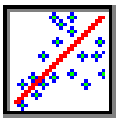
CHAPTER 16



Visualizing Regression Assumptions

CONCEPTS

- Regression Assumptions, Homoskedasticity, Mutually Independent Disturbances, Normally Distributed Disturbances, Zero Mean, Independence of Fixed X and Disturbance, Autocorrelation, Heteroskedasticity, Unbiased, Efficient, Consistent, Normally Distributed

OBJECTIVES

- Know the five assumptions that are required to ensure desirable properties of least-squares regression estimates
- Learn the various ways in which each assumption may be violated
- Be able to identify effects of each violation on the residuals and the estimated parameters
- Understand the types of violations that are likely to arise in time series and cross-sectional data
- Recognize how the properties of estimators may be affected by each type of violation

Overview of Concepts

The regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$ is based on five **regression assumptions**: (1) The disturbances u_i have constant variance, i.e. $E(u_i^2) = \sigma^2$ (**homoskedasticity,** also spelled homoscedasticity); (2) The disturbances are independent of one another, i.e., $E(u_i u_j) = 0$ for $i \neq j$ (**mutual independence**); (3) The disturbances u_i are **normally distributed**; (4) The disturbances u_i have **zero mean**, i.e., $E(u_i) = 0$; and (5) The values of the independent variable X_i are **fixed and independent of the disturbance**. Under these assumptions, the ordinary least squares (OLS) estimators for β_0 and β_1 are **unbiased** (the expected value is the true parameter), **consistent** (the variance of the estimator approaches zero as sample size increases, so the estimator collapses on the true parameter), **efficient** (the smallest variance among all unbiased estimators), **normally distributed**, and BLUE (Best Linear Unbiased Estimators).

Heteroskedasticity (non-constant error variance) is a major violation. The OLS estimators are unbiased, consistent, and normally distributed, but are neither efficient nor BLUE. In a given sample, the validity of a confidence interval for E(Y|X) is in doubt over some regions of the X range. If the underlying constant variance is denoted σ^2, some commonly assumed patterns are $k X_i^c \sigma^2$ (variance proportional to a power of X_i), $k E(Y)_i^c \sigma^2$ (variance proportional to expected value of Y), $k Z_i^c \sigma^2$ (variance proportional to a power of another variable), or $k i^c \sigma^2$ (variance proportional to a power of the observation order). Homoskedastic errors should exhibit no discernible pattern in the residuals on the Y-X scatter plot. Heteroskedasticity may be seen as a "fan-out" pattern of residuals (increasing variance as we move to the right) or a "funnel-in pattern" of residuals (decreasing variance as we move to the right).

Autocorrelation (non-independent disturbances) is common in time series data, and is a major violation. The OLS estimators are unbiased, consistent, and normally distributed, but are neither efficient nor BLUE. The first-order autocorrelation model is $u_t = \rho u_{t-1} + v_t$ where ρ is a constant such that $-1 \leq \rho \leq +1$ and v_t is a "well-behaved" disturbance that fulfills all assumptions. If $\rho < 0$ (negative autocorrelation) then u_{t-1} tends to be followed by u_t of the opposite sign. If ρ is near 0 (non-autocorrelation) then u_t is unrelated to u_{t+1}. If $\rho > 0$ (positive autocorrelation) then u_{t-1} tends to be followed by u_t of the same sign. Runs of residuals of the same sign (e.g., + + + – – – + + +) suggest *positive* autocorrelation (common) while runs of residuals of alternating sign (e.g., + – + – + – + – + –) suggest *negative* autocorrelation (rare).

With *non-normal disturbances*, the OLS estimators are unbiased, consistent, and BLUE, but are neither efficient nor normally distributed. However, they are asymptotically efficient, so in large samples confidence intervals for E(Y|X) may be acceptable. Non-normal disturbances are not a major violation, because the OLS method is fairly robust to non-normality. Evidence may be sought in the probability plot of residuals, which is nearly linear under normality.

