hypothesis tests, you should be able to digest the following points quite comfortably:

- All other factors constant, a false null hypothesis ($H_0: \rho = 0$) is more likely to be rejected as we increase the sample size (as the size of the rejection region increases due to lower and lower absolute values of t_{crit}).
- The smaller the size of the sample, the greater the value of sample correlation required to reject the null hypothesis of zero correlation (in order to make the value of the test statistic sufficiently large so that it exceeds the absolute value of t_{crit} at the given level of significance).
- When the relation between two variables is very strong, a false null hypothesis ($H_0: \rho = 0$) may be rejected with a relatively small sample size (as r would be sufficiently large to push the test-statistic beyond the absolute value of t_{crit}). Note that this is the case in Example 3.
- With large sample sizes, even relatively small correlation coefficients can be significantly different from zero (as a high value of n increases the test statistic and reduces the size of the rejection region for the hypothesis test).

Uses of Correlation Analysis

Correlation analysis is used for in:

- Investment analysis (e.g. evaluating the accuracy of inflation forecasts in order to apply the forecasts in predicting asset prices).
- Identifying appropriate benchmarks in the evaluation of portfolio manager performance.
- Identifying appropriate avenues for effective diversification of investment portfolios.
- Evaluating the appropriateness of using other measures (e.g. net income) as proxies for cash flow in financial statement analysis.

LOS 11d: Distinguish between the dependent and independent variables in a linear regression. Vol 1, pg 300 - 303

Linear Regression with One Independent Variable

Linear regression is used to summarize the relationship between two variables that are linearly related. It is used to make predictions about a **dependent variable**, Y (also known as the **explained variable**, **endogenous variable** and **predicted variable**) using an **independent variable**, X (also known as the **explanatory variable**, **exogenous variable** and **predicting variable**), to test hypotheses regarding the relation between the two variables, and to evaluate the strength of the relationship between them. The dependent variable is the variable whose variation we are seeking to explain, while the independent variable is the variable that is used to explain the variation in the dependent variable. The following linear regression model describes the relation between the dependent and the independent variables.

$$\text{Regression model equation} = Y_i = b_0 + b_1X_i + \varepsilon_i, i = 1, \dots, n$$

- b_1 and b_0 are the regression coefficients.
- b_1 is the slope coefficient.
- b_0 is the intercept term.
- ε is the error term that represents the variation in the dependent variable that is not explained by the independent variable.

Based on this model, the regression process estimates the line of best fit for the data in the sample. The regression line takes the following form:

$$\text{Regression line equation} = \hat{Y}_i = \hat{b}_0 + \hat{b}_1X_i, i = 1, \dots, n$$

Linear regression computes the line of best fit that minimizes the sum of the **regression residuals** (the squared vertical distances between actual observations of the dependent variable and the regression line). What this means is that it looks to obtain estimates, \hat{b}_0 and \hat{b}_1 , for b_0 and b_1 respectively, that minimize the sum of the squared differences between the actual values of Y, Y_i , and the predicted values of Y, \hat{Y}_i , according to the regression equation ($\hat{Y}_i = \hat{b}_0 + \hat{b}_1X_i$). Therefore, linear regression looks to minimize the expression:

$$\sum_{i=1}^n [Y_i - (\hat{b}_0 + \hat{b}_1X_i)]^2$$

where:

Y_i = Actual value of the dependent variable

$\hat{b}_0 + \hat{b}_1X_i$ = Predicted value of dependent variable

Another way to look at simple linear regression is that it aims to explain the variation in the dependent variable in terms of the variation in the independent variable. Note that variation refers to the extent that a variable deviates from its mean value. Do not confuse variation with variance.

Hats over the symbols for regression coefficients indicate estimated values. Note that it is these estimates that are used to conduct hypothesis tests and to make predictions about the dependent variable.

The sum of the squared differences between actual and predicted values of Y is known as the **sum of squared errors**, or **SSE**).

LOS 11e: Explain the assumptions underlying linear regression, and interpret the regression coefficients. Vol 1, pg 303 - 306

Assumptions of the Linear Regression Model

The following six assumptions are known as the **classic normal linear regression assumptions**:

1. The relationship between the dependent (Y) and the independent variable (X) is linear in the parameters, b_1 and b_0 . This means that b_1 and b_0 are raised to the first power only and neither of them is multiplied or divided by another regression parameter.
2. The independent variable, X, is not random.
3. The expected value of the error term is zero: $E(\varepsilon) = 0$
4. The variance of the error term is constant for all observations ($E(\varepsilon_i^2) = \sigma_\varepsilon^2, i = 1, \dots, n$). This is known as the homoskedasticity assumption.
5. The error term is uncorrelated across observations.
6. The error term is normally distributed.

Example 4: Linear Regression

For the money supply growth and inflation rate data that we have been working with in this reading, determine the slope coefficient and the intercept term of a simple linear regression using money supply growth as the independent variable and the inflation rate as the dependent variable. The data provided below is excerpted from Example 2:

Country	Money Supply Growth Rate (X_i)	Inflation Rate (Y_i)	Cross Product ($(X_i - \bar{X})(Y_i - \bar{Y})$)	Squared Deviations ($(X_i - \bar{X})^2$)	Squared Deviations ($(Y_i - \bar{Y})^2$)
Sum	0.607	0.4306	0.003735	0.004921	0.003094
Average	0.1012	0.0718			
Covariance			0.000747		
Variance				0.000984	0.000619
Std. Dev (s)				0.031373	0.024874

Solution:

The regression equation can be stated as:

$$\text{Inflation rate} = b_0 + b_1 (\text{Money supply growth rate}) + \varepsilon$$

Typically, regression software is used to determine the regression coefficients. However, for illustrative purposes we perform the calculations to make the source of the numbers clear.

$$\text{Slope coefficient} = \hat{b}_1 = \text{Cov}(X, Y) / \text{Var}(X) = 0.000747 / 0.000984 = 0.7591$$

Note that the slope coefficient can be computed in this manner only when there is one independent variable.

In a linear regression, the regression line passes through the coordinates corresponding to the mean values of the independent and dependent variables. Using the mean values for money supply growth ($\bar{X} = 0.1012$) and the inflation rate ($\bar{Y} = 0.0718$) we can compute the intercept term as:

$$\bar{Y} = \hat{b}_0 + \hat{b}_1 \bar{X}$$

$$\hat{b}_0 = 0.0718 - (0.7591)(0.1012) = -0.005$$

Therefore, the relationship between the money supply growth rate and the inflation rate for our sample data can be expressed as:

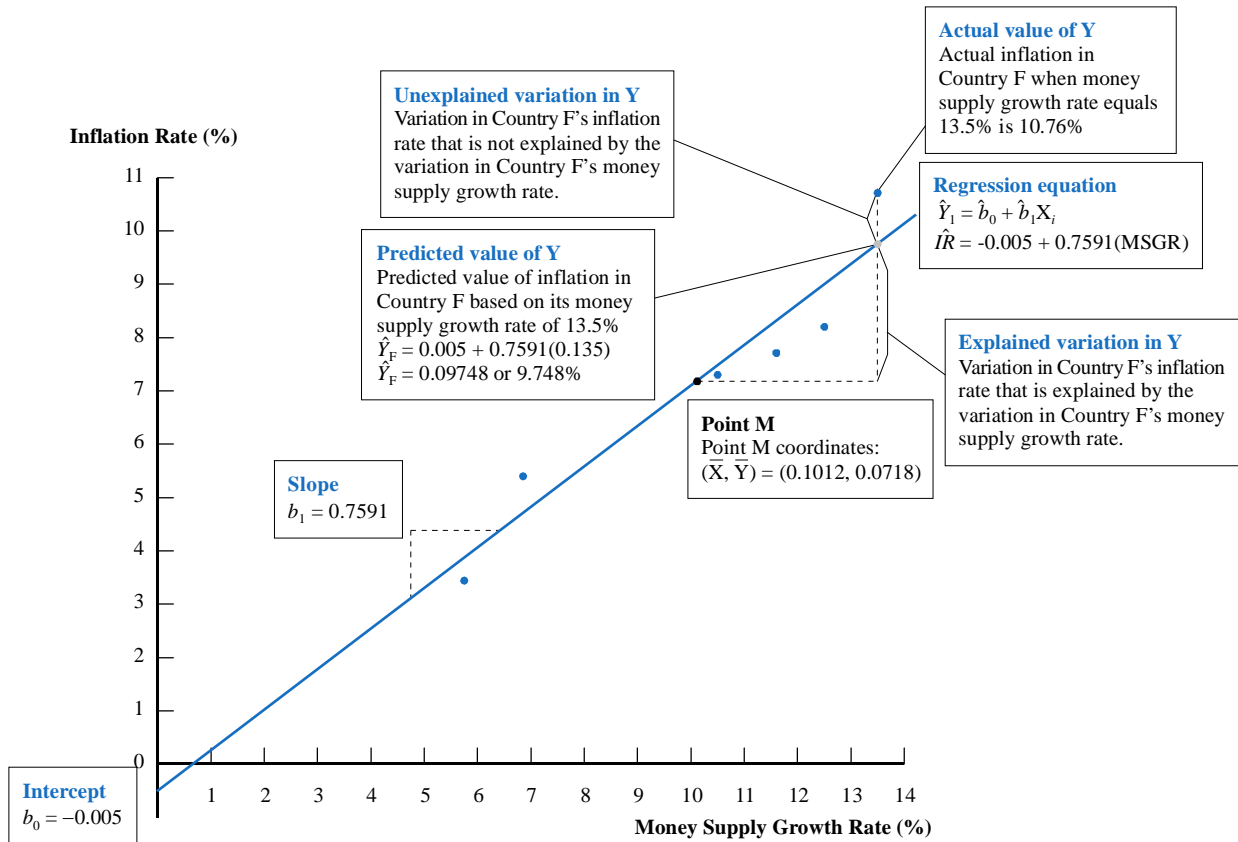
$$\text{Inflation rate} = -0.005 + 0.7591(\text{Money supply growth rate})$$

The regression equation implies that:

- For every 1-percentage-point increase in the money supply growth rate, the inflation rate is predicted to increase by 0.7591 percentage points. The slope coefficient (\hat{b}_1) is interpreted as the estimated change in the dependent variable for a 1-unit change in the independent variable.
- If the money supply growth rate in a country equals 0, the inflation rate in the country will be -0.5%. The intercept (\hat{b}_0) is an estimate of the dependent variable when the independent variable equals 0.

Figure 5 illustrates the regression line (blue line) along with the scatter plot of the actual sample data (blue dots). Using Country F as an example, we also illustrate the difference between the actual observation of the inflation rate in Country F, and the predicted value of inflation (according to the regression model) in Country F.

Figure 5: Regression Line and Scatter Plot



To determine the importance of the independent variable in the regression in explaining the variation in the dependent variable, we need to perform hypothesis tests or create confidence intervals to evaluate the statistical significance of the slope coefficient. Just looking at the magnitude of the slope coefficient does not tell us anything about the importance of the independent variable in explaining the variation in the dependent variable.

LOS 11f: Calculate and interpret the standard error of estimate, the coefficient of determination, and a confidence interval for a regression coefficient.

Vol 1, pg 306 - 310

LOS 11g: Formulate a null and alternative hypothesis about a population value of a regression coefficient, select the appropriate test statistic, and determine whether the null hypothesis is rejected at a given level of significance.

Vol 1, pg 310 - 318

The Standard Error of Estimate

The **standard error of estimate (SEE)**, also known as the **standard error of the regression**, is used to measure how well a regression model captures the relationship between the two variables. It indicates how well the regression line ‘fits’ the sample data and is used to determine how certain we can be about a particular prediction of the dependent variable (\hat{Y}_i) based on a regression equation. A good way to look at the SEE is that it basically measures the standard deviation of the residual term ($\hat{\epsilon}_i$) in the regression. At the risk of stating the obvious, the smaller the standard deviation of the residual term (the smaller the standard error of estimate), the more accurate the predictions based on the model.

The formula for the SEE for a linear regression with one independent variable is:

$$\text{SEE} = \left(\frac{\sum_{i=1}^n (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2}{n - 2} \right)^{1/2} = \left(\frac{\sum_{i=1}^n (\hat{\epsilon}_i)^2}{n - 2} \right)^{1/2} = \left(\frac{\text{SSE}}{n - 2} \right)^{1/2}$$

Note:

- In the numerator of the SEE equation we are essentially calculating the sum of the squared differences between actual and predicted (based on the regression equation) values of the dependent variable.
- We divide the numerator by n-2 (degrees of freedom) to ensure that the estimate of SEE is unbiased.

Example 5: Computing the Standard Error of Estimate

Based on the regression equation: Inflation rate = $-0.005 + 0.7591(\text{Money supply growth rate})$ compute the standard error of estimate (SEE) for the regression.

The squared residuals in the last column can also be denoted by $(\hat{\epsilon}_i)^2$.

The sum of the squared residuals (0.000259) is known as the sum of squared errors (SSE) or sum of squared residuals.

Country	Money Supply Growth Rate (X_i)	Inflation Rate (Y_i)	Predicted Inflation Rate (\hat{Y}_i)	Regression Residual ($Y_i - \hat{Y}_i$)	Squared Residual ($Y_i - \hat{Y}_i$) ²
A	0.0685	0.0545	0.0470	0.0075	0.000057
B	0.1160	0.0776	0.0830	-0.0054	0.000029
C	0.0575	0.0349	0.0386	-0.0037	0.000014
D	0.1050	0.0735	0.0747	-0.0012	0.000001
E	0.1250	0.0825	0.0899	-0.0074	0.000054
F	0.1350	0.1076	0.0974	0.0102	0.000103
Sum					0.000259

Candidates get VERY confused between SEE and SSE so we have introduced both of them here. SEE is the standard deviation of the error term in the regression while SSE equals the sum of the squared residuals in the regression. SSE (as you will see later in the reading) is used to calculate R^2 and the F-stat for the regression. Also note that the two are related by the following equation. $SEE = (SSE/n-2)^{0.5}$

Just to illustrate how we obtained the values in this table, let's perform the calculations for Country A:

$$\text{Predicted inflation rate} = -0.005 + 0.7591(0.0685) = 0.046998$$

$$\text{Regression residual} = 0.0545 - 0.046998 = 0.0075$$

$$\text{Squared residual} = 0.0075^2 = 0.000057$$

From the table (by aggregating the values in the last column) we obtain a figure of 0.000259 as the sum of the squared residuals (SSE). This figure is then plugged into the SEE formula to determine the standard error of estimate.

$$SEE = \left(\frac{0.000259}{6 - 2} \right)^{1/2} = 0.00805 \text{ or } 0.8\%$$

The Coefficient of Determination

The coefficient of determination (R^2) tells us how well the independent variable explains the variation in the dependent variable. It measures the fraction of the total variation in the dependent variable that is explained by the independent variable. The coefficient of determination can be calculated in two ways:

1. $R^2 = (r)^2$

For a linear regression with only one independent variable, the coefficient of determination (R^2) can be calculated by squaring the correlation coefficient (r). In Example 2, we calculated the correlation coefficient between inflation rates and money supply growth from 1990-2010 to be 0.9573. Thus, the coefficient of determination for the regression equals 0.9573^2 or 0.9164. What this means is that variation in money supply growth rate explains about 91.64% of the variation in inflation rates across the 6 countries from 1990-2010.

2. The following method can be used to calculate the coefficient of determination for regressions with one or more independent variables.

- The **total variation** in the dependent variable (sum of squared deviations of observed values of Y from the average value of Y) denoted by $\sum_{i=1}^n (Y_i - \bar{Y})^2$ can be broken down into the variation explained by the independent variable(s) and the variation that remains unexplained by the independent variable(s).
- The variation in the dependent variable that cannot be explained by the independent variable(s) (sum of squared deviations of actual values of Y from the values predicted by the regression equation) denoted by $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ is known as **unexplained variation**.
- The variation in the dependent variable that can be explained by the independent variable(s) (sum of squared deviations of predicted values of Y from the average value of Y) denoted by $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ is known as **explained variation**.

The important thing to note is that R^2 measures the percentage of the total variation in the dependent variable that can be explained by the variation in the independent variable.

$$\text{Total variation} = \text{Unexplained variation} + \text{Explained variation}$$

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}}$$

$$= 1 - \frac{\text{Unexplained variation}}{\text{Total variation}}$$

Note that the square root of the coefficient of determination in a one-independent variable linear regression, after attaching the sign of the estimated slope coefficient, gives the correlation coefficient for the regression.

Example 6: Calculating the Coefficient of Determination

From Example 5 we know that the unexplained variation (sum of squared differences between observed and predicted values of the dependent variable) for our regression involving money supply growth rates and inflation rates equals 0.000259 (SSE). Calculate the total variation in inflation rates and then compute the coefficient of determination for the regression.

Solution:

The computation of the total variation in the dependent variable (inflation rates) is illustrated in the table below:

Country	Money Supply Growth Rate (X_i)	Inflation Rate (Y_i)	Deviation from Mean ($Y_i - \bar{Y}$)	Squared Deviation ($(Y_i - \bar{Y})^2$)
A	0.0685	0.0545	-0.0173	0.000298
B	0.1160	0.0776	0.0058	0.000034
C	0.0575	0.0349	-0.0369	0.001359
D	0.1050	0.0735	0.0017	0.000003
E	0.1250	0.0825	0.0107	0.000115
F	0.1350	0.1076	0.0358	0.001284
	Average	$(\bar{Y}) = 0.0718$	Sum	0.003094

Just to illustrate how we obtained the values in this table, let's perform the calculations for Country A:

$$\text{Deviation from mean} = 0.0545 - 0.0718 = -0.0173$$

$$\text{Squared deviation} = -0.0173^2 = 0.000298$$

From the table (by aggregating the values in the last column) we obtain a figure of 0.003094 as the sum of the squared deviations of observed values of the dependent variable from their average value. This figure represents the total variation in the dependent variable, and given the unexplained variation in the dependent variable (SSE = 0.000259) can be used to calculate the coefficient of determination for the regression as follows:

$$R^2 = \frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}} = \frac{0.003094 - 0.000259}{0.003094} = 0.9162 \text{ or } 91.62\%$$

Hypothesis Tests on Regression Coefficients

Hypothesis tests on the population values of the slope coefficient and the intercept have important practical applications. For example, when using the CAPM to price a risky asset, ABC Stock, a hypothesis test may be used to test the belief that ABC Stock has a market-average level of systematic risk (beta = 1).

$$\text{CAPM: } R_{ABC} = R_F + \beta_{ABC}(R_M - R_F)$$

If we hypothesize that ABC Stock has a beta of 1, and therefore its required return in excess of the risk-free rate is the same as the market's required excess return (equity market risk premium), the regression may be structured as follows:

$$R_{ABC} - R_F = \alpha + \beta_{ABC}(R_M - R_F) + \varepsilon$$

- The intercept term for the regression, b_0 , is α .
- The slope coefficient for the regression, b_1 , is β_{ABC}

Example 7: Hypothesis Tests on Regression Coefficients

Suppose we perform a regression on monthly returns data from January 2006 until December 2010 (60 months) for ABC Stock and the market index. We want to test the null hypothesis that the beta for ABC Stock equals 1 to determine whether the stock has the same required return premium as the market as a whole. The results of the regression are provided below:

Regression Statistics

Multiple R	0.5864		
R-squared	0.3439		
Standard error of estimate	0.0945		
Observations	60		
	Coefficients	Standard Error	t-Statistic
Alpha	0.0041	0.0135	0.3037
Beta	1.1558	0.2096	5.5135

The null and alternative hypotheses for this test are structured as follows:

$$H_0: \beta_{ABC} = 1$$

$$H_a: \beta_{ABC} \neq 1$$

Note that this is a two-tailed test as the null hypothesis has the “=” sign.

The overall market has a Beta of 1.