Non-zero mean exists when the disturbances are not centered on the true regression line. It may be an annoyance or a major violation. A *constant* non-zero mean is a trivial problem, since it will merely be reflected in the intercept estimate and will not even be observable. A *non-constant* non-zero mean is a severe problem due to *incorrect functional form* or an *omitted variable* Z, making the OLS slope estimates biased and inconsistent (although if Z and X are uncorrelated, desirable asymptotic properties of the slope may still exist).

Stochastic X (X is not fixed) is a minor violation causing no loss of desirable properties *unless* X is correlated with the disturbance u_t (non-independence of X and disturbances). In the latter case, the estimator for the slope loses all desirable properties.

Illustration of Concepts

Figure 1 illustrates a positive relationship between years of employment (X) and home value (Y) for a sample of 100 public school teachers. Lack of **homoskedasticity** is evident in the "fan-out" pattern of points on the scatter plot. The residual plot in Figure 2 clearly shows the **heteroskedastic** pattern of increasing variance. Despite this violation of the **regression assumptions**, the slope and intercept estimates are **unbiased** and **consistent**. However, they are not **efficient** and their t-values may be unreliable. Specifically, regression estimates of home value will be more precise (smaller variance) for teachers with less experience and less precise (larger variance) for those with more experience. This makes sense, because new teachers generally cannot afford expensive homes, so the Y value varies within a smaller range. Conversely, experienced teachers generally (but not always) own more expensive homes (greater range) so their home values are harder to predict.

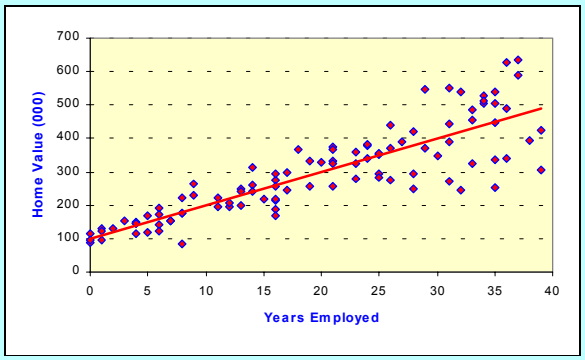


Figure 1: Home Value and Years Employed

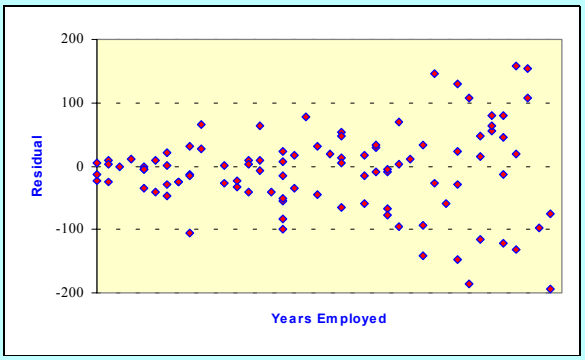


Figure 2: Heteroskedastic Residuals

Another serious problem is **autocorrelation**. Figure 3 shows a scatter plot and fitted regression line relating aggregate personal taxes (PT) to aggregate personal income (PI) for the U.S., 1960-1997 (data are in billions of dollars). The fitted equation is PT = -8.620 + .1337 PI with $R^2 = 0.99$. The excellent fit disguises the fact that the errors lack **mutual independence**. A bar chart of the residuals against time (Figure 4) reveals a pattern with "runs" of positive and negative residuals. Figure 4 also suggests heteroskedasticity (increasing variance). This reminds us that more than one assumption can be violated. In this case, the OLS. estimates are unbiased and consistent, but not efficient. Other violations of assumptions may exist, but are not illustrated here.

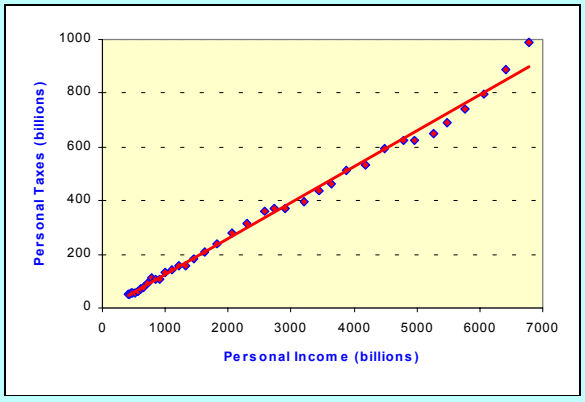


Figure 3: U.S. Income and Taxes, 1960-1997

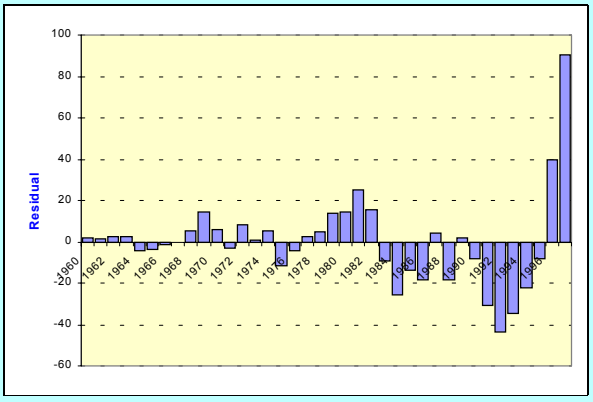


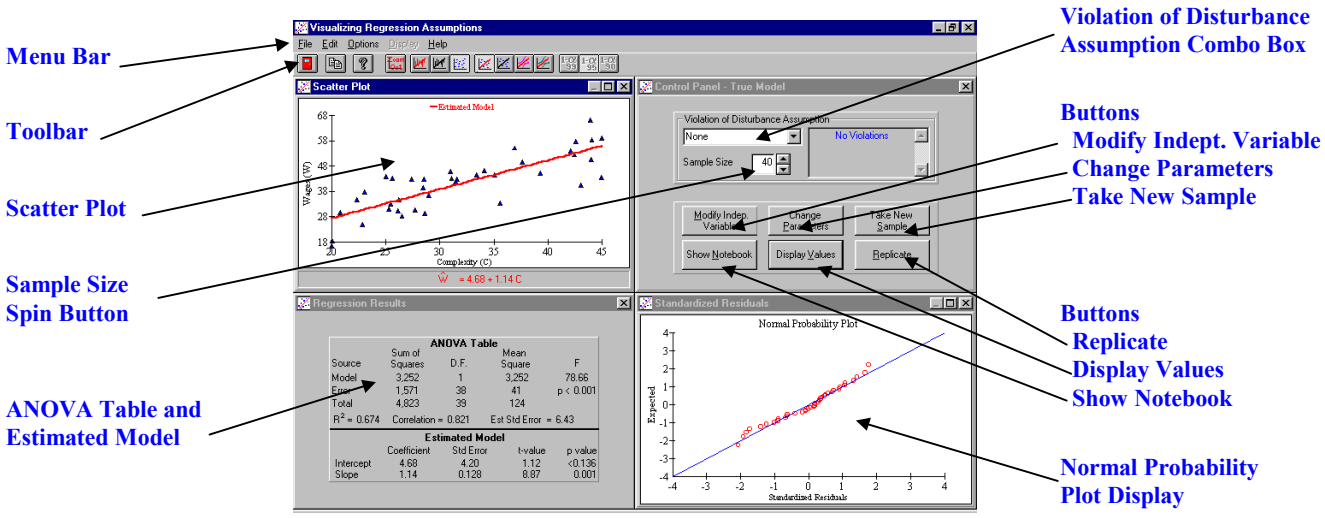
Figure 4: Autocorrelated Residuals

Orientation to Basic Features

This module generates data from a true regression model that meets all five regression assumptions, but then lets you violate any *one* assumption. You can examine similarities and differences between the true and estimated models and between the residuals and disturbances. You can display a confidence interval for E(y|x) and a prediction interval for y|x. You can replicate the experiment and display the estimated models and histograms of the estimated slopes, intercepts, standard errors, R^2, t-statistics, F ratios, and correlation coefficients.