The test statistic for hypothesis tests on the slope coefficient is calculated as:

$$\text{Test statistic} = t = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}}$$

From the regression results, the estimate of the slope coefficient $\hat{\beta}_{ABC}$ equals 1.1558, while the estimated standard error of the slope coefficient, $s_{\hat{\beta}_{ABC}}$ equals 0.2096. Therefore, the test statistic equals $1.1558 - 1/0.20963 = 0.7433$

To determine the critical t-values for the hypothesis test we need to ascertain the degrees of freedom and specify a level of significance for the test. The degrees of freedom are calculated as the number of observations (n) in the sample minus the number of parameters being estimated in the regression. In this example, we are estimating two parameters (coefficient on the independent variable and the intercept term). Therefore, the degrees of freedom equal $60 - 2 = 58$. We shall assume a 5% significance level for this hypothesis test.

With 58 degrees of freedom, and a 5% level of significance, the critical t-values for this two-tailed hypothesis test are -2.00 and +2.00.

Comparing the test statistic (0.7433) to the upper critical value (2.00) we fail to reject the null hypothesis that ABC Stock has the same level of systematic risk (beta) as the overall market. It is important to note that the t-statistic associated with the slope coefficient, beta, in the regression results is 5.5135. This t-value is computed as the slope coefficient divided by its standard error. It basically represents the test statistic were the null hypothesis structured to test whether β_{ABC} equals 0 (not 1 as is the case in our example). Based on this number ($t = 5.5135$) we would be able to reject the null hypothesis that $\beta_{ABC} = 0$ as it exceeds the critical t-value (2.00).

Finally, notice that the coefficient of determination (R^2) in this regression is only 0.3439. This suggests that only 34.39% of the variation in the excess returns on ABC Stock can be explained by excess returns on the overall market. The remaining 65% of ABC's excess returns can be attributed to firm-specific factors.

Confidence Intervals for Regression Coefficients

A confidence interval is a range of values within which we believe the true population parameter (e.g., b_1) lies, with a certain degree of confidence ($1 - \alpha$). Let's work with the same example that we just used to perform the hypothesis test on ABC Stock's beta to illustrate how confidence intervals are computed and interpreted.

According to the results of the regression in Example 7, the estimate of the slope coefficient (\hat{b}_1) is 1.1558 with a standard error ($s_{\hat{b}_1}$) of 0.2096. The hypothesized value of the population parameter (b_1) is 1 (the market's average slope coefficient or beta). Once again, we use a 5% significance level, or a 95% level of confidence to evaluate our hypothesis.

A confidence interval spans the range from $\hat{b}_1 - t_c \hat{s}_{b_1}$ to $\hat{b}_1 + t_c \hat{s}_{b_1}$

The critical value depends on the degrees of freedom for the test. The degrees of freedom when there are 60 observations equal 58 (calculated as $n - 2$). Given a significance level of 0.05, we determine that the critical t-value is 2.00. Therefore, our 95% confidence interval is calculated as:

$$\hat{b}_1 \pm t_c s_{\hat{b}_1} = 1.1558 \pm (2.00)(0.2096) = 0.7366 \text{ to } 1.575$$

Since we are testing whether b_1 equals 1, and this hypothesized value does lie within the computed interval, we fail to reject the null hypothesis. We can be 95% confident that ABC Stock's beta equals 1.

As you can see, we reach the same conclusion regarding our hypothesis using a hypothesis test and using confidence intervals. Confidence intervals and hypothesis test are linked by critical values.

- In a confidence interval, we aim to determine whether the hypothesized value of the population parameter, slope coefficient ($\beta_{ABC} = 1$) lies within a computed interval (where the interval is based around, or centered on the estimated parameter value from sample data, ($\hat{\beta}_{ABC} = 1.1558$) with a particular degree of confidence ($1 - \alpha$). The confidence interval represents the 'fail-to-reject-the-null region'.
- In a hypothesis test, we examine whether the estimate of the parameter ($\hat{\beta}_{ABC}$) lies in the rejection region, or outside an interval (where the interval is based around, or centered on the hypothesized value of the population parameter, ($\beta_{ABC} = 1$)) at a particular level of significance (α).

Some Important Points Relating to Hypothesis Tests on Regression Coefficients

- The choice of significance level is a matter of judgment. A lower level of significance increases the absolute value of t_{crit} resulting in a wider confidence interval and a lower likelihood of rejecting the null hypothesis.
- Increasing the significance level increases the probability of a Type I error, but decreases the probability of a Type II error.
- The p -value is the lowest level of significance at which the null hypothesis can be rejected. Note that the p -value that is generally reported by statistical software packages as part of the regression results applies to a null hypothesis that the true population parameter equals zero, versus an alternative hypothesis that the true population parameter does not equal zero (given the estimated coefficient and standard error of the coefficient). For example, a p -value of 0.007 tells us that we can reject the null hypothesis that the true population parameter equals zero at the 0.7% level of significance (or with a 99.3% level of confidence).
- The smaller the standard error of an estimated parameter, the stronger the results of the regression and the narrower the resulting confidence intervals.

LOS 11i: Describe the use of analysis of variance (ANOVA) in regression analysis, interpret ANOVA results, and calculate and interpret an F -statistic.
Vol 1, pg 319 - 322

Analysis of variance (ANOVA) is a statistical procedure that is used to determine the usefulness of the independent variable(s) in explaining the variation in the dependent variable. An important part of ANOVA is the F -test, which tests whether all the slope coefficients in the regression are equal to zero. Since we are working with linear regression with one independent variable in this reading, the F -test basically tests a null hypothesis of $b_1 = 0$ versus an alternate of $b_1 \neq 0$.

In order to calculate the F -stat (to perform the F -test) we need the following information:

- The total number of observations (n).
- The total number of parameters to be estimated. In a one-independent variable regression, there are two parameters that are estimated: the intercept term and the slope coefficient.
- **The regression sum of squares (RSS)** - the amount of variation in the dependent variable that is explained by the independent variable. It equals the sum of the squared deviations of predicted values of the dependent variable (based on the regression equation) from the mean value of the dependent variable.

$$RSS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \rightarrow \text{Explained variation}$$

- **The sum of squared errors or residuals (SSE)** - the amount of variation in the dependent variable that cannot be explained by the independent variable. It equals the sum of the squared deviations of actual values of the dependent variable from their predicted values (based on the regression equation).

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \rightarrow \text{Unexplained variation}$$

Note that total variation in the dependent variable is the sum of SSE and RSS. It is calculated as the sum of the squared deviations of the actual values of the dependent variable from the mean value of the dependent variable.

$$\sum_{i=1}^n (Y_i - \bar{Y})^2$$

A typical ANOVA table for a simple linear regression is presented in Figure 6

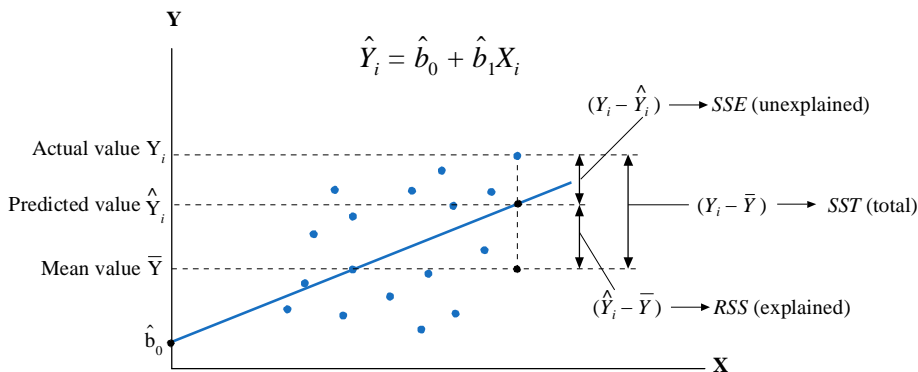
Figure 6: ANOVA Table

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares
Regression (explained)	k	RSS	$MSR = \frac{RSS}{k} = \frac{RSS}{1} = RSS$
Error (unexplained)	$n - (k + 1)$	SSE	$MSE = \frac{SSE}{n - 2}$
Total	$n - 1$	SST	

k = the number of slope coefficients in the regression.

Figure 7 illustrates the components of total variation in the dependent variable.

Figure 7: Components of Total Variation



Given all the pieces of information mentioned on the previous page, the F-stat is calculated as follows:

$$F\text{-stat} = \frac{MSR}{MSE}$$

where:

$$MSR = \frac{RSS}{k}$$

$$MSE = \frac{SSE}{(n - 2)}$$

k = Number of slope parameters estimated in the regression. In a one-independent variable regression, this number equals 1.

Degrees of freedom for F-test = k for numerator, $n-2$ for denominator.

The F-test is a one-tailed test. The decision rule for the test is that we reject H_0 if $F\text{-stat} > F_{\text{crit}}$

If the regression model does a good job of explaining the variation in the dependent variable, explained variation should be relatively high, so MSR should be high relative to MSE, and the value of the F-stat should be higher.

For a one-independent variable regression, the F-stat is basically the same as the square of the t-stat for the slope coefficient. Since it duplicates the t-test for the significance of the slope coefficient for a one-independent variable regression, analysts only use the F-stat for multiple-independent variable regressions. However, we have introduced it in this reading so that you build a solid foundation before moving on to multiple regression in the next reading.

The ANOVA output can also be used to calculate the standard error of estimate (SEE) for the regression. SEE can be calculated as:

$$SEE = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}$$

Example 8: ANOVA

The results of the regression of monthly returns data from January 2006 until December 2010 (60 months) for ABC Stock on the market index, along with the ANOVA table for the regression are provided below. Use the F-stat to determine whether the slope coefficient in the regression equals 0.

Multiple R	0.5864
R-squared	0.3439
Standard error of estimate	0.0404
Observations	60

ANOVA	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Sum of Squares (MSS)	F
Regression	1	0.04953	0.04953	30.39937
Residual	58	0.0945	0.00162931	
Total	59	0.14403		

	Coefficients	Standard Error	t-Statistic
Alpha	0.0041	0.0135	0.3037
Beta	1.1558	0.2096	5.51356

Solution:

$$H_0: \beta_{ABC} = 0$$

$$H_a: \beta_{ABC} \neq 0$$

The F-stat is calculated as:

$$F = \frac{RSS/1}{SSE/(n-2)} = \frac{MSR}{MSE} = \frac{0.04953/1}{0.0945/(60-2)} = 30.399$$

The F-critical value at the 5% level of significance, $F_{crit,1,58} = 4.00$

Since the F-stat for the test (30.399) exceeds the F-critical value (4.00), we reject the null hypothesis and conclude that the slope coefficient for the regression (ABC Stock beta) is significantly different from zero.

Note that the t-test we performed earlier on ABC Stock beta (in Example 7) was different because the null hypothesis then was that the slope coefficient was equal to 1. In this example, in using the F-test, we are testing a null hypothesis that the slope coefficient equals 0.

Also notice that the t-stat in the regression results in Example 7 (5.51356), which as we mentioned earlier, is calculated based on a null hypothesis that the slope coefficient equals zero, equals the square root of our computed F-stat (30.399). Based on this t-stat, which assumes a hypothesized population parameter value of 0, we would reject the null hypothesis that ABC Stock beta equals 0.

Finally, notice the value of the standard error of estimate SEE can be calculated as:

$$SEE = \sqrt{MSE} = \sqrt{0.00162931} = 0.0404$$

LOS 11h: Calculate a predicted value for the dependent variable, given an estimated regression model and a value for the independent variable, and calculate and interpret a confidence interval for the predicted value of a dependent variable. Vol 1, pg 322 - 324

Prediction Intervals

Regression analysis is often used to make predictions or forecasts of the dependent variable based on the regression equation. Typically, analysts construct confidence intervals around the regression forecasts. Note that when we illustrated confidence intervals earlier in the reading, we constructed a confidence interval for the slope coefficient of the independent variable in the regression. Now we shall attempt to create a confidence interval for the dependent variable.

There are two sources of uncertainty when we use a regression model to make a prediction regarding the value of the dependent variable.

- The uncertainty inherent in the error term, ε .
- The uncertainty in the estimated parameters, b_0 and b_1 . If there was no uncertainty regarding the estimates of the regression parameters (i.e., the true population values of the parameters were known with certainty) the variance of any forecast (given an assumed value of X) would simply be s^2 , the squared standard error of estimate, as the error term would be the only source of uncertainty.

Given that the parameters of the regression must be estimated and that their true population values are not known, the estimated variance of the prediction error, s_f^2 , of Y is calculated as:

$$s_f^2 = s^2 \left[1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{(n-1)s_x^2} \right]$$

Once we have computed the estimate of the variance of the prediction error (s_f^2), the prediction interval is constructed using a similar procedure to that of constructing confidence intervals for the slope coefficient. The $(1-\alpha)$ percent prediction interval is constructed as:

$$\hat{Y} \pm t_c s_f$$

Example 9: Computing the Prediction Interval

Go back to the inflation rate and money supply growth rate example that we were working with earlier in the reading. Determine the 95% prediction interval for the inflation rate given that the money supply growth rate is 9%.

Solution:

First, based on the estimate regression equation determine the predicted value of the dependent variable (inflation rate) given the value for the independent variable (money supply growth rate = 0.09)

$$\begin{aligned} \hat{Y} &= \text{Inflation rate} = -0.005 + 0.7591(\text{Money supply growth rate}) \\ \hat{Y} &= \text{Inflation rate} = -0.005 + (0.7591)(0.09) = 0.0633 \end{aligned}$$

To compute the variance of the prediction error we need to calculate (1) the standard error of estimate for the equation, (2) the mean money supply growth rate (\bar{X}) and (3) the variance of the money supply growth rate (Var X).

$$\begin{aligned} \text{SEE} &= 0.0080 \text{ (Computed in Example 5)} \\ \bar{X} &= 0.1012 \text{ (Computed in Example 2)} \\ \text{Var (X)} &= 0.000984 \text{ (Computed in Example 2)} \end{aligned}$$

$$s_f^2 = s^2 \left[1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{(n-1)s_x^2} \right] = 0.008^2 \left[1 + \frac{1}{6} + \frac{(0.09 - 0.1012)^2}{(6-1)(0.000984)} \right] = 0.0000763$$

$$s_f = 0.0087$$

The critical t-value for a 95% confidence interval with 4 degrees of freedom is 2.776. Therefore, the confidence interval extends from $0.0633 - (2.776)(0.0087) = 0.039$ to $0.0633 + (2.776)(0.0087) = 0.087$.

Interpretation: If the money supply growth rate is 9%, the 95% prediction interval for the inflation rate will extend from 3.9% to 8.7%. Since the sample size is small, the prediction interval is relatively wide.

LOS 11j: Discuss the limitations of regression analysis. Vol 1, pg 325

Limitations of Regression Analysis

- Regression relations can change over time. For example, a time series regression estimating beta for a stock may come up with a different estimate of beta depending on the time period selected. This problem is referred to as **parameter instability**.
- Public knowledge of regression relationships (especially in the investment arena) may negate their usefulness going forward as more and more market participants make their investment decisions based on the perceived relationships.
- If the assumptions of regression analysis do not hold, the predictions based on the model will not be valid. These violations of regression assumptions and their repercussions on the analysis are discussed in the next reading.

MULTIPLE REGRESSION AND ISSUES IN REGRESSION ANALYSIS

LOS 12a: Formulate a multiple regression equation to describe the relation between a dependent variable and several independent variables, determine the statistical significance of each independent variable, and interpret the estimated coefficients and their p -values. Vol 1, pg 350-356

LOS 12b: Formulate a null and an alternative hypothesis about the population value of a regression coefficient, calculate the value of the test statistic, determine whether to reject the null hypothesis at a given level of significance by using a one-tailed or two-tailed test, and interpret the results of the test. Vol 1, pg 350-356

Multiple regression is a statistical procedure that allows us to evaluate the impact of more than one (multiple) independent variable on a dependent variable. A multiple linear regression model has the following general form:

$$\text{Multiple regression equation} = Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \dots + b_kX_{ki} + \varepsilon_i, i = 1, 2, \dots, n$$

where:

- Y_i = the i th observation of the dependent variable Y
- X_{ji} = the i th observation of the independent variable $X_j, j = 1, 2, \dots, k$
- b_0 = the intercept of the equation
- b_1, \dots, b_k = the slope coefficients for each of the independent variables
- ε_i = the error term for the i th observation
- n = the number of observations

The slope coefficient, b_1 , measures how much the dependent variable, Y , changes in response to a one-unit change in the independent variable, X_1 , holding all other independent variables constant. For example, if b_1 equals -1 , and all the other independent variables in the regression are held constant, a one unit increase in the independent variable, X_1 , will result in a one-unit decrease in the dependent variable, Y .

Note:

- There are k slope coefficients in a multiple regression.
- The k slope coefficients and the intercept, b_0 , are all known as **regression coefficients**. There are $k+1$ regression coefficients in a multiple regression.
- The residual term, ε_i , equals the difference between the actual value of Y (Y_i) and the predicted value of Y (\hat{Y}_i)

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - (\hat{b}_0 + \hat{b}_1X_{1i} + \hat{b}_2X_{2i} + \dots + \hat{b}_kX_{ki})$$

Example 1: Determining the Significance of the Coefficients in a Multiple Regression

Amy is interested in predicting the GMAT scores of students looking to gain admission into MBA programs around the United States. She specifies a regression model with the GMAT score as the dependent variable and the number of hours spent studying for the test, and the student's college GPA as the independent variables. The regression is estimated from using data from 50 students and is formulated as:

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \varepsilon_i$$

where:

Y_i = A student's GMAT score

b_0 = Intercept term

X_{1i} = Independent Variable 1: The number of hours a student spends preparing for the test.

X_{2i} = Independent Variable 2: The student's undergraduate college GPA.

Amy believes that the higher the number of hours spent preparing for the test, the higher the score obtained on the test (i.e., a positive relationship exists between the two variables). Therefore, she sets up her null and alternative hypotheses for testing the significance of the slope coefficient of X_{1i} (the number of hours spent studying) as follows:

$$H_0: b_1 \leq 0$$

$$H_a: b_1 > 0$$

Amy also believes that the higher a student's college GPA, the higher the score obtained on the test (i.e., a positive relationship exists between the two variables). Therefore she formulates the following hypotheses relating to the slope coefficient of X_{2i} (undergraduate GPA):

$$H_0: b_2 \leq 0$$

$$H_a: b_2 > 0$$

Table 1 shows the results of the regression.

The null hypothesis is the position the researcher is looking to reject. The alternative hypothesis is the condition whose existence the researcher is trying to validate.

For both tests we use a t-test, not the z-test, because we only know sample variances for b_1 and b_2 ; we do not know the population variances for b_1 and b_2 .

Table 1: Results from Regressing GMAT Scores on Hours of Prep and College GPA

	Coefficient	Standard Error	t-Statistic		
Intercept	231.3476	47.3286	4.8881		
Number of hours of study	1.103	0.0939	11.7465		
College GPA	68.3342	16.5938	4.1181		

ANOVA	df	SS	MS	F	Significance F
Regression	2	444866.09	222433.04	73.12	0
Residual	47	142983.91	3042.21		
Total	49	587850			

Standard Error	55.1562
R Square	0.7568
Observations	50
Adjusted R Square	0.7464

As the first step in multiple regression analysis, an analyst should evaluate the overall significance of the regression. The ANOVA section of the regression results provides us with the data that is used to evaluate the overall explanatory power and significance of the regression. We will get into this in detail in LOS 12e. For now, we will move directly into tests relating to the significance of the individual regression coefficients and assume that overall, the regression is significant.

Just like in simple linear regression, the magnitude of the regression coefficients in a multiple regression does not tell us anything about their significance in explaining the variation in the dependent variable. Hypothesis tests must be performed on these coefficients to evaluate their importance in explaining the variation in the dependent variable.

First we evaluate the belief that the higher the number of hours spent studying for the test, the higher the score obtained.

$$H_0: b_1 \leq 0$$

$$H_a: b_1 > 0$$

$$t\text{-stat} = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = 1.103 - 0 / 0.0939 = 11.7465$$

If the regression is not significant overall, there is no point in proceeding to interpret the individual regression coefficients.