1. Opening Screen

Start the module by clicking its Run Module button in the *Visual Statistics* menu. When the module is loaded, you will be on the introduction page of the Notebook. Click the Introduction and the Concepts tabs to see topics that are covered in this module. Click on the Scenarios tab. Click on No Violations, select the Wage vs Job Complexity scenario, read it, and press OK.



2. Control Panel

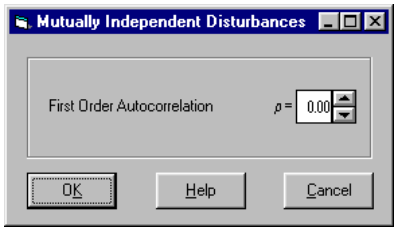
The control panel is in the upper right quadrant. The Sample Size spin button sets the sample size. Press the Modify Indept. Variable button to bring up a new window to change the independent variable's minimum value, maximum value, type of variable (integer or decimal), and whether the X end points are included. Press the Change Parameters button to see a control panel that allows you to change the model. Press the Take New Sample button to draw a new sample. The Show Notebook button reveals the Notebook allowing you to change scenarios or use the Do-It-Yourself controls. Press the Display Values button to see a table of the x values, y values, predicted y values, residuals, and standardized residuals. Click on the Violation of Disturbance Assumption combo box to change the assumption being violated (this is specified by the scenario). The text window states the violation that has been chosen. You may only choose one violation at a time (otherwise, the violations are too muddled to be distinguished).

3. Scatter Plot and Regression Results

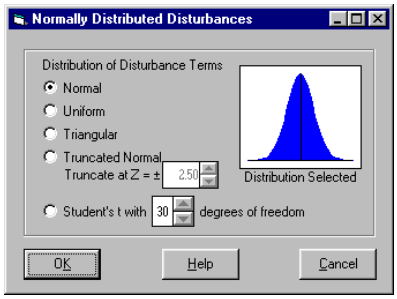
The other three quadrant displays are identical to Chapter 15, Visualizing Simple Regression. When you click on the scatter plot, the menu bar buttons become active, to display residuals, disturbances, true model, estimated model, confidence intervals, or prediction intervals.

4. **Normal Probability Plot Display**
Activate the normal probability plot display by clicking on it. Five toolbar buttons appear to the right of the ? button. They are identical to Chapter 15 except for the third button, which shows the probability plot. The fourth button is inactive unless the histogram of residuals is displayed. Any graph (scatter plot, probability plot, residual histogram) can be maximized (or minimized) with the windows buttons in its upper right corner.

5. **Assumption of Homoskedasticity**
Click on the Violation of Disturbance Assumption combo box and select Homoskedastic to produce the control panel for this assumption. Use the scroll bar to select the desired degree of "fan-out" (right side of the scale) of "funnel-in" (left end of the scale). Do *not* click Advanced Options at this time.



6. **Assumption of Mutual Independence**
Click on the Violation of Disturbance Assumption combo box and select Mutual Independence to produce the control panel for this assumption. Use the spin button (or type in any ρ value from -1 to +1) to select the desired degree of autocorrelation. The default is ρ = 0 (none).



7. **Assumption of Normally Distributed Disturbances**
Click on the Violation of Disturbance Assumption combo box and select Normal Distribution to produce the control panel for this assumption. Instead of Normal (the default) you can choose Uniform, Triangular, Truncated Normal (you specify the truncation point), or Student's t (you specify the degrees of freedom). All are symmetric. Click on each and watch how the diagram varies, then click Cancel to return to the main control panel.

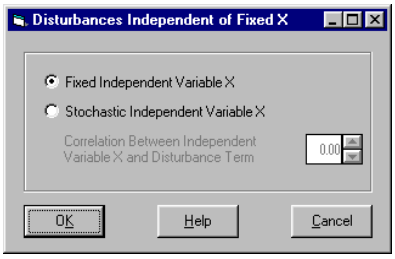
8. **Copying a Display**
Click on any graph or the ANOVA Table. Press the Copy button on the toolbar or select Copy from the Edit menu on the menu bar. The copied display can be pasted into another application.

9. **Help**
Press the ? button on the toolbar or click Help on the menu bar at the top of the screen. Search for Help lets you search an index, Contents shows a table of contents for this module, Using Help gives instructions on Help, and About gives licensing and copyright information.

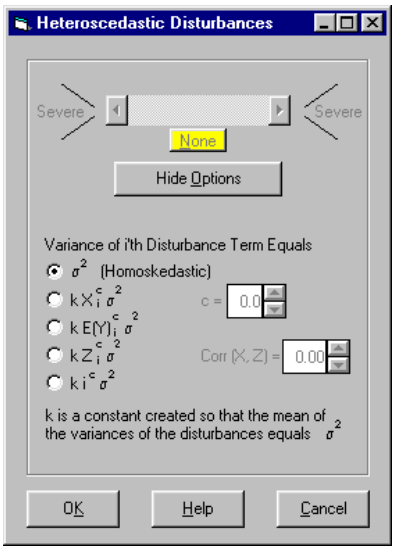
10. **Exit**
Close the module by selecting Exit in the File menu (or click ☒ in the upper right-hand corner of the window). You will be returned to the *Visual Statistics* main menu.

Orientation to Additional Features

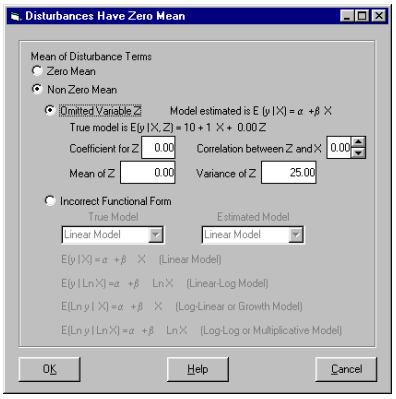
1. **Assumption of Independent Fixed X**
   Click on the Violation of Disturbance Assumption combo box and select Independent of Fixed X to produce the control panel for this assumption. The default is Fixed Independent Variable X (no violation). You can select Stochastic Independent Variable X and (to achieve the worst violation) you can have X be correlated with the disturbance (the default is no correlation). *Note: The latter violation can cause such severe bias as to make the graphs difficult to read.*



3. **Assumption of Homoskedasticity**
   Click on the Violation of Disturbance Assumption combo box, select Homoskedastic, and press Advanced Options to produce the control panel to the right. Select a form of the heteroskedasticity and a value for the exponent c. If the form is $kZ_i^c\sigma^2$, select the correlation between Z and X. Press the Cancel button to return to the main screen.



4. **Assumption of Zero Mean of the Disturbances**
   Click on the Violation of Disturbance Assumption combo box and select Zero Mean to see the control panel for this assumption (bottom figure to the right). If you select Non-Zero Mean you must specify if the violation is caused by omitting a variable Z or if the wrong functional form was estimated. If the Omitted Variable Z option is selected, you must specify its coefficient, correlation with X, mean and variance. If the Incorrect Functional Form option is selected you must select the true and estimated functional form from the four options provided. Press the Cancel button to return to the main screen.



4. **Replication and Do-It-Yourself Controls**
   The Replicate button works the same as in Chapter 15, except that three additional replication histograms are available (F statistics and t-statistics for slope and intercept). The Do-It-Yourself tab is identical to Chapter 15, allowing you to set the intercept, slope, and standard error (or, if you prefer, you can set the desired R^2 rather than the standard error, by clicking on Options on the menu bar and select Set Error Using and then Desired R-Squared).

Basic Learning Exercises Name _____

Visualizing Heteroskedasticity

1. Press the Show Notebook button and select the Scenarios tab. Select No Violation and select the Wage vs. Job Complexity scenario. Read it and press the OK button. Click on the Violation of Disturbance Assumption combo box and select Homoskedastic. Move the slider bar to the right to generate severe "fan-out" heteroskedasticity. Press OK. Click on the Normal Probability Plot to select it and use the toolbar below the menu bar to select a bar graph (4^{th} icon from the left). The bar graph displays the residuals *against the ordered X values*. Press the Take New Sample button a couple of times and describe the bar graph and the scatter plot. Why does this "fan-out" pattern indicate that the disturbances are heteroskedastic?