Notice that the t-stat used in the hypothesis test is the same as the number listed in the regression results. This is because (as mentioned in Reading 11) the t-stats presented in the regression results are computed on the basis of a hypothesized parameter value of 0.

k represents the number of independent variables in the regression, and 1 is added to account for the intercept term. Therefore, the degrees of freedom equal $n - (k+1)$.

The critical t-value at the 5% level of significance for this one-tailed test with 47 (calculated as $n - (k+1) = 50 - 3$) degrees of freedom is 1.678.

The t-stat (11.7465) is greater than the critical t-value (1.678). Therefore, at the 5% level of significance, we can reject the null hypothesis and conclude that the higher the number of hours a student spends studying for the GMAT, the higher the score obtained.

Next, we evaluate the belief that the higher the student's college GPA, the higher the GMAT score obtained.

$$H_0: b_2 \leq 0$$

$$H_a: b_2 > 0$$

$$t\text{-stat} = \frac{\hat{b}_2 - b_2}{s_{\hat{b}_2}} = 68.3342 - 0 / 16.5938 = 4.1181$$

The critical t-value at the 5% level of significance for this one-tailed test with 47 degrees of freedom is 1.678.

The t-stat (4.1181) is greater than the critical t-value (1.678). Therefore, at the 5% level of significance we can reject the null hypothesis and conclude that the higher the student's undergraduate GPA, the higher the GMAT score obtained.

Most software programs also report a *p*-value for each regression coefficient. The *p*-value represents the lowest level of significance at which a null hypothesis that the population value of the regression coefficient equals 0 can be rejected in a two-tailed test. For example, if the *p*-value for a regression coefficient equals 0.03, the null hypothesis that the coefficient equals 0 can be rejected at the 5% level of significance, but not at the 2% significance level. The lower the *p*-value, the stronger the case for rejecting the null hypothesis.

Based on the results of the regression, our estimated regression equation is:

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_{1i} + \hat{b}_2 X_{2i} = 231.35 + 1.103X_{1i} + 68.3342X_{2i}$$

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_{1i} + \hat{b}_2 X_{2i} = 231.35 + 1.103(\text{no. of hours}) + 68.3342(\text{college GPA})$$

Note that \hat{Y}_i stands for the predicted value of Y_i , and \hat{b}_0 , \hat{b}_1 , and \hat{b}_2 are estimates of the values of b_0 , b_1 and b_2 respectively.

Before moving into interpreting the results of a multiple regression, let's take a step back. Suppose Amy were to start off the process of explaining an individual's GMAT score with a one-independent variable regression model with the number of hours spent studying (X_{1i}) as the only independent variable. The regression equation for her one-independent variable regression is given as:

$$\hat{Y}_i = 260.54 + 2.134X_{1i}$$

The appropriate interpretation of the slope coefficient for this regression equation is that if an individual studies for 1 additional hour, we would expect her GMAT score to increase by 2.134 points.

Then Amy decides to add a second independent variable, a student's college GPA, to her regression model. The equation for her two-independent-variable regression model (obtained through the regression data in Table 1) is given as:

$$\hat{Y}_i = 231.35 + 1.103X_{1i} + 68.3342X_{2i}$$

Notice that the estimated slope coefficient for X_1 has changed from 2.134 (in the one-independent-variable regression equation) to 1.103 (in the two-independent-variable regression equation) when we add X_2 to the regression. This is a fairly typical outcome when another variable is added to a regression (unless X_1 is uncorrelated with X_2) because when X_1 changes by 1 unit, we would expect X_2 to be different as well. The results of the multiple regression capture this relationship between X_1 and X_2 in predicting \hat{Y} .

- In interpreting the slope coefficient for X_1 for the one-independent variable regression model, we state that if an individual studies for 1 additional hour, we would expect her GMAT score to increase by 2.134 points when X_2 is not held constant.
- In interpreting the slope coefficient for X_1 for the two-independent variable regression model, we state that if an individual studies for 1 more hour, we would expect her GMAT score to increase by 1.103 points, *holding her college GPA constant*. This is why the slope coefficients of a multiple regression model are also known as **partial regression coefficients** or **partial slope coefficients**.

Based on the results of her two-independent variable regression, Amy must be careful not to expect the difference in the expected GMAT scores of two individuals whose total number of hours of prep differed by one hour to be 1.103 points. This is because in all likelihood, the college GPAs of the two individuals would differ as well, which would have an impact on their GMAT scores. Therefore, 1.103 points is the expected net effect of each additional hour spent studying for the test (net of the impact of the student's GPA) on her expected GMAT score.

Interpreting the intercept term of the multiple regression equation is fairly straightforward. It represents the expected value of the dependent variable if all the independent variables in the regression equal 0.

LOS 12c: Calculate and interpret 1) a confidence interval for the population value of a regression coefficient and 2) a predicted value for the dependent variable, given an estimated regression model and assumed values for the independent variables. Vol 1, pg 361-363

Confidence Intervals

A confidence interval for a regression coefficient in a multiple regression is constructed in the same manner as we demonstrated in the previous reading, when we constructed a confidence interval for a regression coefficient in a simple linear regression. The confidence interval is constructed as follows:

$$\hat{b}_j \pm (t_c \times s_{\hat{b}_j})$$

estimated regression coefficient \pm (critical t -value)(coefficient standard error)

The critical t -value is a two-tailed value computed based on the significance level ($1 - \text{confidence level}$) and $n - (k+1)$ degrees of freedom.

Note that in the t -test pertaining to b_1 in Example 1, we were testing whether the slope coefficient was greater than zero. In Example 2, when working with a confidence interval, we are testing the hypothesis that the slope coefficient, b_1 , is simply different from zero.

A t -test with a null hypothesis of "equal to zero" at a significance level of α , and a confidence interval with a $(1-\alpha)$ level of confidence will always give the same result.

Example 2: Confidence Interval for a Regression Coefficient in a Multiple Regression

Calculate the 95% confidence interval for the estimated coefficient of number of hours spent studying in our GMAT score example.

Solution:

The critical t -value for a two-tailed test at the 5% level of significance with 47 degrees of freedom is 2.012. Therefore, the confidence interval for the slope coefficient b_1 is:

$$1.103 \pm (2.012)(0.0939) = 0.914 \text{ to } 1.291$$

Since the hypothesized value (0) of the slope coefficient (b_1) of the independent variable, number of hours spent studying (X_1), does not lie within the computed 95% confidence interval, we reject the null hypothesis that the slope coefficient, b_1 , equals 0 at the 5% level of significance.

Predicting the Dependent Variable

Predicting the value of the dependent variable from the multiple regression equation given forecasted or assumed values of the independent variables in the regression is quite straightforward. We simply follow the steps listed below:

- Obtain estimates for $\hat{b}_0, \hat{b}_1, \hat{b}_2, \dots, \hat{b}_k$ of regression parameters $b_0, b_1, b_2, \dots, b_k$.
- Determine the assumed values for independent variables $\hat{X}_1, \hat{X}_2, \dots, \hat{X}_k$.
- Compute the value of the dependent variable, \hat{Y}_1 using the equation

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 \hat{X}_{1i} + \hat{b}_2 \hat{X}_{2i} + \dots + \hat{b}_k \hat{X}_{ki}$$

Do keep in mind that all the independent variables in the regression equation (regardless of whether or not their estimated slope coefficients are significantly different from 0), must be used in predicting the value of the dependent variable.

Example 3: Predicting the Dependent Variable

Amy has put in 270 hours of study for her upcoming GMAT test. Her undergraduate college GPA was 3.64. Based on her regression model, what score should she expect to obtain on her test?

Solution:

$$\begin{aligned}\text{GMAT score} &= 231.35 + 1.103(\text{no. of hours}) + 68.3342(\text{college GPA}) \\ &= 231.35 + 1.103(270) + 68.3342(3.64) \\ &= 777.90 \text{ or approximately } 778\end{aligned}$$

Amy's regression model predicts a score of approximately 778 for her on the GMAT based on 270 hours of prep and a 3.64 college GPA.

Note that when using the estimated regression equation to make predictions of the dependent variable:

- We should be confident that the assumptions of the regression model are met.
- We should be cautious about predictions based on out-of-sample values of the independent variables (values that are outside the range of data on which the model was estimated) as these predictions can be unreliable.

LOS 12d: Explain the assumptions of a multiple regression model.

Vol 1, pg 356-357

Assumptions of the Multiple Linear Regression Model

The classical normal multiple linear regression model makes the following 6 assumptions:

- The relationship between the dependent variable (Y) and the independent variables (X_1, X_2, \dots, X_k) is linear.
- The independent variables (X_1, X_2, \dots, X_k) are not random and no linear relationship exists between two or more independent variables.
- The expected value of the error term, conditioned on the independent variables, is zero: $E(\varepsilon | X_1, X_2, \dots, X_k) = 0$.
- The variance of the error term is the same for all observations. $E(\varepsilon_i^2) = \sigma_\varepsilon^2$.
- The error term is uncorrelated across observations. $E(\varepsilon_i \varepsilon_j) = 0, j \neq i$.
- The error term is normally distributed.

LOS 12e: Calculate and interpret the *F*-statistic, and discuss how it is used in regression analysis. Vol 1, pg 363-365

Testing whether all the Population Regression Coefficients Equal Zero

In Example 1, we illustrated how to conduct hypothesis tests on the individual regression coefficients. We deferred the discussion relating to evaluation of the significance of the estimated regression model as a whole. To address the question, “How well do the independent variables as a group explain the variation in the dependent variable?”, we perform an F-test with a null hypothesis that all the slope coefficients in the regression simultaneously equal zero versus an alternative hypothesis that at least one of the slope coefficients in the regression does not equal zero.

$H_0: b_1 = b_2 = \dots = b_k = 0$
 $H_a: \text{At least one of the slope coefficients} \neq 0$

If none of the independent variables significantly explain the variation in the dependent variable, none of the slope coefficients should be significantly different from zero. However, in a multiple regression, we cannot test the hypothesis that all the slope coefficients equal zero based on t-tests on the individual slope coefficients. This is because the individual t-tests do not account for the effects of the correlation or interaction between the independent variables. The F-test and individual t-tests on the slope coefficients may offer conflicting conclusions in the following scenarios:

1. We may be able to reject the null hypothesis that all the slope coefficients equal zero based on the F-test (and conclude that the regression model significantly explains the variation in the dependent variable) even though none of the individual slope coefficients appear significant based on the individual t-tests. (This is a classic symptom of **multicollinearity**, which we discuss in detail later in the reading).
2. We may fail to reject the null hypothesis that all the slope coefficients equal zero based on the F-test (and conclude that the regression model does not significantly explain the variation in the dependent variable) even though the individual slope coefficients appear to be statistically different from zero based on the individual t-tests.

Details for the ANOVA table for multiple regression are discussed in LOS 12g.

To calculate the F-stat (test statistic when testing the hypothesis that all the slope coefficients in a multiple regression are jointly equal to zero) we need the following inputs, which are typically included in the ANOVA section of the regression results.

- Total number of observations, n .
- Total number of regression coefficients that must be estimated ($k + 1$) where k equals the number of slope coefficients.
- The sum of squared errors or residuals (SSE) which represents unexplained variation.
$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \hat{\epsilon}_i^2$$
- The regression sum of squares (RSS) which represents explained variation.
$$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

The F-stat measures how well the regression model explains the variation in the dependent variable. The greater the F-stat, the better the performance of the regression model in explaining the variation in the dependent variable. From Reading 11, recall that the *F*-stat measures the ratio of the mean regression sum of squares (MSR) to the mean squared error (MSE). It is calculated as follows:

$$F\text{-stat} = \frac{MSR}{MSE} = \frac{RSS/k}{SSE/[n - (k + 1)]}$$

Note that the F-test is a one-tailed test (even though the null hypothesis contains the “=” sign) with the critical F-value computed at the desired level of significance with *k* and *n* – (*k* + 1) degrees of freedom for the numerator and denominator respectively.

Example 4: Testing whether all the Population Regression Coefficients Equal Zero

Evaluate the significance of Amy’s two independent-variable regression in explaining students’ GMAT scores at the 5% level of significance. An excerpt from the regression results is reproduced below:

Table 1: Excerpt

ANOVA	df	SS	MS	Significance	
				F	F
Regression	2	444866.09	222433.04	73.12	0.00
Residual	47	142983.91	3042.21		
Total	49	587850.00			

Solution:

$$H_0: b_1 = b_2 = 0$$

H_a : At least one of the slope coefficients $\neq 0$

$$F\text{-stat} = \frac{444866.09/2}{142983.91/47} = 73.12$$

At the 5% significance level, the critical F-value with 2 and 47 degrees of freedom for the numerator and denominator respectively is between 3.15 and 3.23.

Since the F-stat (73.12) is greater than the critical F-value, we reject the null hypothesis that the slope coefficients on both the independent variables equal zero. We conclude that at least one of the slope coefficients in the regression is significantly different from 0, which basically implies that at least one of the independent variables in the regression explains the variation in the dependent variable to a significant extent. The *p*-value of the F-stat (0) means that the smallest level of significance at which the null hypothesis can be rejected is practically 0. The *p*-value also (as we might expect) suggests that there is a strong case for rejecting the null hypothesis.

LOS 12f: Distinguish between and interpret the R^2 and adjusted R^2 in multiple regression. Vol 1, pg 365-366

R^2 and Adjusted R^2

From Reading 11, recall that the coefficient of determination (R^2) measures how much of the variation in the dependent variable is captured by the independent variables in the regression collectively. It is calculated as:

$$R^2 = \frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}} = \frac{\text{SST} - \text{SSE}}{\text{SST}} = \frac{\text{RSS}}{\text{SST}}$$

In multiple regression analysis, as more and more independent variables are added to the mix, the total amount of unexplained variation will decrease (as the amount of explained variation increases) and the R^2 measure will reflect an improvement on the previous model in terms of the variation explained by the group of independent variables as a proportion of total variation in the dependent variable. This will be the case as long as each newly-added independent variable is even slightly correlated with the dependent variable and is not a linear combination of the other independent variables already in the regression model.

Therefore, when evaluating a multiple regression model, analysts typically use adjusted R^2 . Adjusted R^2 does not automatically increase when another variable is added to the regression as it is adjusted for degrees of freedom. It is calculated as:

$$\text{Adjusted } R^2 = \bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1} \right) (1 - R^2)$$

Note:

- If $k = 1$, R^2 will be greater than adjusted R^2
- Adjusted R^2 will decrease if the inclusion of another independent variable in the regression model results in a nominal increase in explained variation (RSS) and R^2
- Adjusted R^2 can be negative (in which case we consider its value to equal 0) while R^2 can never be negative.
- If we use adjusted R^2 to compare two regression models, we must ensure that the dependent variable is defined in the same manner in the two models and that the sample sizes used to estimate the models are the same.

Example 5: R^2 versus Adjusted R^2

Amy now decides to add a third independent variable (X_{3t}) to her regression model. Upon adding the variable “number of practice tests taken” to the regression, the regression sum of squares (RSS) in the ANOVA increases to 487,342.64, while the sum of squared errors (SSE) falls to 100,507.36. Calculate the R^2 and adjusted R^2 for the new (three-independent variable) regression model and comment on the values.

Solution:

The R^2 and adjusted R^2 for the two-independent variable regression are provided in Table 1 (Example 1).

$$R^2 = 0.7568 \text{ or } 75.68\%$$

$$\text{Adjusted } R^2 = 0.7464 \text{ or } 74.64\%$$

For the new (three-independent variable) regression, R^2 and adjusted R^2 are calculated as:

$$R^2 = \text{RSS}/\text{SST} = 487342.64/587850 = 0.8290 \text{ or } 82.90\%$$

$$\text{Adjusted } R^2 = 1 - \left(\frac{n - 1}{n - k - 1} \right) (1 - R^2) = 0.8217 \text{ or } 82.17\%$$

The R^2 of the three-independent variable regression is higher (82.9% versus 75.68% earlier), but more importantly, the adjusted R^2 of the three-independent variable regression is also higher (82.17% versus 74.64% earlier), which suggests that the new model should be preferred. The addition of the third independent variable has improved the model.

Note that total variation in the independent variable SST is the same in both (two-independent variable and three-independent variable) regression models. This should make sense because the total variation in the dependent variable remains the same regardless of the number of independent variables employed in the regression.

LOS 12g: Infer how well a regression model explains the dependent variable by analyzing the output of the regression equation and an ANOVA table.

This LOS basically covers all the LOSs that we have already covered in this reading. Below we summarize the process of analyzing the output of a regression.

Regression Equation

- Shows the relationship between the dependent variable and the independent variables. Can be used to predict the value of the dependent variable given specific values for the independent variables.
- The significance of the individual regression coefficients is evaluated using t-tests or p-values.
- The t-stat for each regression coefficient is calculated by dividing the value of the coefficient by its standard error.

Table 2: ANOVA Table

Source of Variation	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Sum of Squares (SS/df)
Regression	k	RSS	MSR
Error	n - (k + 1)	SSE	MSE
Total	n - 1	SST	

ANOVA Table

- Lists the regression sum of squares (RSS), sum of squared errors (SSE), and total sum of squares (SST) along with associated degrees of freedom.
- Also includes calculated values for mean regression sum of squares (MSR) and mean squared error (MSE), .
- The F-stat can be calculated by dividing MSR by MSE. The F-test is used to test whether at least one of the slope coefficients on the independent variables in the regression is significantly different from 0.
- R^2 (and adjusted R^2) can be calculated from the data in the ANOVA table by dividing RSS by SST. R^2 is used to determine the goodness of fit of the regression equation to the data.
- The standard error of estimate (SEE) can also be computed from the information in the ANOVA table. $SEE = \sqrt{MSE}$

LOS 12h: Formulate a multiple regression equation by using dummy variables to represent qualitative factors, and interpret the coefficients and regression results. Vol 1, pg 366-371

Using Dummy Variables in a Regression

Dummy variables in regression models help analysts determine whether a particular qualitative variable explains the variation in the model's dependent variable to a significant extent.

- A dummy variable must be binary in nature i.e., it may take on a value of either 0 or 1.
- If the model aims to distinguish between n categories, it must employ $n-1$ dummy variables. The category that is omitted is used as a reference point for the other categories.
- The intercept term in the regression indicates the average value of the dependent variable for the omitted category.
- The slope coefficient of each dummy variable estimates the difference (compared to the omitted category) a particular dummy variable makes to the dependent variable.

If we use n dummy variables (instead of $n-1$) we would be violating the regression assumption of no linear relationship between the independent variables.

Example 6: Hypothesis Testing with Dummy Variables

Let's suppose we are trying to evaluate the seasonality of a company's annual sales. Management believes that sales are significantly different in the fourth quarter compared to the other three quarters. Therefore, we use the fourth quarter as the reference point (i.e., the fourth quarter represents the omitted category) in our regression. The results of the regression based on quarterly sales data for the last 15 years are presented in Table 3:

Table 3: Results from Regressing Sales on Quarterly Dummy Variables

	Coefficient	Standard Error	t-Statistic	
Intercept	4.27	0.97	4.4021	
Sales _{Q1}	-2.735	0.83	-3.295	
Sales _{Q2}	-2.415	0.83	-2.91	
Sales _{Q3}	-2.69	0.83	-3.241	
<hr/>				
ANOVA	df	SS	MS	F
Regression	3	37.328	12.443	26.174
Residual	56	26.623	0.4754	
Total	59	63.951		
<hr/>				
Standard Error	0.6895			
R Square	0.5837			
Observations	60			

Let's first state the regression equation to understand what the variables actually represent

$$Y_t = b_0 + b_1(\text{Sales}_{Q1}) + b_2(\text{Sales}_{Q2}) + b_3(\text{Sales}_{Q3}) + \varepsilon$$

$$\text{Quarterly sales} = 4.27 - 2.735 (\text{Sales}_{Q1}) - 2.415 (\text{Sales}_{Q2}) - 2.690 (\text{Sales}_{Q3}) + \varepsilon$$

- b_0 (4.27) is the intercept term. It represents average sales in the fourth quarter (the omitted category).
- b_1 is the slope coefficient for sales in the first quarter (Sales_{Q1}). It represents the average difference in sales between the first quarter and the fourth quarter (the omitted category). According to the regression results, sales in Q1 are on average 2.735m less than sales in the fourth quarter. Sales in Q1 equal $4.27\text{m} - 2.735\text{m} = 1.535\text{m}$ on average.
- Similarly, sales in Q2 are on average 2.415m less than sales in the fourth quarter, while sales in Q3 are on average 2.69m less than sales in the fourth quarter. Average sales in Q2 and Q3 are 1.855m and 1.58m respectively.

The F-test is used to evaluate the null hypothesis that jointly, the slope coefficients all equal 0.

$$H_0: b_1 = b_2 = b_3 = 0$$

H_a : At least one of the slope coefficients $\neq 0$

The F-stat is given in the regression results (26.174). The critical F-value at the 5% significance level with 3 and 56 degrees of freedom for the numerator and denominator respectively lies between 2.76 and 2.84. Given that the F-stat for the regression is higher, we can reject the null hypothesis that all the slope coefficients in the regression jointly equal 0.

When working with dummy variables, t-stats are used to test whether the value of the dependent variable in each category is different from the value of the dependent variable in the omitted category. In our example, the t-stats can be used to test whether sales in each of the first three quarters of the year are different from sales in the fourth quarter on average.

$H_0: b_1 = 0$ versus $H_a: b_1 \neq 0$ tests whether Q1 sales are significantly different from Q4 sales.
 $H_0: b_2 = 0$ versus $H_a: b_2 \neq 0$ tests whether Q2 sales are significantly different from Q4 sales.
 $H_0: b_3 = 0$ versus $H_a: b_3 \neq 0$ tests whether Q3 sales are significantly different from Q4 sales.

The critical t-values for a two-tailed test with 56 (calculated as $n - (k + 1)$) degrees of freedom are -2.0 and $+2.0$. Since the absolute values of t-stats for the coefficients on each of the three quarters are higher than $+2.0$, we reject all three null hypotheses (that Q1 sales equal Q4 sales, that Q2 sales equal Q4 sales, and that Q3 sales equal Q4 sales) and conclude that sales in each of the first 3 quarters of the year are significantly different from sales in the fourth quarter on average.

LOS 12i: Discuss the types of heteroskedasticity and the effects of heteroskedasticity and serial correlation on statistical inference.

Vol 1, pg 371-382

Violations of Regression Assumptions

Heteroskedasticity

Heteroskedasticity occurs when the variance of the error term in the regression is not constant across observations. Figure 1 shows the scatter plot and regression line for a model with **homoskedastic errors**. There seems to be no systematic relationship between the regression residuals (vertical distances between the data points and the regression line) and the independent variable. Figure 2 shows the scatter plot and regression line for a model with **heteroskedastic errors**. Notice that the regression residuals appear to increase in size as the value of the independent variable increases.

Figure 1: Regression with Homoskedasticity

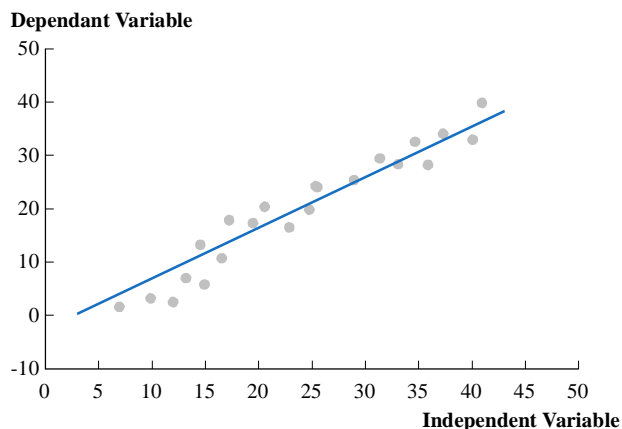
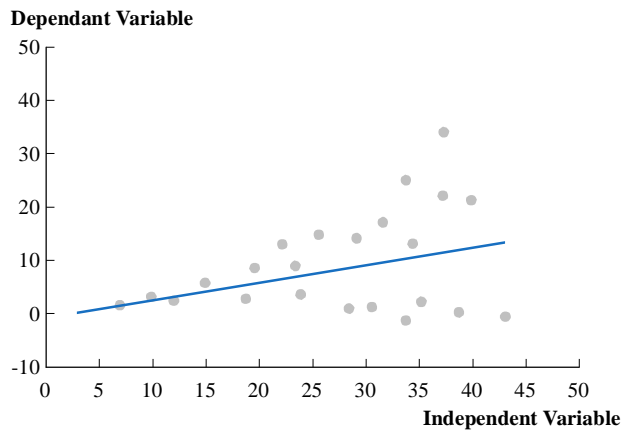


Figure 2: Regression with Heteroskedasticity

Effects of Heteroskedasticity

- Heteroskedasticity does not affect the consistency of estimators of regression parameters.
- However, it can lead to mistakes in inferences made from parameter estimates.
 - The F-test for the overall significance of the regression becomes unreliable as the MSE becomes a biased estimator of the true population variance.
 - The t-tests for the significance of each regression coefficient become unreliable as the estimates of the standard errors of regression coefficients become biased.
 - Typically, in regressions with financial data, standard errors of regression coefficients are underestimated and t-stats are inflated due to heteroskedasticity. Therefore, ignoring heteroskedasticity results in significant relationships being found when none actually exist. (Null hypotheses are rejected too often).
 - Sometimes however, heteroskedasticity leads to standard errors that are too large, which makes t-stats too small.

Note that heteroskedasticity does not affect estimates of the regression coefficients.

Types of Heteroskedasticity

- **Unconditional heteroskedasticity** occurs when the heteroskedasticity of the variance in the error term is not related to the independent variables in the regression. Unconditional heteroskedasticity does not create major problems for regression analysis.
- **Conditional heteroskedasticity** occurs when the heteroskedasticity in the error variance is correlated with the independent variables in the regression. While conditional heteroskedasticity does create problems for statistical inference, it can be easily identified and corrected.

Testing for Heteroskedasticity- The Breusch-Pagan (BP) Test

The BP test requires a regression of **the squared residuals from the original estimated regression equation** (in which the dependent variable is regressed on the independent variables) on the independent variables in the regression.

- If conditional heteroskedasticity does not exist, the independent variables will not explain much of the variation in the squared residuals from the original regression.
- If conditional heteroskedasticity is present, the independent variables will explain the variation in the squared residuals to a significant extent.

The test statistic for the BP test is a Chi-squared (χ^2) random variable that is calculated as:

$$\chi^2 = nR^2 \text{ with } k \text{ degrees of freedom}$$

n = Number of observations

R^2 = Correlation coefficient of the **second regression** (the regression when the squared residuals of the original regression are regressed on the independent variables).

k = Number of independent variables

H_0 : The original regression's squared error term is uncorrelated with the independent variables.

H_a : The original regression's squared error term is correlated with the independent variables.

Note: The BP test is a one-tailed Chi-squared test because conditional heteroskedasticity is only a problem if it is too large.

Example 7: Testing for Heteroskedasticity

An analyst wants to test a hypothesis suggested by Irving Fisher that nominal interest rates increase by 1% for every 1% increase in expected inflation. The Fisher effect assumes the following relationship:

$$i = r + \pi^e$$

where:

i = Nominal interest rate

r = real interest rate (assumed constant)

π^e = Expected inflation

The analyst specifies the regression model as: $i_i = b_0 + b_1\pi^e + \varepsilon_i$

Since the Fisher effect basically asserts that the coefficient on the expected inflation (b_1) variable equals 1, the hypotheses are structured as:

$$H_0: b_1 = 1$$

$$H_a: b_1 \neq 1$$

Quarterly data for 3-month T-bill returns (nominal interest rate) are regressed on inflation rate expectations over the last 25 years. The results of the regression are presented in Table 4:

Table 4: Results from Regressing T-Bill Returns on Expected Inflation

	Coefficient	Standard Error	t-Statistic
Intercept	0.04	0.0051	7.843
Expected inflation	1.153	0.065	17.738
Residual standard error	0.029		
Multiple R-squared	0.45		
Observations	100		
Durbin-Watson statistic	0.547		

To determine whether the data support the assertions of the Fisher relation, we compute the t-stat for the slope coefficient on expected inflation as:

$$\text{Test statistic} = t = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{1.153 - 1}{0.065} = 2.35$$

The critical t-values with 98 degrees of freedom at the 5% significance level are approximately -1.98 and +1.98. Comparing the test statistic (+2.35) to the upper critical t-value (+1.98) we reject the null hypothesis and lean towards concluding that the Fischer Effect does not hold because the coefficient on expected inflation appears to be significantly different from 1.

However, before accepting the validity of the results of this test, we should test the null hypothesis that the regression errors do not suffer from conditional heteroskedasticity. A regression of the squared residuals from the original regression on expected inflation rates yields an R^2 of 0.193. The test statistic for the BP test is calculated as:

$$\chi^2 = nR^2 = (100)(0.193) = 19.3$$

The critical χ^2 value at the 5% significance level for a one-tailed test with 1 degree of freedom is 3.84. Since the test-statistic (19.3) is higher, we reject the null hypothesis of no conditional heteroskedasticity in the error terms. Since conditional heteroskedasticity is present in the residuals (of the original regression) the standard errors computed in the original regression are incorrect and we cannot accept the result of the t-test above (which provides evidence against the Fisher relation) as valid.

Correcting Heteroskedasticity

There are two ways to correct for conditional heteroskedasticity in linear regression models:

- Use **robust standard errors** (**White-corrected standard errors** or **heteroskedasticity-consistent standard errors**) to recalculate the t-statistics for the original regression coefficients based on corrected-for-heteroskedasticity standard errors.
- Use **generalized least squares**, where the original regression equation is modified to eliminate heteroskedasticity.

Example 8: Using Robust Standard Errors to Adjust for Conditional Heteroskedasticity

The analyst corrects the standard errors obtained in the initial regression of 3-month T-bill returns (nominal interest rates) on expected inflation rates for heteroskedasticity and obtains the results presented in Table 5:

Table 5: Results from Regressing T-Bill Returns on Expected Inflation. (Standard Errors Corrected for Conditional Heteroskedasticity)

	Coefficient	Standard Error	t-Statistic
Intercept	0.04	0.0048	8.333
Expected inflation	1.153	0.085	13.565
Residual standard error	0.029		
Multiple R-squared	0.45		
Observations	100		

Compared to the regression results in Table 4 (Example 7) notice that the standard error for the intercept does not change significantly, but the standard error for the coefficient on expected inflation increases by about 30% (from 0.065 to 0.085). Further, the regression coefficients remain the same (0.04 for the intercept and 1.153 for expected inflation).

Using the adjusted standard error for the slope coefficient, the test-statistic for our hypothesis test is calculated as:

$$\text{Test statistic} = t = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{1.153 - 1}{0.085} = 1.8$$

Comparing this test statistic to the upper critical t-value (1.98) leads us to fail to reject the null hypothesis. The conditional heteroskedasticity in the data was so significant that the result of our hypothesis test changed (compared to Example 7) once the standard errors were corrected for heteroskedasticity (in Example 8). We now conclude that the Fischer Effect does hold as the slope coefficient on the expected inflation independent variable does not significantly differ from 1.

Serial Correlation

Serial correlation (autocorrelation) occurs when regression errors are correlated across observations. It typically arises in time series regressions.

- **Positive serial correlation** occurs when a positive (negative) error for one observation increases the chances of a positive (negative) error for another.
- **Negative serial correlation** occurs when a positive (negative) error for one observation increases the chances of a negative (positive) error for another.

Effects of Serial Correlation

- Serial correlation does not affect the consistency of the estimated regression coefficients unless one of the independent variables in the regression is a lagged value of the dependent variable. For example, when examining the Fisher relation, if we were to use the T-bill return for the previous month as an independent variable (even though the T-bill return that represents the nominal interest rate is actually the dependent variable in our regression model) serial correlation would cause the parameter estimates from the regression to be inconsistent. In this reading, we assume that none of the independent variables is a lagged value of the dependent variable.
- When a lagged value of the dependent variable is not an independent variable in the regression, positive (negative) serial correlation:
 - Does not affect the consistency of the estimated regression coefficients.
 - Causes the F-stat (which is used to test the overall significance of the regression) to be inflated (deflated) because MSE will tend to underestimate (overestimate) the population error variance.
 - Causes the standard errors for the regression coefficients to be underestimated (overestimated), which results in larger (smaller) t-values. Consequently, analysts may reject (fail to reject) null hypotheses incorrectly, make Type I errors (Type II errors) and attach (fail to attach) significance to relationships that are in fact not significant (significant).

In this reading we also make the common assumption that serial correlation takes the form of first-order serial correlation i.e., serial correlation only exists between adjacent observations.

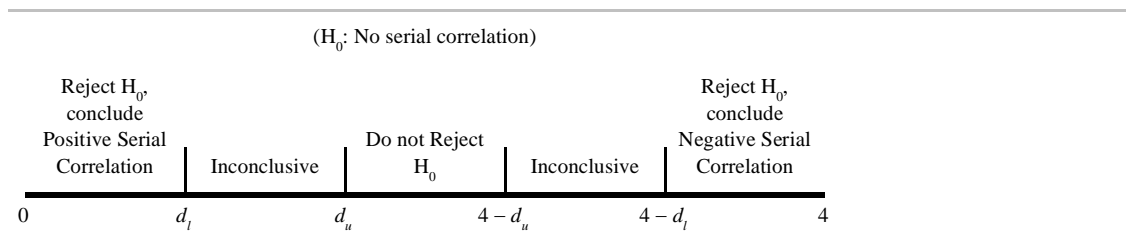
Testing for Serial Correlation- The Durbin-Watson (DW) Test

The DW test-statistic is approximated as:

$DW \approx 2(1 - r)$; where r is the sample correlation between squared residuals from one period and those from the previous period.

- The DW-stat can range from 0 (when serial correlation equals +1) to 4 (when serial correlation equals -1)
- If the regression has no serial correlation, the DW stat equals 2.
- If the regression residuals are positively serially correlated, the DW stat will be less than 2.
- If the regression residuals are negatively serially correlated, the DW stat will be greater than 2.
- For a given sample, the critical DW value (d^*) is not known with certainty. We only know that it lies between two values (d_l and d_u). Figure 3 depicts the lower and upper values for d^* as they relate to the results of the DW test.

Figure 3: Value of Durbin-Watson Statistic



Decision rules for Durbin-Watson tests:

When testing for positive serial correlation:

- Reject H_0 of no positive serial correlation if the DW stat is lower than d_l . Conclude that there is positive serial correlation.
- The test is inconclusive if the DW stat lies between d_l and d_u .
- Fail to reject H_0 of no positive serial correlation when DW stat is greater than d_u .

When testing for negative serial correlation:

- Reject H_0 of no negative serial correlation if the DW stat is higher than $4 - d_l$. Conclude that there is negative serial correlation.
- The test is inconclusive if the DW stat lies between $4 - d_u$ and $4 - d_l$.
- Fail to reject H_0 of no negative serial correlation when DW stat is less than $4 - d_u$.

Example 9: Testing for Serial Correlation

Let's go back to the regression in Table 2 (Example 7) where we were examining the Fisher relation. We are given a Durbin-Watson statistic of 0.547 for the regression. Based on the DW stat formula on the previous page, we use this value to first determine whether the regression residuals are positively or negatively serially correlated.

$$DW \approx 2(1 - r) = 0.547$$

$$r = 1 - DW/2$$

$$r = 0.7265$$

This positive value of r raises a concern that the standard errors of the regression may suffer from positive serial correlation, which may cause the OLS regression standard errors to be underestimated. Therefore, we must determine whether the observed value of the DW stat provides enough evidence to reject the null hypothesis of no positive serial correlation.

Given that the Fisher relation regression has one independent variable and 100 observations, the critical DW value lies between 1.65 (d_l) and 1.69 (d_u). Since the DW test stat (0.547) lies below d_l , we reject the null hypothesis of no positive serial correlation. The results of the test suggest that the standard errors of the original regression are positively serially correlated and are therefore, too small.

Correcting Serial Correlation

There are two ways to correct for serial correlation in the regression residuals:

- Adjust the coefficient standard errors to account for serial correlation using [Hansen's method](#) (which incidentally also corrects for heteroskedasticity). The regression coefficients remain the same but the standard errors change. After correcting for positive serial correlation, the robust standard errors are larger than they were originally. Note that the DW stat still remains the same.
- [Modify the regression equation](#) to eliminate the serial correlation.

Example 10: Correcting for Serial Correlation

Table 6 shows the results of correcting the standard errors of the original regression for serial correlation and heteroskedasticity using Hansen’s method.

Table 6: Results from Regressing T-Bill Returns on Expected Inflation. (Standard Errors Corrected for Conditional Heteroskedasticity and Serial Correlation)

	Coefficient	Standard Error	t-Statistic
Intercept	0.04	0.0088	4.545
Expected inflation	1.153	0.155	7.439
Residual standard error	0.029		
Multiple R-squared	0.45		
Observations	100		
Durbin-Watson statistic	0.547		

Note that the coefficients for the intercept and slope are exactly the same (0.04 for the intercept and 1.153 for expected inflation) as in the original regression (Example 7). Further, note that the DW stat is the same (0.547), but the standard errors have been corrected (they are now much larger) to account for the positive serial correlation.

Given these new and more accurate coefficient standard errors let’s once again test the null hypothesis that the coefficient on the expected inflation independent variable equals 1. The test statistic for the hypothesis test is computed as:

$$\text{Test statistic} = t = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{1.153 - 1}{0.155} = 0.987$$

The critical t-values with 98 degrees of freedom at the 5% significance level are approximately -1.98 and +1.98. Comparing the test statistic (0.987) to the upper critical t-value (+1.98) we fail to reject the null hypothesis and conclude that the Fischer Effect does hold as the slope coefficient on the expected inflation independent variable does not significantly differ from 1. Note that the result of this hypothesis test is different from the test we conducted using the standard errors of the original regression (which were affected by serial correlation and heteroskedasticity) in Example 7. Further, the result is the same as the test conducted on White-corrected standard errors (which were corrected for heteroskedasticity) in Example 8.

LOS 12j: Describe multicollinearity, and discuss its causes and effects in regression analysis. Vol 1, pg 382-385

Multicollinearity

Multicollinearity occurs when two or more independent variables (or combinations of independent variables) in a regression model are highly (but not perfectly) correlated with each other.

Perfect collinearity is much less of a practical concern than multicollinearity.

Effects of Multicollinearity

- Multicollinearity does not affect the consistency of OLS estimates and regression coefficients, but makes them inaccurate and unreliable.
- It becomes difficult to isolate the impact of each independent variable on the dependent variable.
- The standard errors for the regression coefficients are inflated, which results in t -stats becoming too small and less powerful (in terms of their ability to reject null hypotheses).

Detecting Multicollinearity

It has been suggested that high pair-wise correlations between the independent variables may indicate the existence of multicollinearity. However, this is not always the case. Only when there are exactly two independent variables in the regression is the magnitude of the correlation between the independent variables a reasonable indicator of multicollinearity (especially when the correlation between them is greater than 0.7). Otherwise; low pair-wise correlations between the independent variables in the regression do not mean that multicollinearity is not present, and high pair-wise correlations between the independent variables in the regression are not necessary for multicollinearity to exist.

A high R^2 and a significant F -stat (both of which indicate that the regression model overall does a good job of explaining the dependent variable) coupled with insignificant t -stats of slope coefficients (which indicate that the independent variables individually do not significantly explain the variation in the dependent variable) provide the classic case of multicollinearity. The low t -stats on the slope coefficients increase the chances of Type II errors: failure to reject the null hypothesis when it is false.

Bear in mind that multicollinearity may be present even when we do not observe insignificant t -stats and a highly significant F -stat for the regression model.