2. Click again on the Assumption combo box and again select Homoskedastic. Move the scroll bar to the left to generate "funnel-in" heteroskedasticity. Press the OK button. Press the Take New Sample button a couple of times and describe the bar graph and the scatter plot. Why does this "funnel-in" pattern indicate that the disturbances are heteroskedastic?

3. Press the Show Notebook button and click on Next page twice to see scenarios with heteroskedasticity. Select the Heteroskedasticity Related to X scenario read it and press OK. a) Press the Take New Sample button a couple of times and describe the bar graph and scatter plot. b) Why would the variance of the error term increase with Money Growth? c) In many regression programs, the bar graph is replaced with a scatter plot of the residuals *against X*. What is a possible advantage and disadvantage of such a scatter plot? d) Press the Show Notebook button, select the Heteroskedasticity Related to E(Y) scenario, read it and press OK. Press the Take New Sample button a couple of times and describe the bar graph and the scatter plot. e) Rather than this bar graph, a scatter plot of the residuals *against predicted Y* (since E(Y) is not observed) is usually used to reveal this type of heteroskedasticity. Why?

4. Press the Show Notebook button, select the Heteroskedasticity Related to 3rd Variable scenario, read it and press OK. Press the Take New Sample button a couple of times and describe the bar graph and the scatter plot. Why does neither graph reveal the heteroskedasticity?

Visualizing Autocorrelation

5. Press Show Notebook and click on Previous page for scenarios with autocorrelation. Select Severe Positive Autocorrelation. Read it, and press OK. a) Press the Take New Sample button a couple of times and describe the bar graph and the scatter plot. b) Take 10 samples, in what percent is the autocorrelation evident? c) Why is this called first-order autocorrelation? d) Why does this violation occur most often in time series data? e) In many regression programs, the bar graph is replaced with a scatter plot of the residuals *against the observation number*. Why would a graph versus the *observation number* generally be more reliable at revealing first-order autocorrelation than a graph versus the *ordered X values*.

6. Press the Show Notebook button, select the Moderate Positive Autocorrelation scenario, read it, and press OK. a) Press the Take New Sample button a couple of times and describe the bar graph and the scatter plot. b) Take 10 samples, in what percent is the autocorrelation evident? What is the value of rho (read the description in the panel above the buttons)? c) Click on the Assumptions combo box, select Mutual Independence, reduce ρ (rho) to 0.2, and press the OK button. Press Take New Sample. Take 10 samples. In what percent is the autocorrelation evident? d) What can you conclude about the value of rho and the ability to detect first-order autocorrelation on a graph?

7. Press the Show Notebook button, select the Negative Autocorrelation scenario, read it, and press OK. a) Press the Take New Sample button a couple of times and describe the bar graph and the scatter plot. b) Increase rho using the Mutual Independence dialog box (as in question 6). Is it still harder to detect visually when rho is closer to zero? c) Why do you think negative autocorrelation would be more unusual in real world data?

Visualizing Non-normally Distributed Disturbances

8. Press Show Notebook and click on Previous page for scenarios with no violations. Select the State Income Tax vs. Income scenario. Read it, and press OK. Replace the bar graph of residuals with a histogram of the standardized residuals (use the toolbar after clicking on the bar graph). a) Press the Take New Sample button a couple of times and describe the histogram. b) Replace the histogram with the normal probability plot. Press the Take New Sample button a couple of times and describe the probability plot. **Hint:** If you don't know how to use a normal probability plot, click Help in the menu bar, select Contents, and select Displays: Normal Probability Plot of Residuals under the heading Using the Program.

9. Press Show Notebook and click on the Scenarios tab. Select Non-normality and select the Uniformly Distributed Error Term scenario. Read it, and press OK. a) Press the Take New Sample button a couple of times and describe the normal probability plot. b) Replace the normal probability plot with the histogram. Press the Take New Sample button a couple of times and describe the histogram. c) Do you think either graph could be effectively used to detect residuals that are uniformly distributed?