Example 11: Multicollinearity

An individual is trying to determine how closely associated the investment strategy followed by her portfolio manager is with the returns of a value index and the returns of a growth index over the last 60 years. She regresses the historical annual returns of her portfolio on the historical returns of the S&P 500/BARRA Growth Index, S&P 500/BARRA Value Index and the S&P 500. Results of her regression are given in Table 7:

Table 7: Results from Regressing Portfolio Returns against S&P 500/BARRA Growth and Value Indexes and the S&P500.

Regression Coefficient	t-Stat
Intercept	1.250
S&P 500/BARRA Growth Index	-0.825
S&P 500/BARRA Value Index	-0.756
S&P 500 Index	1.520

<i>F</i> -Stat	35.17
R ²	82.34%
Observations	60

Evaluate the results of the regression.

Solution:

The absolute values of the t-stats for all the regression coefficients- the intercept (1.25), slope coefficient on the growth index (0.825), slope coefficient on the value index (0.756) and the slope coefficient on the S&P 500 (1.52) are lower than the absolute value of t_{crit} (2.00) at the 5% level of significance (df = 56). This suggests that none of the coefficients on the independent variables in the regression are significantly different from 0.

However, the *F*-stat (35.17) is greater than the *F* critical value of 2.76 ($\alpha = 0.05$, df = 3, 56), which suggests that the slope coefficients on the independent variables do not jointly equal zero (at least one of them is significantly different from 0). Further, the R² (82.34%) is quite high, which means that the model as a whole does a good job of explaining the variation in the portfolio's returns.

This regression therefore, clearly suffers from the classic case of multicollinearity as described earlier.

Correcting for Multicollinearity

Analysts may correct for multicollinearity by excluding one or more of the independent variables from the regression model. **Stepwise regression** is a technique that systematically removes variables from the regression until multicollinearity is eliminated.

Example 12: Correcting for Multicollinearity

Given that the regression in Example 11 suffers from multicollinearity, the independent variable-return on the S&P 500 is removed from the regression. Results of the regression with only the return on the S&P 500/BARRA Growth Index and the return on the S&P 500/BARRA Value Index as independent variables are given in Table 8:

Table 8: Results from Regressing Portfolio Returns against S&P 500/BARRA Growth and Value Indexes.

Regression Coefficient	t-Stat
Intercept	1.35
S&P 500/BARRA Growth Index	6.53
S&P 500/BARRA Value Index	-1.16

<i>F</i> -Stat	57.62
R ²	82.12%
Observations	60

Evaluate the results of this regression.

Solution:

The t -stat of the slope coefficient on the growth index (6.53) is greater than the t -critical value (2.00) indicating that the slope coefficient on the growth index is significantly different from 0 at the 5% significance level. However, the t -stat of the value index (-1.16) is not different from 0 at the 5% significance level. This suggests that returns on the portfolio are linked to the returns on the growth index, but not closely related to the returns on the value index.

The F -stat (57.62) is greater than the F critical value of 3.15 ($\alpha = 0.05$, $df = 2, 57$), which suggests that the slope coefficients on the independent variables do not jointly equal zero. Further, the R^2 (82.12%) is quite high, which means that the model as a whole does a good job of explaining the variation in the portfolio's returns.

Removing the return on the S&P 500 as an independent variable in the regression corrected the multicollinearity problem in the initial regression. The significant relationship between the portfolio's returns and the return on the growth index was uncovered as a result.

Table 7: Problems in Linear Regression and Solutions¹

Problem	Effect	Solution
Heteroskedasticity	Incorrect standard errors	Use robust standard errors (corrected for conditional heteroskedasticity)
Serial correlation	Incorrect standard errors (additional problems if a lagged value of the dependent variable is used as an independent variable)	Use robust standard errors (corrected for serial correlation)
Multicollinearity	High R^2 and low t -statistics	Remove one or more independent variables; often no solution based in theory

¹ Table 11, pg 386, Vol 1, Level II CFA Program Curriculum 2012

LOS 12k: Discuss the effects of model misspecification on the results of a regression analysis, and explain how to avoid the common forms of misspecification. Vol 1, pg 386-401

Model Specification

Principles of Model Specification

- The model should be backed by solid economic reasoning. Data mining (where the model is based on the characteristics of the data) should be avoided.
- The functional form for the variables in the regression should be in line with the nature of the variables.
- Each variable in the model should be relevant, making the model ‘parsimonious’.
- The model should be tested for violations of regression assumptions before being accepted.
- The model should be found useful out of sample.

Model Specification Errors

In describing different model specification errors, we shall work with the following regression equation, which represents the ‘true regression model’ for explaining the variation in a particular dependent variable.

$$Y_i = b_0 + b_1 \ln X_{1i} + b_2 X_{2i} + \varepsilon$$

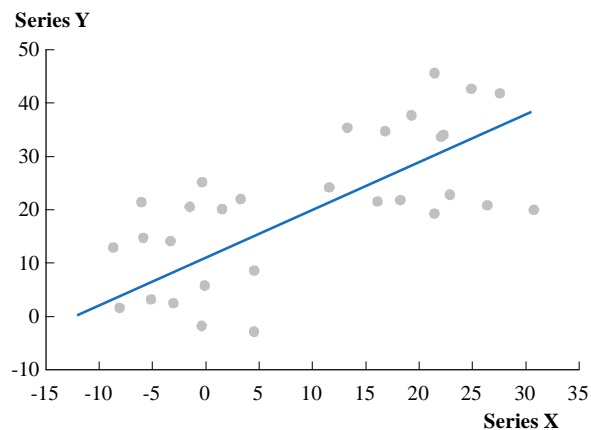
Misspecified Functional Form

1. One or more important variables may have been omitted from the regression (e.g. if we were to leave out X_2 from our model altogether and estimate the following regression equation: $Y_i = a_0 + a_1 \ln X_{1i} + \varepsilon$. If the omitted variable (X_2) is correlated with the included variable ($\ln X_1$), the error term in the model would be correlated with $\ln X_1$, the estimates of the regression coefficients (a_0 and a_1) would be biased and inconsistent, while their estimated standard errors would also be inconsistent. As a result, neither the coefficients nor their standard errors would be useful for statistical analysis.
2. A wrong form of the data may be used in the regression. One or more of the variables may need to be transformed before estimating the regression.
 - If the relationship between the dependent and independent variables is nonlinear, but becomes linear when one or more of the variables is presented as a proportional change in the variable, the misspecification may be corrected by using the natural logarithm of the variable(s). For example, in our true regression equation, Variable X_1 has been transformed to $\ln X_1$.

Note that when X_2 is omitted from the regression, the values of the intercept and slope coefficient on $\ln X_1$ are different than in our true regression. Hence we represent the regression coefficients as a_0 and a_1 instead of b_0 and b_1 .

- Sometimes it is more appropriate to use a scaled version of the data as opposed to unscaled data. When comparing financial statements of companies, analysts often use common size statements to ensure that the results of the regression are based on companies' underlying economics, not on differences in their respective sizes. For example, a regression seeking to identify a relationship between cash flow from operations and free cash flow for different companies should scale the data by dividing both (the dependent and independent) variables by sales so that differences in company size are accounted for, and the regression is correctly specified.
3. The model may pool data from different sources that should not have been pooled. Figure 4 illustrates this type of misspecification. In each cluster of data, there seems to be no clear relationship between X and Y. However, if both these clusters are treated as one sample when estimating the regression, the analyst may find a 'statistical' relationship (spurious correlation) when no economic relationship actually exists.

Figure 4: Regression Line Based on Two Different Sources of Data



Time Series Misspecification

Time-series misspecification results from the kinds of independent variables included in the regression. Time series misspecification can result from:

1. **Including lagged dependent variables as independent variables in regressions with serially correlated errors.** The lagged dependent variable (which serves as an independent variable in the regression) will be correlated with the error term (violating the regression assumption that independent variables must be uncorrelated with the error term). When such a misspecification occurs, estimates of regression coefficients will be biased and inconsistent. An example of this type of misspecification was mentioned earlier in the reading (in our discussion regarding the effects of serial correlation), when we proposed adding the previous month's T-bill rate (lagged version of the dependent variable, nominal interest rates) as an independent variable to predict the 3-month T-Bill rate. In our true regression model, adding Y_{t-1} as a third independent variable in the regression would create a similar problem.
2. **Using the regression to forecast the dependent variable at time, $t+1$ based on independent variables that are a function of the dependent variable at time, $t+1$.** When this occurs, the independent variable is correlated with the error term so the model is misspecified. For example, suppose an analyst builds a model to predict the returns of various stocks over 2010 based on their P/BV ratios at the end of 2010. In this case, a high return (the dependent variable) for 2010 actually causes the high P/BV ratio (independent variable) at the end of 2010 rather than the other way round. The same variable (price return) effectively appears on both sides of the regression equation so it would be incorrect to assert (based on this model) that returns can be predicted based on P/BV ratios.
3. **Independent variables are measured with error.** For example, in the Fisher relation equation, if we were to use actual inflation instead of expected inflation as the independent variable in the regression, the independent variable (actual inflation rates) would be correlated with the error term and the regression coefficients would be biased and inconsistent.
4. Another source of misspecification in time series models is **nonstationarity** (which is discussed in detail in the next reading) which occurs when a variable's properties, (e.g. mean and variance) are not constant over time.

LOS 12l: Discuss models with qualitative dependent variables.**Vol 1, pg 401-403**

Qualitative Dependent Variables

A qualitative dependent variable is basically a dummy variable used as a dependent variable instead of as an independent variable (as we discussed earlier in the reading) in the regression. For example, whether or not a company will go bankrupt can be modeled as a qualitative dependent variable (1 = Will go bankrupt; 0 = Will not go bankrupt) based on various independent variables like its return on equity, leverage ratios, interest coverage ratios, etc. A linear regression model cannot be used to capture the relationship between the variables because the value that the dependent variable can take under such a model could be less than 0 or even greater than 1 (which is not empirically possible as the probability of going bankrupt cannot possibly be less than 0% or greater than 100%). Therefore, probit, logit, or discriminant models are used to model such regressions.

The **probit model** is based on the normal distribution. It estimates the probability that a qualitative condition is fulfilled ($Y = 1$) given the value of the independent variable (X). The **logit model** is similar except that it is based on the logistic distribution. Both models use maximum likelihood methodologies.

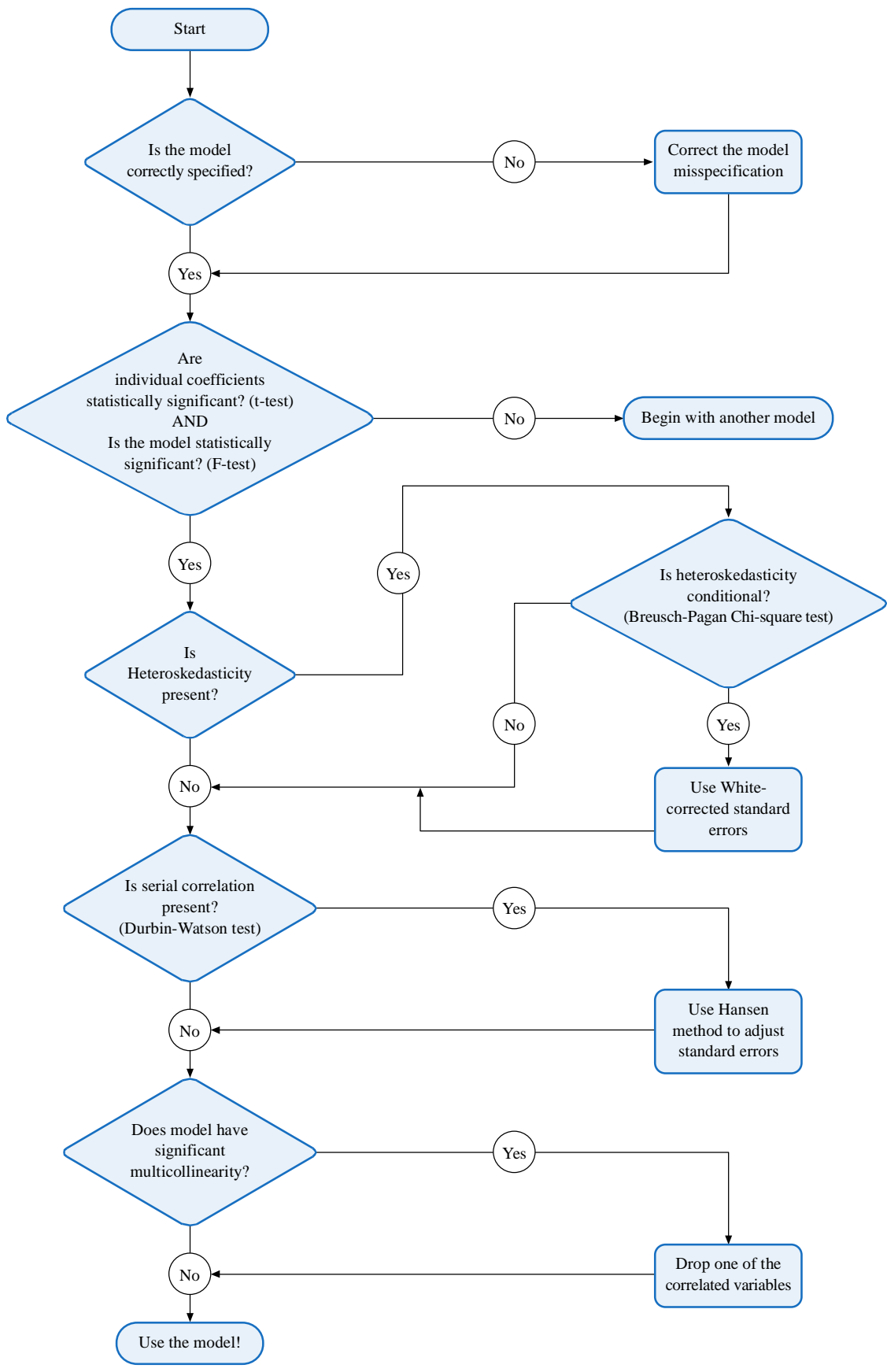
Discriminant analysis offers a linear function (similar to a regression equation) that is used to create an overall score on the basis of which an observation can be classified qualitatively (e.g. into a bankrupt or not bankrupt category).

The analysis of these models is very similar to the analysis of linear regression models as illustrated in this reading. The significance of the individual coefficients is evaluated using t -tests, while the strength of the overall model is judged on the basis of the F -test and R^2 . Analysts must also watch out for heteroskedasticity, serial correlation and multicollinearity in the regression.

LOS 12m: Interpret the economic meaning of the results of multiple regression analysis and critique a regression model and its results.

This LOS basically covers everything we have covered in this reading. Just to help you review, we list the steps in assessing a multiple regression model in the flowchart in Figure 5:

Figure 5: Steps in Assessing a Multiple Regression Model



TIME-SERIES ANALYSIS

A **time series** is a set of observations of the outcomes for a particular variable over a period of time (e.g. the quarterly sales of a company over the last 10 years). Time series analysis is undertaken (1) to explain the past and (2) to predict the future.

LOS 13a: Calculate and evaluate the predicted trend value for a time series, modeled as either a linear trend or a log-linear trend, given the estimated trend coefficients. Vol 1, pg 440-449

LOS 13b: Describe factors that determine whether a linear or a log-linear trend should be used with a particular time series, and evaluate the limitations of trend models. Vol 1, pg 440-449

TREND MODELS

Linear Trend Models

A **linear trend model** is one in which the dependent variable changes by a constant amount in each period. On a graph, a linear trend is presented as a straight line, with a positively-sloped line indicating an upward trend, and a negatively-sloped line indicating a downward trend. Linear trends can be modeled with the following regression equation:

$$y_t = b_0 + b_1t + \varepsilon_t, \quad t = 1, 2, \dots, T$$

where:

y_t = the value of the time series at time t (value of the dependent variable)

b_0 = the y -intercept term

b_1 = the slope coefficient/ trend coefficient

t = time, the independent or explanatory variable

ε_t = a random-error term

Ordinary least squares (OLS) regression is used to estimate the regression coefficients (\hat{b}_0 and \hat{b}_1) and the resulting regression equation is used to predict the value of the time series (y_t) for any period (t). Notice that this model is very similar to the simple linear regression model that we studied earlier. In a linear trend model, the independent variable is the time period.

Another thing to note is that in a linear trend model, the value of the dependent variable changes by b_1 (the trend coefficient) in each successive time period (as t increases by 1 unit) irrespective of the level of the series in the previous period.

Example 1: Linear Trend Models

Keiron Gibbs wants to estimate the linear trend in inflation in Gunnerland over time. He uses monthly observations of the inflation rate (expressed as annual percentage rates) over the 30 year-period from January 1981 to December 2010 and obtains the following regression results:

Table 1: Estimating a Linear Trend for Monthly Inflation Data

Regression Statistics			
R-squared	0.0537		
Standard error	2.3541		
Observations	360		
Durbin-Watson	1.27		
	Coefficient	Standard Error	t-Stat
Intercept	4.2587	0.4132	10.3066
Trend	-0.0087	0.0029	-3

Evaluating the Significance of Regression Coefficients

At the 5% significance level with 358 (calculated as 360-1-1) degrees of freedom, the critical t-value for a two-tailed test is 1.972. Since the absolute values of the t-statistics for both the intercept (10.3066) and the trend coefficient (-3.00) are greater than the critical t-value, we conclude that both the regression coefficients ($\hat{b}_0 = 4.2587$, $\hat{b}_1 = -0.0087$) are statistically significant.

Estimating the Regression Equation

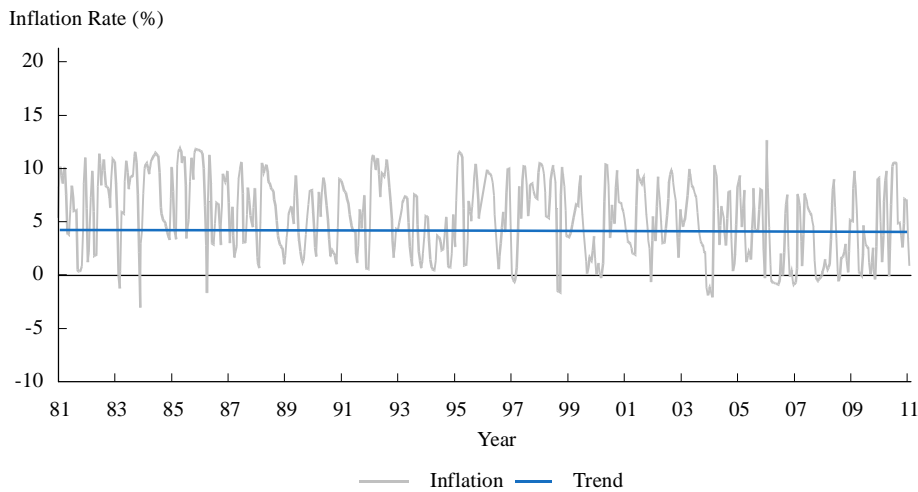
Based on these results, the estimated regression equation would be written as:

$$y_t = 4.2587 - 0.0087t$$

Using the Regression Results to Make Forecasts

The regression equation can be used to make **in-sample forecasts** (e.g. inflation for $t = 12$, December 1981 is estimated at $4.2587 - 0.0087(12) = 4.1543\%$) and **out-of-sample forecasts** (e.g. inflation for $t = 384$, December 2012 is estimated at $4.2587 - 0.0087(384) = 0.9179\%$). The regression equation also tells us that the inflation rate decreased by approximately 0.0087% (the trend coefficient) each month during the sample period.

Figure 1 shows a plot of the **actual time series** (monthly observations of the inflation rate during the sample period) along with the **estimated regression line**. Notice that the residuals appear to be uncorrelated over time and are not persistent. Therefore, use of the linear trend to model the time series seems appropriate. However, the low R^2 of the model (5.37%) suggests that inflation forecasts from the model are quite uncertain, and that a better model may be available.