10. Press Show Notebook and select the Peaked Error Term scenario. Read it, and press OK. a) Press the Take New Sample button a couple of times and describe the histogram. b) Replace the histogram with the normal probability plot. Press the Take New Sample button a couple of times and describe the probability plot. c) Do you think either graph could be effectively used to detect residuals that are more peaked than a normal distribution?

11. Click on the Assumptions combo box and select Normal Distribution. The sketch shows that the disturbances currently are distributed with a Student's t distribution with 3 degrees of freedom. Increase the degrees of freedom to get an idea of how peaked this distribution really is. Select Triangular distribution and press the OK button. a) Press the Take New Sample button a couple of times and describe the normal probability plot. b) Replace the normal probability plot with the histogram. Press the Take New Sample button a couple of times and describe the histogram. c) Superimpose a normal distribution on the histogram. What do you now observe? d) Do you think either graph could be effectively used to detect residuals that are triangularly distributed? **Hint:** Use the Assumption combo box to bring up normally distributed disturbances and see what you observe.

Intermediate Learning Exercises Name _____

Visualizing Unbiasedness, Consistency, Normality and Efficiency

12. Press Show Notebook and click on the Scenarios tab. Select No Violation and select the Job Performance vs. Test Scores scenario. Read it, and press OK. a) Write down the true model. Press the Replicate button. Read the Hint that appears and press OK. Set the number of replications to 1000 by using its combo box. Reduce the sample size to 10 using its spin button. Press the Start Experiment button. Press the Pause Experiment button after a few replications. The current estimate of the slope and intercept are entered into their respective histograms in red. The corresponding estimated regression line is shown in red in the upper left quadrant. Press the Continue Experiment button to restart the experiment. Press the Finish Experiment button to fast forward to the end of the experiment. b) The true standard error of this model is 40 (variance is 1600). Since all of the assumptions are true, what is the theoretical distribution of the estimated slope histogram? Be sure and specify its parameters. c) Click on the histogram for the estimated slope and superimpose the true distribution (toolbar below the Menu bar). How does this show that the estimated slope is unbiased? d) normally distributed? e) Is the estimated intercept normally distributed? f) unbiased? **Hint:** Click on Search for Help on in Help and type "slope" to find the true distribution,

13. To be consistent, an estimator *must* be unbiased (or the bias must go to zero as the sample size n increases to infinity) and its variance must decrease as n increases. Write down the minimum and maximum value for the estimated slope and intercept. Increase the sample size to 100. Press the Start Experiment button and then the Finish Experiment button. How does this experiment show that both estimators are consistent? (The sophisticated user will recognize this as *mean square error consistency*.)

    Slope: Minimum _____ Maximum _____ Intercept: Minimum _____ Maximum ____

14. To be efficient an estimator *must* to be unbiased and either have the minimum variance of *all* unbiased estimators or achieve the Cramer-Rao lower bound. In this case, this lower bound can be verified visually by comparing the histogram of the estimated variances for the model, to its theoretical distribution. The theoretical distribution is a scaled χ^2, where the scale factor is $\sigma^2 / (n - 2)$. Right click on the Estimates of Model window (upper left) and select Histogram – Estimates of Variance. Superimpose its theoretical distribution. Does the theoretical distribution outline the histogram? If it does, then it achieves the lower bound and the slope and intercept estimators are efficient.