Figure 1: Monthly CPI Inflation with Trend

Log-Linear Trend Models

A linear trend would not be appropriate to model a time series that exhibits exponential growth (i.e., constant growth at a particular rate) because the regression residuals would be persistent. Use of a log-linear trend may be more appropriate as such a model typically fits a time series that exhibits exponential growth quite well. A series that grows exponentially can be described using the following equation:

$$y_t = e^{b_0 + b_1 t}$$

where:

y_t = the value of the time series at time t (value of the dependent variable)

b_0 = the y -intercept term

b_1 = the slope coefficient

t = time = 1, 2, 3 ... T

In this equation, the dependent variable (y_t) is an exponential function of the independent variable, time (t). We take the natural logarithm of both sides of the equation to arrive at the equation for the log-linear model:

$$\ln y_t = b_0 + b_1 t + \varepsilon_t, \quad t = 1, 2, \dots, T$$

The equation linking the variables, y_t and t , has been transformed from an exponential function to a linear function (the equation is linear in the coefficients, b_0 and b_1) so we can now use linear regression to model the series.

Exponential growth is growth at a constant rate ($e^{b_1} - 1$) with continuous compounding.

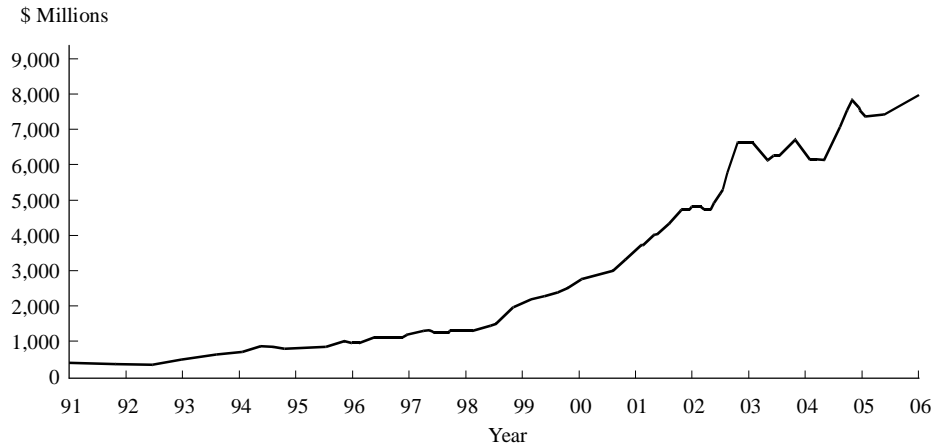
No time series grows at an exact exponential rate so we add the error term to the log linear model equation.

Example 2: Linear versus Log-Linear Trend Model

Samir Nasri wants to model the quarterly sales made by ABC Company over the 15-year period from 1991 to 2005. Quarterly sales data over the period is illustrated in Figure 2 below:

If we plot the data from a time series with positive exponential growth, the observations will form a convex curve like in Figure 2. Negative exponential growth means that the observed values of the series decrease at a constant rate, so the time series forms a concave curve.

Figure 2: ABC Company Quarterly Sales



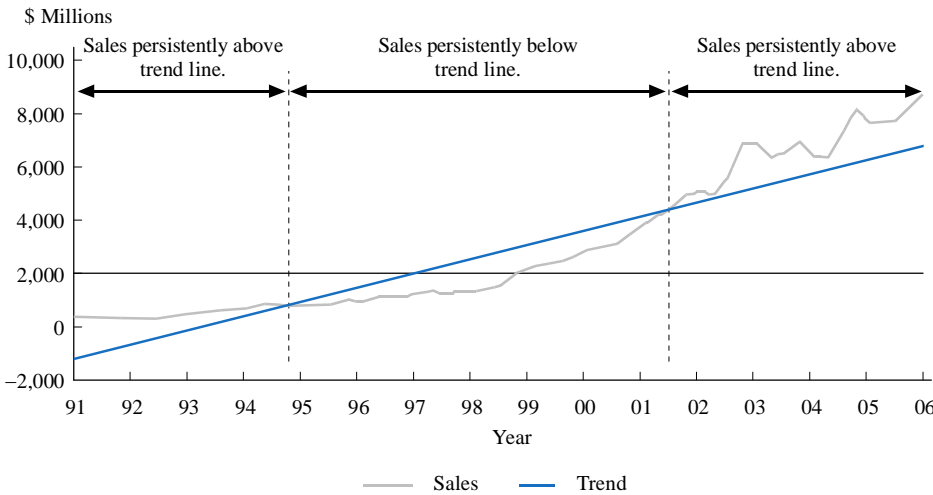
Initially, Nasri uses a linear trend model to capture the data. The results from the regression are presented in Table 2:

Table 2: Estimating a Linear Trend for ABC Company Sales

Regression Statistics			
R-squared	0.8443		
Standard error	786.32		
Observations	60		
Durbin-Watson	0.15		
	Coefficient	Standard Error	t-Stat
Intercept	-1,212.46	335.8417	-3.6102
Trend	125.3872	6.3542	19.733

The results of the regression seem to support the use of a linear trend model to fit the data. The absolute values of the t-stats of both the intercept and the trend coefficient (-3.61 and 19.73 respectively) appear statistically significant as they exceed the critical t-value of 2.0 ($\alpha = 0.05$, $df = 58$). However, when quarterly sales are plotted along with the trend line (Figure 3), the errors seem to be persistent (the residuals remain above or below the trend line for an extended period of time), which suggests that they are positively serially correlated. The persistent serial correlation in the residuals makes the linear regression model inappropriate (even though the R^2 is quite high at 84.43%) to fit ABC's sales as it violates the regression assumption of uncorrelated residual errors.

Figure 3: ABC Company Quarterly Sales with Trend



Since the sales data plot (Figure 2) is curved upwards, Nasri’s supervisor suggests that he use the log-linear model. Nasri then estimates the following log-linear regression equation:

$$\ln y_t = b_0 + b_1t + \varepsilon_t, \quad t = 1, 2, \dots, 60$$

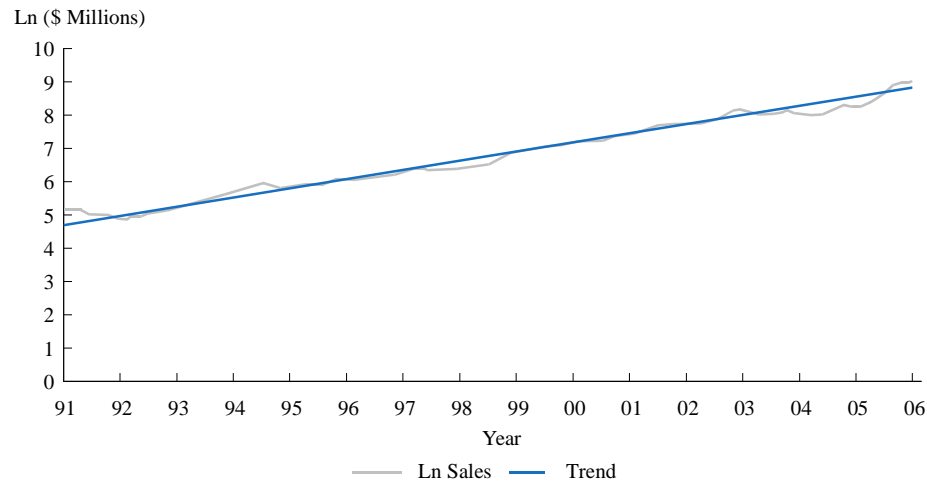
Table 3 presents the log-linear regression results.

Table 3: Estimating a Linear Trend in Lognormal ABC Company Sales

Regression Statistics			
R-squared	0.9524		
Standard error	0.1235		
Observations	60		
Durbin-Watson	0.37		
	Coefficient	Standard Error	t-Stat
Intercept	4.6842	0.0453	103.404
Trend	0.0686	0.0008	85.75

Notice that the R^2 (95.24%) is now much higher than in Table 2 (linear trend model regression results where the R^2 was 84.43%). This suggests that the log-linear model fits the sales data much better than the linear trend model. Figure 4 plots the **linear trend line** suggested by the log-linear regression along with the **natural logs of the sales data**. Notice that the vertical distances between the lines are quite small, and that the residuals are not persistent (log actual sales are not above or below the trend line for an extended period of time). Consequently, Nasri concludes that the log-linear trend model is more suitable for modeling ABC’s sales compared to the linear trend model.

An R^2 of 0.9524 means that 95.24% of the variation in the natural log of ABC’s sales is explained solely by a linear trend.

Figure 4: Natural Log of ABC Company Quarterly Sales

To illustrate how log-linear trend models are used in making forecasts, let's calculate ABC's expected sales for Q3 2006, or Quarter 63 (an out-of-sample forecast).

$$\ln \hat{y} = 4.6842 + 0.0686(63)$$

$$\hat{y}_{63} = \$8,151.849 \text{ million}$$

Compared to the forecast for Quarter 63 sales based on the linear trend model $(-1,212.46 + 125.3872(63) = \$6,686.93 \text{ million})$ the log linear regression model offers a much higher forecast.

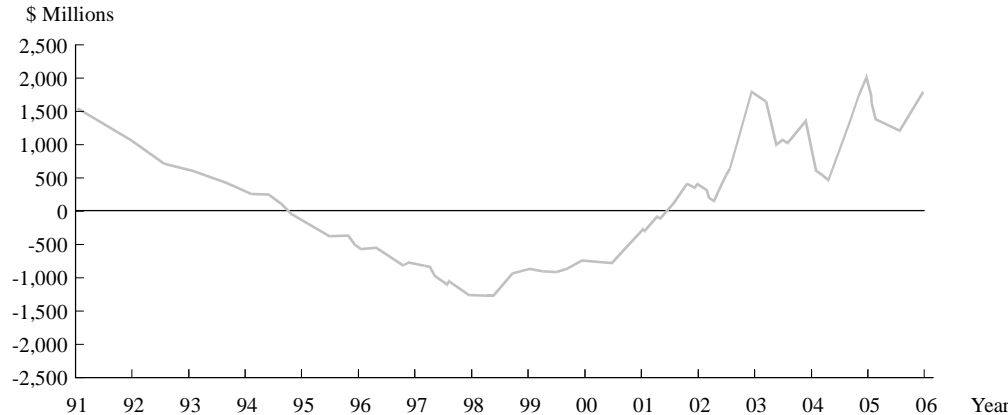
An important difference between the linear and log-linear trend models lies in the interpretation of the slope coefficient, b_1 .

- A linear trend model predicts that y_t will grow by a **constant amount** (b_1) each period. For example, if b_1 equals 0.1%, y_t will grow by 0.1% in each period.
- A log-linear trend model predicts that $\ln y_t$ will grow by a constant amount (b_1) in each period. This means that y_t itself will witness a **constant growth rate** of $e^{b_1} - 1$ in each period. For example, if b_1 equals 0.1% then the predicted growth rate of y_t in each period equals $e^{0.01} - 1 = 0.01005$ or 1.005%.

Also, in a linear trend model the predicted value of y_t is $\hat{b}_0 + \hat{b}_1 t$, but in a log-linear trend model the predicted value of y_t is $e^{\hat{b}_0 + \hat{b}_1 t}$ because $e^{\ln y_t} = y_t$.

Testing for Correlated Errors in Trend Models

If a regression model is correctly specified, the regression error for one time period will be uncorrelated with the regression errors for other time periods. One way to determine whether the error terms are correlated across time periods is to inspect the plot of residuals. Figure 5 plots the residuals when a **linear trend** is used to model ABC Company's sales. The figure clearly indicates that there is persistent serial correlation in the residuals of the model.

Figure 5: Residual from Predicting ABC Company Sales with a Linear Trend

However, a more formal test to determine whether the regression errors are serially correlated is the [Durbin-Watson \(DW\) test](#).

The DW stat for the **log-linear trend model** (Table 3 in Example 2) equals 0.37. To test the null hypothesis of no positive serial correlation in the residuals at the 5% level of significance, the critical value (d_1) equals 1.55 ($k = 1, n = 60$). Since the value of the DW stat is less than d_1 , we reject the null hypothesis and conclude that the log-linear model does suffer from positive serial correlation. Consequently, we need to build a different kind of model to represent the relation between time and ABC Company's sales.

Figure 5 shows us that the errors of the linear trend model are serially correlated.

The Durbin-Watson test shows us that the errors of the log-linear trend are also serially correlated.

Existence of serial correlation suggests that we can build better forecasting models than trend models to fit the series.

LOS 13c: Explain the requirement for a time series to be covariance stationary, and describe the significance of a series that is not stationary.

Vol 1, pg 450-451

LOS 13d: Describe the structure of an autoregressive (AR) model of order p , and calculate one- and two-period-ahead forecasts given the estimated coefficients. Vol 1, pg 450-459

LOS 13e: Explain how autocorrelations of the residuals can be used to test whether the autoregressive model fits the time series. Vol 1, pg 450-459

LOS 13f: Explain mean reversion, and calculate a mean-reverting level. Vol 1, pg 455

AUTOREGRESSIVE (AR) TIME-SERIES MODELS

An autoregressive (AR) model is a time series that is regressed on its own past values. Since the same variable essentially shows up on both sides of the equation (as a dependent and an independent variable), we drop the normal notation of y_t as the dependent variable and only use x_t . For example, an AR(1) model (first-order autoregressive model) is represented as:

$$x_t = b_0 + b_1x_{t-1} + \varepsilon_t$$

Note: In an AR(1) model, only the single most-recent past value of x_t is used to predict the current value of x_t .

A p th order autoregressive model is represented as:

$$x_t = b_0 + b_1x_{t-1} + b_2x_{t-2} + \dots + b_px_{t-p} + \varepsilon_t$$

Note: An AR(p) model uses p past values of x_t to predict the current value of x_t .

Covariance Stationary Series

When an independent variable in the regression equation is a lagged value of the dependent variable (as is the case in autoregressive time series models) statistical inferences based on OLS regression are not always valid. In order to conduct statistical inference based on these models, we must assume that the time series is **covariance stationary** or **weakly stationary**. There are three basic requirements for a time-series to be covariance stationary:

1. The expected value or mean of the time series must be constant and finite in all periods.
2. The variance of the time series must be constant and finite in all periods.
3. The covariance of the time series with itself for a fixed number of periods in the past or future must be constant and finite in all periods.

If an AR model is used to model a time series that is not covariance stationary, the analyst would obtain biased estimates of the slope coefficient(s), which would render the results of any hypothesis tests invalid. Such a model would only yield spurious results.

One way to determine whether a time series is covariance stationary is by looking at a graphical plot of the data. The inflation data in Example 1 appears to be covariance stationary (see Figure 1). The data seem to have the same mean and variance over the sample period. On the other hand, ABC Company's quarterly sales appear to grow steadily over time, which implies that the mean is not constant and therefore, the series is not covariance stationary (see Figure 2).

Other (more sophisticated) ways to determine whether a time series is covariance stationary are presented under LOS 13j. If a time series is not covariance stationary, there is a way to convert it into a stationary time series. This method (**first differencing**) is also demonstrated later in the reading.

Detecting Serially Correlated Errors in an AR Model

An AR model can be estimated using ordinary least squares if (1) the time series is covariance stationary and (2) the errors are uncorrelated. Tests for stationarity are discussed later in the reading (**examining that time series autocorrelations at various lags** and the **unit-root test**), but first let's discuss how we can test whether the residuals are serially correlated. The Durbin-Watson test cannot be used to test for serial correlation in an AR model because the independent variables include past values of the dependent variable. However, another test based on the **autocorrelations of the error term** can be used to determine if the errors in the AR time series model are serially correlated.

Aside from being covariance stationary and having uncorrelated residuals, an appropriately-specified AR model should have homoskedastic (not heteroskedastic) errors. We learn to test the residuals of an AR model for heteroskedasticity in LOS 13m

Note that this test for serial correlation focuses on the **autocorrelations of the error term**, which are different from the autocorrelations of the time series itself.

We determine whether the residuals of the time series model are serially correlated by testing whether the **autocorrelations of the error terms** (**error autocorrelations** or **residual autocorrelations**) are significantly different from 0.

- If any of the error autocorrelations are significantly different from 0, the errors are serially correlated and the model is not specified correctly.
- If all the error autocorrelations are not significantly different from 0, the errors are not serially correlated and the model is specified correctly.

To determine whether an error autocorrelation for a particular lag is significantly different from 0, we perform a t-test, where the t-stat is calculated as the error autocorrelation for that particular lag divided by the standard error of the residual autocorrelation (which equals $1/\sqrt{T}$).

$$\text{t-stat} = \frac{\text{Residual autocorrelation for lag}}{\text{Standard error of residual autocorrelation}}$$

where:

Standard error of residual autocorrelation = $1/\sqrt{T}$

T = Number of observations in the time series

There are three basic steps (illustrated in Example 3) for detecting serially correlated errors in an AR time-series model:

1. Estimate a particular AR model.
2. Compute the autocorrelations of the residuals from the model.
3. Determine whether the residual autocorrelations significantly differ from 0.

Example 3: Testing whether an AR Time Series Model has Serially Correlated Errors

Jack Wilshire uses a time-series model to predict ABC Company's gross margins. He uses quarterly data from Q1 1981 to Q4 1995. Since he believes that the gross margin in the current period is dependent on the gross margin in the previous period, he starts with an AR(1) model:

$$\text{Gross margin}_t = b_0 + b_1(\text{Gross margin}_{t-1}) + \varepsilon_t$$

Table 4 presents the results from estimating the AR(1) model while Table 5 presents the autocorrelations of the residuals from the model.

Table 4: AR(1) Model Regression Results

Regression Statistics				
R-squared	0.7521			
Standard error	0.0387			
Observations	60			
Durbin-Watson	1.9132			
	Coefficient	Standard Error	t-Stat	
Intercept	0.0795	0.0352	2.259	
Lag 1	0.8524	0.0602	14.159	

Table 5: Autocorrelations of the Residuals from the AR(1) Model

Lag	Coefficient	Standard	
		Error	t-Stat
1	0.0583	0.1291	0.4516
2	0.0796	0.1291	0.6166
3	-0.1921	0.1291	-1.4880
4	-0.1285	0.1291	-0.9954

The first lag of a time series is the value of the time series in the previous period.

From Table 4 notice that the intercept ($\hat{b}_0 = 0.0795$) and the coefficient on the first lag ($\hat{b}_1 = 0.8524$) are highly significant in this regression. The t-stat of the intercept (2.259) and that of the coefficient on the first lag of the gross margin (14.159) are both greater than the critical t-value at the 5% significance level with 58 degrees of freedom ($t_c = 2.0$).

Even though Wilshire concludes that both the regression coefficients individually do not equal 0 (or are statistically significant), he must still evaluate the validity of the model by determining whether the residuals from his model are serially correlated. Since this is an AR model (the independent variables include past values of the dependent variable) the Durbin-Watson test for serial correlation cannot be used.

Table 5 lists the first four autocorrelations of the residual along with their standard errors and t-statistics. Since there are 60 observations, the standard error for each of the residual autocorrelations equals 0.1291 (calculated as $1/\sqrt{60}$). None of the t-stats in Table 5 is greater than 2.0 (critical t-value) in absolute value, which indicates that none of the residual autocorrelations significantly differs from 0. Wilshire concludes that the regression residuals are not serially correlated and that his AR(1) model is correctly specified. Therefore, he can use ordinary least squares to estimate the parameters and the standard errors of the parameters in his model.

Note that if any of the lag autocorrelations were significantly different from zero (if they had t-stats that were greater than the critical t-value in absolute value) the model would be misspecified due to serial correlation between the residuals. If the residuals of an AR model are serially correlated, the model can be improved by adding more lags of the dependent variable as explanatory (independent) variables. More and more lags of the dependent variable must be added as independent variables in the model until all the residual autocorrelations are insignificant.

Once it has been established that the residuals are not serially correlated and that the model is correctly specified, it can be used to make forecasts. The estimated regression equation in this example is given as:

$$\text{Gross margin}_t = 0.0795 + 0.8524(\text{Gross margin}_{t-1})$$

From the regression equation, note that:

- If the gross margin is currently 50%, the model predicts that next quarter's gross margin will *increase* to 0.5057 or 50.57%.
- If the gross margin is currently 60%, the model predicts that next quarter's gross margin will *decrease* to 0.5909 or 59.09%.

As we will learn in the next section, the model predicts an increase in the gross margin during a particular quarter if the gross margin in the previous quarter was less than 53.86%, and a decrease in the gross margin during a particular quarter if the gross margin in the previous quarter was greater than 53.86%.

Mean Reversion

A time series is said to exhibit **mean reversion** if it tends to fall when its current level is above the mean and tends to rise when its current level is below the mean. The mean-reverting level, x_t , for a time series is given as:

$$x_t = \frac{b_0}{1 - b_1}$$

- If a time series is currently at its mean-reverting level, the model predicts that its value will remain unchanged in the next period.
- If a time series is currently above its mean-reverting level, the model predicts that its value will decrease in the next period.
- If a time series is currently below its mean-reverting level, the model predicts that its value will increase in the next period.

In the case of gross margins for ABC Company (Example 3), the mean reverting level is calculated as $0.0795/(1 - 0.8524) = 0.5386$ or 53.86%.

Important: All covariance stationary time series have a finite mean-reverting level. An AR(1) time series will have a finite mean-reverting level if the absolute value of the lag coefficient, b_1 , is less than 1.

Multiperiod Forecasts and the Chain Rule of Forecasting

The **chain rule of forecasting** is used to make multi-period forecasts based on an autoregressive time series model. For example, a one-period forecast (\hat{x}_{t+1}) based on an AR(1) model is calculated as:

$$\hat{x}_{t+1} = \hat{b}_0 + \hat{b}_1 x_t$$

Using this one-period forecast (\hat{x}_{t+1}), the two-period forecast is calculated as:

$$\hat{x}_{t+2} = \hat{b}_0 + \hat{b}_1 \hat{x}_{t+1}$$

Since we do not know x_{t+1} in period t , we must start by forecasting x_{t+1} using x_t as an input and then forecast x_{t+2} using our forecast of x_{t+1} as an input.

An AR(1) time series is said to have a unit root if b_1 equals 1, and is said to have an explosive root if b_1 is greater than 1. Only if the time series has a finite mean-reverting level ($b_1 < 1$) can standard regression analysis be applied to estimate an AR(1) model to fit the series. More on this in LOS 13 i, j and k.

Note that multi-period forecasts entail more uncertainty than single-period forecasts because each period's forecast (used as an input to eventually arrive at the multi-period forecast) entails uncertainty. Generally speaking, the more periods a forecast has, the greater the uncertainty.

Example 4: Chain Rule of Forecasting

Assume that ABC Company's gross margin for the current quarter is 65%. Using the AR(1) model in Example 3, forecast ABC's gross margin in two quarters.

Solution:

First we forecast next quarter's gross margin based on the current quarter's gross margin:

$$\text{Gross margin}_{t+1} = 0.0795 + 0.8524(\text{Gross margin}_t) = 0.0795 + 0.8524(0.65) = 0.6336 \text{ or } 63.36\%$$

Then we forecast the gross margin in two quarters based on next period's gross margin forecast:

$$\text{Gross margin}_{t+2} = 0.0795 + 0.8524(\text{Gross margin}_{t+1}) = 0.0795 + 0.8524(0.6336) = 0.6196 \text{ or } 61.96\%$$

Notice that since x_t and x_{t+1} are greater than the mean reverting level (0.5386), the value of the series falls in subsequent periods.

LOS 13g: Contrast in-sample and out-of-sample forecasts, and compare the forecasting accuracy of different time-series models based on the root mean squared error criterion. Vol 1, pg 459-461

Comparing Forecast Model Performance

One way to evaluate the forecasting performance of two models is by comparing their standard errors. The standard error for the time series regression is typically reported in the statistical output for the regression. The model with the smaller standard error will be more accurate as it will have a smaller forecast error variance (s_f^2).

When comparing the forecasting performance of various models, analysts distinguish between **in-sample forecast errors** and **out-of-sample forecast errors**. In-sample forecast errors are differences between the actual values of the dependent variable and predicted values of the dependent variable (based on the estimated regression equation) **for data from within the sample period**. In essence, in-sample forecasts are the residuals from a fitted time-series model. For instance, in Example 1, the residuals of the regression (differences between actual inflation and forecasted inflation for the months lying in the January 1981-December 2010 sample period) represent in-sample forecast errors. If we were to predict inflation for a month **outside the sample period** (e.g. July 2012) based on this model, the difference between actual and predicted inflation would represent an out-of-sample forecast error. Out-of-sample forecasts are important in evaluating the model's contribution and applicability in the real world.

The out-of-sample forecasting performance of autoregressive models is evaluated on the basis of their **root mean square error (RMSE)**. The RMSE for each model under consideration is calculated based on out-of-sample data. The model with the lowest RMSE has the lowest forecast error and hence carries the most predictive power.

For example, consider a data set that includes 35 observations of historical annual unemployment rates. Suppose we considered only the first 30 years as the sample period in developing our time series models, and we came up with an AR(1) and an AR(2) model to fit the 30-year unemployment data. The remaining 5 years of data from Year 31 to Year 35 (the out-of-sample data) would be used to calculate the RMSE for the two models, and the model with the lower RMSE would be judged to have greater predictive power. Bear in mind that a model with the lower RMSE (more accuracy) for in-sample data will not necessarily have a lower RMSE for out-of-sample data.

In addition to the forecast accuracy of a model, the stability of the regression coefficients (discussed in the next LOS) is an important consideration when evaluating a model.

LOS 13h: Explain instability of coefficients of time-series models.

Vol 1, pg 461-464

Instability of Regression Coefficients

The choice of sample period is a very important consideration when constructing time series models. This is because:

- Regression estimates from time series models based on different sample periods can be quite different.
- Regression estimates obtained from models based on longer sample periods can be quite different from estimates from models based on shorter sample periods.

There are no clear-cut rules that define an ideal length for the sample period. Based on the fact that models are only valid if the time series is covariance stationary, analysts look to define sample periods as times during which important underlying economic conditions have remained unchanged. For example, data from a period when exchange rates were fixed should not be combined with data from a period when they were floating as the variance of the exchange rate would be different under the two regimes. Usually, analysts look at graphs of the data to see if the series looks stationary. If there has been a significant shift in governmental policy during the period, analysts use their experience and judgement to determine whether the time series relation has remained the same before and after the shift. If the relation has changed and the series is not covariance stationary, models based on the data will not be valid.

The point here is that even if the autocorrelations of the residuals of a time series model are statistically insignificant, analysts cannot conclude that the sample period used is appropriate (and hence deem the model valid) until they are, at the same time, confident that the series is covariance stationary and that important external factors have remained constant during the sample period used in the study.

LOS 13i: Describe characteristics of random walk processes, and contrast them to covariance stationary processes. Vol 1, pg 464-468

Random Walks

A **random walk**, **simple random walk** or **random walk without a drift** is a time series in which the value of the series in one period equals its value in the previous period plus an unpredictable random error, where the error has a constant variance and is uncorrelated with its value in previous periods.

$$x_t = x_{t-1} + \varepsilon_t, E(\varepsilon_t) = 0, E(\varepsilon_t^2) = \sigma^2, E(\varepsilon_t \varepsilon_s) = 0 \text{ if } t \neq s$$

It is important to note the following regarding random walks:

- The random walk equation is a special case of the AR(1) model where b_0 equals 0, and b_1 equals 1.
- The best forecast of x_t is essentially x_{t-1} as the expected value of the error term is 0.

Standard regression analysis cannot be applied to estimate an AR(1) model for a time series that follows a random walk. Statistical conclusions based on such a model would be incorrect because AR models cannot be used to model any time series that is not covariance stationary. Random walks are not covariance stationary as:

- They do not have a finite mean-reverting level. For a random walk, the mean reverting level is undefined. $b_0/(1 - b_1) = 0/(1-1) = 0/0$
- They do not have a finite variance. As t increases, the variance of x_t grows with no upper bound (it approaches infinity).

Fortunately, a random walk can be converted to a covariance stationary time series. This is done through **first differencing**, which subtracts the value of the time series in the previous period from its value in the current period. The new time series, y_t , is calculated as x_t minus x_{t-1} . The first difference of the random walk equation is given as:

$$y_t = x_t - x_{t-1} = x_{t-1} + \varepsilon_t - x_{t-1} = \varepsilon_t, E(\varepsilon_t) = 0, E(\varepsilon_t^2) = \sigma^2, E(\varepsilon_t \varepsilon_s) = 0 \text{ for } t \neq s$$

By using the first-differenced time series, we are essentially modeling the change in the value of the dependent variable ($\Delta x_t = x_t - x_{t-1}$) rather than the value of the variable itself (x_t). From the first differenced random walk equation, note that:

- Since the expected value of the error term is 0, the best forecast of y_t is 0, which implies that there will be no change in the value of the current time series, x_{t-1} .
- y_t , the first-differenced variable, is covariance stationary with a finite mean-reverting level of 0 (calculated as $0/(1-0)$) as b_0 and b_1 both equal 0, and a finite variance ($\text{Var}(\varepsilon_t) = \sigma^2$).
- Therefore, we can use linear regression to model the first-differenced series.

Modeling the first-differenced time series with an AR(1) model does not hold predictive value (as b_0 and b_1 both equal 0). It only serves to confirm a suspicion that the original time series is indeed a random walk.

By definition, changes in a random walk (y_t or $x_t - x_{t-1}$) are unpredictable.

Example 5: Determining whether a Time Series is a Random Walk

Aaron Ramsey develops the following AR(1) model for the Japanese Yen/USD exchange rate (x_t) based on monthly observations over 30 years.

$$x_t = x_{t-1} + \varepsilon_t$$

Table 6 contains data relating to his AR(1) model.

Table 6: AR(1) Model for JPY/USD Exchange Rate

Regression Statistics			
R-squared	0.9852		
Standard error	5.8623		
Observations	360		
Durbin-Watson	1.9512		
	Coefficient	Standard Error	t-Stat
Intercept	1.0175	0.9523	1.0685
Lag 1	0.9954	0.0052	191.42

Autocorrelations of the Residuals from the AR(1) Model

Lag	Coefficient	Standard Error	t-Stat
1	0.0745	0.0527	1.4137
2	0.0852	0.0527	1.6167
3	0.0321	0.0527	0.6091
4	0.0525	0.0527	0.9962

Notice the following:

- The intercept term is not significantly different from 0. The low t-stat of 1.06 does not allow you to reject the null hypothesis that the intercept term equals 0 as it is less than the critical t-value of 1.972 at the 5% significance level.
- The coefficient on the first lag of the exchange rate is significantly different from 0 (high t-stat of 191.42 allows you to reject the null hypothesis that it equals 0) and is actually very close to 1 (coefficient on the first lag of the time series equals 0.9954).

However, we cannot use the t-stats in Table 6 to determine whether the exchange rate is a random walk (by conducting a hypothesis test on $H_0: b_0 = 0$ and $H_0: b_1 = 1$) because the standard errors of this AR model would be invalid if the model is based on a time series that is not covariance stationary. Recall that a random walk is not covariance stationary.

In order to determine whether the time series is indeed a random walk, we must run a regression on the first-differenced time series. If the exchange rate is, in fact, a random walk then:

1. The first-differenced time series will be covariance stationary as b_0 and b_1 would equal 0; and
2. The error term will not be serially correlated.

Table 7 presents the regression results for the first-differenced AR(1) model for the JPY/USD exchange rate.

Table 7: First-Differenced AR(1) Model for JPY/USD Exchange Rate

Regression Statistics			
R-squared	0.0052		
Standard error	5.8751		
Observations	360		
Durbin-Watson	1.9812		
	Coefficient	Standard Error	t-Stat
Intercept	-0.4855	0.3287	-1.477
Lag 1	0.0651	0.0525	1.240

Autocorrelations of the Residuals from the AR(1) Model

Lag	Coefficient	Standard Error	t-Stat
1	0.0695	0.0527	1.3188
2	-0.0523	0.0527	-0.9924
3	0.0231	0.0527	0.4383
4	0.0514	0.0527	0.9753

From Table 7, notice that:

- The intercept term (b_0) and the coefficient on the first lag of the first-differenced exchange rate (b_1) both individually do not significantly differ from 0. The absolute values of their t-stats (1.477 and 1.24 respectively) are lower than the absolute value of t_{crit} (1.972) at the 5% significance level.
- None of the residual autocorrelations significantly differs from 0. All the t-stats are lower than the critical t-value (1.972).

Since b_0 and b_1 for the first-differenced AR(1) model both equal 0, and there is no serial correlation in the error terms of the first-differenced time series, we can conclude that the JPY/USD exchange rate is a random walk.

Just one minor point before we move ahead. The R^2 in Table 6 for the AR(1) model on the original time series ($R^2 = 0.9852$) is much higher than the R^2 in Table 7 for the AR(1) model on the first-differenced time series ($R^2 = 0.0052$). If we were to base our choice of model on R^2 alone, we would make the incorrect choice and go with the AR(1) model on the original time series (which is not covariance stationary) instead of the AR(1) model on the first-differenced time series (which actually is covariance stationary). The interpretations of the R^2 's of the two models are fundamentally different:

- The R^2 in Table 6 measures how well the exchange rate in one period predicts the exchange rate in the next period. If the exchange rate is a random walk, this number should be extremely high (which it is).

- The R^2 in Table 7 measures how well the change in the exchange rate in one period predicts the change in the exchange rate in the next period. If the exchange rate is a random walk, changes in the exchange rate should be completely unpredictable and this number should be relatively low (which it is).

Random Walk with a Drift

A **random walk with a drift** is a time series that increases or decreases by a constant amount in each period. The equation for a random walk with a drift is given as:

$$\begin{aligned} x_t &= b_0 + b_1 x_{t-1} + \varepsilon_t \\ b_1 &= 1, b_0 \neq 0, \text{ or} \\ x_t &= b_0 + x_{t-1} + \varepsilon_t, E(\varepsilon_t) = 0 \end{aligned}$$

Unlike a simple random walk (which has $b_0 = 0$), a random walk with a drift has $b_0 \neq 0$. Similar to a simple random walk, a random walk with a drift also has an undefined mean-reverting level (because $b_1 = 1$) and is therefore, not covariance stationary. Consequently, an AR(1) model cannot be used to analyze a random walk with a drift without first-differencing it. The first-difference of the random walk with a drift equation is given as:

$$y_t = x_t - x_{t-1}, y_t = b_0 + \varepsilon_t, b_0 \neq 0$$

LOS 13j: Describe implications of unit roots for time-series analysis, explain when unit roots are likely to occur and how to test for them, and demonstrate how a time series with a unit root can be transformed so it can be analyzed with an AR model. Vol 1, pg 468-472

LOS 13k: Describe the steps of the unit root test for nonstationarity, and explain the relation of the test to autoregressive time-series models. Vol 1, pg 468-472

The Unit Root Test of Nonstationarity

When we introduced covariance stationarity earlier in this reading (under LOS 13c), we stated that one way to determine whether a time series is covariance stationary is by examining a graph that plots the data. There are two (more sophisticated) ways to determine whether a time series is covariance stationary:

1. **Examine the autocorrelations of the time series at various lags.** For a stationary time series, either the time series autocorrelations at all lags do not significantly differ from 0, or the autocorrelations drop off abruptly to 0 as the number of lags becomes large. For a nonstationary time series, the time series autocorrelations do not exhibit either of these characteristics.

Here we are not talking about the residual autocorrelations (which are used to test for serial correlation as in Example 3), but are referring to the autocorrelations of the **actual time series**.

2. **Conduct the Dicky-Fuller test for unit root (preferred approach).** A time series is said to have a unit root when the estimated value of the lag coefficient equals 1. A time series that has a unit root is a random walk, which is not covariance stationary. As we have mentioned before, for statistical reasons, simple t-tests cannot be used to test whether the coefficient on the first lag of the time series in an AR(1) model is significantly different from 1. However, the Dicky-Fuller test can be used to test for a unit root.

The Dicky-Fuller test starts by converting the lag coefficient, b_1 , in a simple AR(1) model into g_1 , which effectively represents $b_1 - 1$, by subtracting x_{t-1} from both sides of the AR(1) equation:

$$\begin{aligned}x_t &= b_0 + b_1x_{t-1} + \varepsilon_t \\x_t - x_{t-1} &= b_0 + b_1x_{t-1} - x_{t-1} + \varepsilon_t \\x_t - x_{t-1} &= b_0 + (b_1 - 1)x_{t-1} + \varepsilon_t \\x_t - x_{t-1} &= b_0 + g_1x_{t-1} + \varepsilon_t\end{aligned}$$

We have already introduced first differencing in an earlier LOS.

Note that the dependent variable ($x_t - x_{t-1}$) is first difference of the time series and the independent variable (x_{t-1}) is the first lag of the time series.

- The null hypothesis for the Dicky-Fuller test is that $g_1 = 0$ (effectively means that $b_1 = 1$) and that the time series has a unit root, which makes it nonstationary.
- The alternative hypothesis for the Dicky-Fuller test is that $g_1 < 0$, (effectively means that $b_1 < 1$) and that the time series is covariance stationary (i.e., it does not have a unit root).
- The t-stat for the Dicky-Fuller test is calculated in the same manner that we have been using in the reading so far, but the critical values used in the test are different. Dicky-Fuller critical values are larger in absolute value than conventional critical t-values.

Example 6: Using First Differenced Data to Make Forecasts

Samir Nasri (the analyst from Example 2) is convinced, after looking at Figures 2 and 4, that the logs of ABC Company's quarterly sales do not represent a covariance stationary time series. He therefore, first differences the log of ABC's quarterly sales.

Figure 6: Log Difference of ABC Company Quarterly Sales



Figure 6 shows that the first-differenced series does not exhibit any strong trend and appears to be covariance stationary. He therefore decides to model the first-differenced time series as an AR(1) model:

$$\ln(\text{Sales}_t) - (\ln \text{Sales}_{t-1}) = b_0 + b_1[\ln(\text{Sales}_{t-1}) - (\ln \text{Sales}_{t-2})] + \varepsilon_t$$

Table 8 shows the results of the regression:

Table 8: Log Differenced Sales: AR(1) Model for ABC Company Quarterly Sales

Regression Statistics			
R-squared	0.1065		
Standard error	0.0617		
Observations	60		
Durbin-Watson	1.9835		
	Coefficient	Standard Error	t-Stat
Intercept	0.0485	0.0152	3.1908
Lag 1	0.3728	0.1324	2.8158

Autocorrelations of the Residuals from the AR(1) Model

Lag	Coefficient	Standard Error	t-Stat
1	-0.0185	0.1291	-0.1433
2	-0.0758	0.1291	-0.5871
3	-0.0496	0.1291	-0.3842
4	0.2026	0.1291	1.5693

From Table 8 notice the following:

- At the 5% significance level, both the regression coefficients ($\hat{b}_0 = 0.0485$, $\hat{b}_1 = 0.3728$) of the first-differenced series are statistically significant as their t-stats (3.19 and 2.82 respectively) are greater than t_{crit} (2.00) with $df = 58$.
- The four autocorrelations of the residuals are statistically insignificant. Their t-stats are smaller in absolute value than t_{crit} so we fail to reject the null hypotheses that each of the residual autocorrelations equals 0. We therefore conclude that there is no serial correlation in the residuals of the regression.

These results suggest that the model is correctly specified and can be used to make predictions of ABC Company's quarterly sales. The value of the intercept ($\hat{b}_0 = 0.0485$) indicates that if sales have not changed in the current quarter ($\ln \text{Sales}_t - \ln \text{Sales}_{t-1} = 0$) sales will grow by 4.85% in the next quarter ($\ln \text{Sales}_{t+1} - \ln \text{Sales}_t$). If sales have changed in the current quarter, the slope coefficient ($\hat{b}_1 = 0.3728$) tells us that in the next quarter, sales will grow by 4.85% plus 0.3728 times sales growth in the current quarter.

Suppose we want to predict sales for the first quarter of 2006 based on the first-differenced model. We are given the following pieces of information:

Sales Q4 2005 = Sales_t = \$8,157m
 Sales Q3 2005 = Sales_{t-1} = \$7,452m
 Sales Q1 2006 = Sales_{t+1} = ?

Our regression equation is given as:

$$\ln \text{Sales}_t - \ln \text{Sales}_{t-1} = 0.0485 + 0.3728 (\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2})$$

Therefore:

$$\begin{aligned} \ln \text{Sales}_{t+1} - \ln \text{Sales}_t &= 0.0485 + 0.3728 (\ln \text{Sales}_t - \ln \text{Sales}_{t-1}) \\ \ln \text{Sales}_{t+1} - \ln 8,157 &= 0.0485 + 0.3728 (\ln 8,157 - \ln 7,452) \\ \ln \text{Sales}_{t+1} &= 0.0485 + (0.3728)(0.0904) + 9.0066 \\ \ln \text{Sales}_{t+1} &= 9.0888 \\ \text{Sales}_{t+1} &= \$8,855.56\text{m} \end{aligned}$$

Therefore, based on Q4 2005 sales of \$8,157m the model predicts that ABC's sales in Q1 2006 would be \$8,855.56m.

Moving-Average Time Series Models Vol 1, pg 472-477

Smoothing Past Values with Moving Averages

Moving averages are generally calculated to eliminate the 'noise' from a time series in order to focus on the underlying trend. An n -period moving average is based on the current value and previous $n - 1$ values of a time series. It is calculated as:

$$\frac{x_t + x_{t-1} + \dots + x_{t-(n-1)}}{n}$$

One of the weaknesses of the moving average is that it always lags large movements in the underlying data. Further, even though moving averages are useful in smoothing out a time series, they do not hold much predictive value (as they give equal weight to all observations). In order to enhance the forecasting performance of moving averages, analysts use moving-average time series models.

Moving Average Time Series Models for Forecasting

A moving average (MA) model of order 1 is given as:

$$x_t = \varepsilon_t + \theta\varepsilon_{t-1}, E(\varepsilon_t) = 0, E(\varepsilon_t^2) = \sigma^2, E(\varepsilon_t\varepsilon_s) = 0 \text{ for } t \neq s$$

x_t is a moving average of ε_t and ε_{t-1} , and ε_t and ε_{t-1} are uncorrelated random variables that have an expected value of 0. Note that in contrast to the simple moving average model equation (where all observations receive an equal weight) this moving-average model attaches a weight of 1 on ε_t and a weight of θ on ε_{t-1} .

An MA(q) moving average model (a moving average model of order q) is given as:

$$x_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \dots + \theta_q\varepsilon_{t-q}, \quad E(\varepsilon_t) = 0, \quad E(\varepsilon_t^2) = \sigma^2, \quad E(\varepsilon_t\varepsilon_s) = 0 \text{ for } t \neq s$$

To determine whether a time series follows an MA(q) model we examine the autocorrelations of the **original time series** (not the residual autocorrelations that we examine in AR models to determine whether serial correlation exists). For an MA(q) model, the first q autocorrelations will be significant, and all the autocorrelations beyond that will equal 0.

The **time series autocorrelations** can also be used to determine whether an autoregressive or a moving average model is more appropriate to fit the data.

- For most AR models, the time series autocorrelations start out large and then decline gradually.
- For MA models, the first q time series autocorrelations are significantly different from 0, and then suddenly drop to 0 beyond that.

Note that most time series are best modeled with AR models.

LOS 13I: Explain how to test and correct for seasonality in a time-series model, and calculate and interpret a forecasted value using an AR model with a seasonal lag. Vol 1, pg 477-482

Seasonality in Time Series Models

A common problem in time series is **seasonality**. Seasonality may cause us to incorrectly conclude that an autoregressive model is inappropriate to model a particular time series (as seasonality may cause one or more of the residual autocorrelations to be significantly different from 0). To detect seasonality in the time series, we examine the **autocorrelations of the residuals** to determine whether the **seasonal autocorrelation** of the error term is significantly different from 0. The seasonal error autocorrelation corresponds to the **seasonal lag**, which is the value of the time series one year before the current period. For example, if we are working with monthly data, the seasonal lag would be the twelfth lag of the series.

To correct for seasonality, we simply add a seasonal lag to the AR model. Example 7 illustrates the processes of detecting seasonality in a time series, correcting for it and making forecasts once a seasonal lag has been added to the model.

Example 7: Seasonality in a Time Series

Robin Van Persie estimates an AR(1) model based on first-differenced sales data to model XYZ Company's quarterly sales for for 10 years from Q1 1991 to Q4 2000. He comes up with the following regression equation:

$$\ln \text{Sales}_t - \ln \text{Sales}_{t-1} = b_0 + b_1(\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2}) + \varepsilon_t$$

Table 9 presents the results of the regression.

Table 9: AR(1) Model on Log First Differenced Quarterly Sales

Regression Statistics			
R-squared	0.1763		
Standard error	0.0751		
Observations	40		
Durbin-Watson	2.056		
	Coefficient	Standard Error	t-Stat
Intercept	0.0555	0.0087	6.3793
Lag 1	-0.3928	0.1052	-3.7338

Autocorrelations of the Residuals from the AR(1) Model

Lag	Coefficient	Standard Error	t-Stat
1	-0.0695	0.1581	-0.4396
2	-0.1523	0.1581	-0.9633
3	-0.1231	0.1581	-0.7786
4	0.4542	0.1581	2.8729

This regression equation expresses the change in sales in the current quarter as a function of the change in sales in the last (previous) quarter.

seasonal lag

seasonal autocorrelation of the error term.

The intercept term and the coefficient on the first lag appear to be significantly different from 0, but the striking thing about the data in Table 8 is that the fourth error autocorrelation is significantly different from 0. The t-stat of 2.8729 is greater than the critical t-value of 2.024 (significance level = 5%, degrees of freedom = 38) so we reject the null hypothesis that the residual autocorrelation for the fourth lag equals 0. The model is therefore misspecified and cannot be used for forecasting.

The fourth autocorrelation is a seasonal autocorrelation as we are working with quarterly data. The model can be improved (adjusted for the seasonal autocorrelation) by introducing a seasonal lag as an independent variable in the model. The regression equation will then be structured as:

$$\ln \text{Sales}_t - \ln \text{Sales}_{t-1} = b_0 + b_1(\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2}) + b_2(\ln \text{Sales}_{t-4} - \ln \text{Sales}_{t-5}) + \varepsilon_t$$

This regression equation expresses the change in sales in the current quarter as a function of the change in sales in the last (previous) quarter and the change in sales four quarters ago.

Table 10 presents the results of the regression after introducing the seasonal lag.

Table 10: AR(1) Model with Seasonal Lag on Log First-Differenced Quarterly Sales

Regression Statistics			
R-squared	0.3483		
Standard error	0.0672		
Observations	40		
Durbin-Watson	2.031		

	Coefficient	Standard Error	t-Stat
Intercept	0.0386	0.0092	4.1957
Lag 1	-0.3725	0.0987	-3.7741
Lag 2	0.4284	0.1008	4.25

Autocorrelations of the Residuals from the AR(1) Model

Lag	Coefficient	Standard Error	t-Stat
1	-0.0248	0.1581	-0.1569
2	0.0928	0.1581	0.587
3	-0.0318	0.1581	-0.2011
4	-0.0542	0.1581	-0.3428

From the data in Table 10, notice that the intercept, and the coefficients on the first and second lags of the time series, are all significantly different from 0. Further, none of the residual autocorrelations is significantly different from 0 so there is no serial correlation. The model is therefore, correctly specified and can be used to make forecasts. Also notice that the R² in Table 10 (0.3483) is almost two times the R² in Table 9 (0.1763), which means that the model does a much better job in explaining ABC’s quarterly sales once the seasonal lag is introduced.

In order to make predictions based on the model, we need to know sales growth in the previous quarter ($\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2}$) and sales growth four quarters ago ($\ln \text{Sales}_{t-4} - \ln \text{Sales}_{t-5}$). For example, if sales grew 3% in the previous quarter and 5% four quarters ago, the model predicts that sales growth for the current quarter would equal 4.88%, calculated as:

$$\ln \text{Sales}_t - \ln \text{Sales}_{t-1} = 0.0386 - 0.3725(0.03) + 0.4284(0.05) = 0.48845 \text{ or } 4.88\%$$

Just one more thing before we move ahead. It is not necessary for only the residual autocorrelation corresponding to the seasonal lag to appear significant in a time series suffering from seasonality. For example, in Table 9, it may have been the case that the residual autocorrelation for the second lag was significantly different from 0 along with the residual autocorrelation for the fourth lag (the seasonal lag). After incorporating the seasonal lag as an independent variable in the model, if seasonality is present in the time series, the residual autocorrelations for the seasonal lag, and for any other lags for whom the residual autocorrelations were significant previously (before the introduction of the seasonal lag as an independent variable in the model), will equal 0.

The regression underlying Table 9 actually uses sales data for 41 quarters starting Q4 1990 so that we are able to have 40 observations of current quarter sales.

In the new model that includes the seasonal lag, we actually use sales data for 44 quarters starting Q1 1990 so that we have 40 observations of current quarter sales (which can be regressed on previous quarter sales and sales four quarters earlier)

If sales data prior to Q1 1991 is not available, then there would be 39 observations in Table 9 and 36 observations in Table 10.

Autoregressive Moving Average (ARMA) Models

An ARMA model combines autoregressive lags of the dependent variable and moving-average errors in order to provide better forecasts than simple AR models. The equation for an ARMA model with p autoregressive terms and q moving-average terms, denoted ARMA (p,q) is:

$$x_t = b_0 + b_1x_{t-1} + \dots + b_px_{t-p} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \dots + \theta_q\varepsilon_{t-q}$$

$$E(\varepsilon_t) = 0, E(\varepsilon_t^2) = \sigma^2, E(\varepsilon_t\varepsilon_s) = 0 \text{ for } t \neq s$$

ARMA models have the following limitations:

- The parameters of the model can be very unstable.
- There are no set criteria for determining p and q .
- Even after a model is selected, it may not do a good job of forecasting.

LOS 13m: Explain autoregressive conditional heteroskedasticity (ARCH), and describe how ARCH models can be applied to predict the variance of a time series. Vol 1, pg 483-486

Autoregressive Conditional Heteroskedasticity Models (ARCH Models)

Heteroskedasticity occurs when the variance of the error term varies with the independent variable. If heteroskedasticity is present in a time series, one or more of the independent variables in the model may appear statistically significant when they are actually not.

ARCH models are used to determine whether the variance of the error in one period depends on the variance of the error in previous periods. In an ARCH(1) model, the squared residuals from a particular time series model (the model may be an AR, MA or ARMA model) are regressed on a constant and on one lag of the squared residuals. The regression equation takes the following form:

$$\hat{\varepsilon}_t^2 = a_0 + a_1\hat{\varepsilon}_{t-1}^2 + u_t$$

- If a_1 equals 0, the variance of the error term in each period is simply a_0 . The variance is constant over time and does not depend on the error in the previous period. When this is the case, the regression coefficients of the time series model are correct and the model can be used for decision-making.
- If a_1 is statistically different from 0, the error in a particular period depends on the size of the error in the previous period. If a_1 is greater (less) than 0, the variance increases (decreases) over time. Such a time series is ARCH(1) and the time series model cannot be used for decision-making. The error in period $t+1$ can then be predicted using the following formula:

$$\hat{\sigma}_{t+1}^2 = \hat{a}_0 + \hat{a}_1\hat{\varepsilon}_t^2$$

Example 10: Testing for ARCH

Thomas Rosicky develops an AR(1) model for monthly inflation over the last 15 years. The regression results indicate that the intercept and lag coefficient are significantly different from 0. Also, none of the residual autocorrelations are significantly different from 0, so he concludes that serial correlation is not a problem. However, before using his AR(1) model to predict inflation, Rosicky wants to ensure that the time series does not suffer from heteroskedasticity. He therefore estimates an ARCH(1) model using the residuals from his AR(1) model. Table 11 contains the results of the regression.

Table 11: ARCH(1) Regression Results for AR(1) Model Residuals

Regression Statistics			
R-squared	0.0154		
Standard error	12.65		
Observations	180		
Durbin-Watson	1.9855		
	Coefficient	Standard Error	t-Stat
Intercept	4.6386	0.892	5.2
Lag 1	0.1782	0.0658	2.708

The regression results indicate that the coefficient on the previous period's squared residual (a_1) is significantly different from 0. The t-stat of 2.708 is high enough for us to be able to reject the null hypothesis that the errors have no autoregressive conditional heteroskedasticity ($H_0: a_1 = 0$). Since the model does contain ARCH(1) errors, the standard errors for the regression parameters estimated in Rosicky's AR(1) model are inaccurate and he cannot use the AR(1) model to forecast inflation even though (as mentioned in the question) the OLS regression coefficients are different from 0 and the residuals do not suffer from serial correlation.

Rosicky can use his estimated ARCH(1) model to predict the variance of the errors. For example, if the error in predicting inflation in one period is 1%, the predicted variance of the error in the next period is calculated as:

$$\hat{\varepsilon}_t^2 = 4.6386 + 0.1782\hat{\varepsilon}_{t-1}^2 = 4.6386 + 0.1782(1) = 4.8168\%$$

If ARCH errors are found to exist, generalized least squares may be used to correct for heteroskedasticity.

LOS 13n: Explain how time-series variables should be analyzed for nonstationarity and/or cointegration before use in a linear regression.

Vol 1, pg 487-491

Regressions with More than One Time Series

So far we have only been working with time-series models for one time series. Now we discuss whether linear regression can be used to analyze the relationship between more than one time series. Let's assume we are presented with two time series, one corresponding to the dependent variable and the other corresponding to the independent variable. Whether we can use linear regression to model the two series depends on whether the series have a unit root. The Dickey-Fuller test is used to make this determination.

There are several possible scenarios regarding the outcome of the Dickey-Fuller tests on the two series:

- If neither of the time series has a unit root, linear regression can be used to test the relationship between the two series.
- If either of the series has a unit root, the error term in the regression would not be covariance stationary and therefore, linear regression cannot be used to analyze the relationship between the two time series.
- If both the series have unit roots, we must determine whether they are **cointegrated**. Two time series are cointegrated if a long term economic relationship exists between them such that they do not diverge from each other significantly in the long run.
 - If they are not cointegrated, linear regression cannot be used as the error term in the regression will not be covariance stationary.
 - If they are cointegrated, linear regression can be used as the error term will be covariance stationary, the regression coefficients and standard errors will be consistent and they can be used to conduct hypothesis tests. However, analysts should still be cautious in interpreting the results from the regression.

Testing for Cointegration

To test whether two time series that each have a unit root are cointegrated, we perform the following steps:

1. Estimate the regression:

$$y_t = b_0 + b_1x_t + \varepsilon_t$$

2. Test whether the error term (ε_t) has a unit root using the Dickey-Fuller test but with Engle-Granger critical values.
3. H_0 : Error term has a unit root versus H_a : Error term does not have a unit root
4. If we fail to reject the null hypothesis, we conclude that the error term in the regression has a unit root, it is not covariance stationary, the time series are not cointegrated, and the regression relation is spurious.
5. If we reject the null hypothesis, we conclude that the error term does not have a unit root, it is covariance stationary, the time series are cointegrated and therefore, the results of linear regression can be used to test hypotheses about the relation between the variables.

Engle-Granger critical values are adjusted for the effect of the uncertainty about the regression parameters on the distribution of the Dickey-Fuller test.

If there are more than two time series, the following rules apply:

- If at least one time series (the dependent variable or one of the independent variables) has a unit root and at least one time series (the dependent variable or one of the independent variables) does not, the error term cannot be covariance stationary so multiple linear regression cannot be used.
- If all of them have unit roots, the time series must be tested for cointegration using a similar process as outlined previously (except that the regression will have more than one independent variable).
 - If we fail to reject the null hypothesis of a unit root, the error term in the regression is not covariance stationary and we conclude that the time series are not cointegrated. Multiple regression cannot be used in this scenario.
 - If we reject the null hypothesis of a unit root, the error term in the regression is covariance stationary and we conclude that the time series are cointegrated. However, bear in mind that modeling three or more time series that are cointegrated is very complex.

Note that when making forecasts based on time series analysis, we need to consider:

- The uncertainty associated with the error term; and
- The uncertainty about the estimates of the parameters in the model.

LOS 13o: Determine the appropriate time-series model to analyze a given investment problem, and justify that choice. Vol 1, pg 492-493

Suggested Steps in Time Series Forecasting

The following is a step-by-step guide to building a model to predict a time series.

1. Understand the investment problem you have, and make an initial choice of model. There are two options:
 - A regression model that predicts the future behavior of a variable based on hypothesized causal relationships with other variables. We studied these models in Readings 11 and 12.
 - A time-series model that attempts to predict the future behavior of a variable based on the past behavior of the same variable. We studied these models in this Reading.
2. If you go with a time-series model, compile the data and plot it to see whether it looks covariance stationary. The plot might show deviations from covariance stationarity, such as:
 - A linear trend
 - An exponential trend
 - Seasonality
 - A change in mean or variance

3. If you find no significant seasonality or a change in mean or variance, then either a linear trend or an exponential trend may be appropriate to model the time series. In that case, take the following steps:
 - Determine whether a linear or exponential trend seems most reasonable (usually by plotting the series).
 - Estimate the trend.
 - Compute the residuals.
 - Use the Durbin–Watson statistic to determine whether the residuals have significant serial correlation. If you find no significant serial correlation in the residuals, then the trend model is specified correctly and you can use that model for forecasting.

4. If you find significant serial correlation in the residuals from the trend model, use a more complex model, such as an autoregressive model. First however, ensure that the time series is covariance stationary. Following is a list of violations of stationarity, along with potential methods to adjust the time series to make it covariance stationary:
 - If the time series has a linear trend, first-difference the time series.
 - If the time series has an exponential trend, take the natural log of the time series and then first-difference it.
 - If the time series shifts significantly during the sample period, estimate different time-series models before and after the shift.
 - If the time series has significant seasonality, include a seasonal lag (discussed in Step 7).

5. After you have successfully transformed a raw time series into a covariance-stationary time series, you can usually model the transformed series with an autoregressive model. To decide which autoregressive model to use, take the following steps:
 - Estimate an AR(1) model.
 - Test to see whether the residuals from this model have significant serial correlation.
 - If there is no significant serial correlation, you can use the AR(1) model to forecast.

6. If you find significant serial correlation in the residuals, use an AR(2) model and test for significant serial correlation of the residuals of the AR(2) model.
 - If you find no significant serial correlation, use the AR(2) model.
 - If you find significant serial correlation of the residuals, keep increasing the order of the AR model until the residual serial correlation is no longer significant.

7. Check for seasonality. You can use one of two approaches:
 - Graph the data and check for regular seasonal patterns.
 - Examine the data to see whether the seasonal autocorrelations of the residuals from an AR model are significant (for example, if you are using quarterly data, you should check the fourth residual autocorrelation for significance) and whether other autocorrelations are significant. To correct for seasonality, add a seasonal lag of the time series to the model.

8. Test whether the residuals have autoregressive conditional heteroskedasticity. To test for ARCH(1) errors:
 - Regress the squared residuals from your time-series model on a lagged value of the squared residual.
 - Test whether the coefficient on the squared lagged residual differs significantly from 0.
 - If the coefficient on the squared lagged residual does not differ significantly from 0, the residuals do not display ARCH and you can rely on the standard errors from your time-series estimates.
 - If the coefficient on the squared lagged residual does differ significantly from 0, use generalized least squares or other methods to correct for ARCH.
9. As a final step, you may also want to perform tests of the model's out-of-sample forecasting performance to see how the model's out-of-sample performance compares to its in-sample performance.