15. Because the estimators for the slope and intercept are unbiased, the test statistics for the estimated parameters have a Student's t distribution with n – 2 degrees of freedom. Press the Exit Replication Mode button. Press the Change Parameters button. Change the intercept's value from –50 to 0. Press the Change Assumptions button to bring back the assumption panel. Press the Replicate button. Press the Start Experiment button and then the Finish Experiment button. Replace both estimated parameter histograms with their t-statistic histograms. Superimpose the true distribution on both t statistic histograms. Since the true value of the *intercept* is 0, the null hypothesis is true (H_0: $\beta_0 = 0$). The histogram of t-statistics has the traditional t distribution with 98 (n – 2) degrees of freedom. The critical values for $\alpha = 0.05$ are approximately ±2 (shown in magenta). a) Approximately, how many rejections out of 1000 are in each tail? b) Now look at the t-statistic for the slope. In this case the true value of the slope coefficient is 2 (check your answer to question 12a). Therefore, the null hypothesis that $\beta_1 = 0$ is false. The magenta lines are the critical values. What are their approximate values? c) Approximately, what percentage of the 1000 estimates is outside these critical values? Because the null hypothesis is false, the distribution being displayed is the true distribution when $\beta_1 = 2$. It is called a non-central t distribution with n – 2 degrees of freedom. *Notice that the histogram corresponds to what the theory suggests when the assumptions about the disturbances are true.*

Properties of Estimators when Disturbances are Heteroskedastic

16. Click on the Assumptions combo box and click on Homoskedastic. If the advanced options are showing, press the Hide Options button. Move the slider all the way to the right. Press OK. You now have severe heteroskedasticity (fan-out). Make sure n = 100. Use the process outlined in question 12 to determine if the estimated parameters are unbiased and normally distributed (recall that $\beta_0 = 0$ and $\beta_1 = 2$).

17. Use question 13 as a guide to determine if the estimated parameters are efficient. **Hint:** In this module, the average variance of all observations equals σ^2. If this was not done, the form of the heteroskedasticity may shift the histogram. However, the difference in shape between the histogram and the theoretical distribution is what causes inefficiency, not just the shift.

18. Will the t-statistics agree with their theoretical distribution? If yes why, if no why not? Why is this important? Display the t-statistics to see if you were correct. **Hint:** To see a table comparing the expected and actual frequency by interval, select the table icon on the toolbar.

19. Use question 14 as a guide, determine if the estimated parameters are consistent.

20. Select Homoskedastic in the Assumption combo box. Move the slider all the way to the left. Press OK. You now have severe funnel-in heteroskedasticity. Make sure n = 100. Press the Start Experiment button and then the Finish Experiment button. Do you get different conclusions when funnel-in heteroskedasticity is present?

Properties of Estimators when Disturbances are Autocorrelated

21. Click on the Assumptions combo box and click on Mutual Independence. Set ρ to 0.4. Press OK. You now have moderate autocorrelation. Make sure n = 100. Use the process outlined in question 12 to determine if the estimated parameters are unbiased and normally distributed (recall that $\beta_0 = 0$ and $\beta_1 = 2$).

22. Use question 13 as a guide to determine if the estimated parameters are efficient.

23. Will the t-statistics agree with their theoretical distribution? If yes why, if no why not? Why is this important? Display the t-statistics to see if you were correct.

24. Using question 14 as a guide, determine if the estimated parameters are consistent.

25. Click on the Assumptions combo box and click on Mutual Independence. Set ρ to -0.4. Press OK. You now have moderate *negative* autocorrelation. Make sure n = 100. Although the estimators for the slope and intercept are still unbiased, consistent, normally distributed, and inefficient, which histograms are substantially different? What would these differences imply about hypothesis testing?

Properties of Estimators when Disturbances are Not Normally Distributed

26. Click on the Assumptions combo box, click on Normal Distribution, select Student's t, and set degrees of freedom to 3. Press OK. Set n = 10. a) Are the estimated parameters unbiased? b) Are they normally distributed? c) Are the estimators efficient?

27. a) Display the t-statistics. Are the test statistics distributed as Student's t? If yes why, if no why not? b) Why is this important? c) Are the estimated parameters consistent?

28. With sample size set to 100, do both estimators appear to be normally distributed? Are the test statistics distributed as Student's t?

29. Change the distribution to uniform using the Assumptions combo box. Set n = 10. a) Are the estimated parameters normally distributed? b) Are the test statistics distributed as Student's t? c) How can you explain this result given your answer to questions 26 and 27?

Advanced Learning Exercises Name _____

Omitted Variable

30. Press Show Notebook and click on the Scenarios tab. Select Other Violations, select the Omitted Variable scenario, and read it. a) Write down the true model. b) What model will be estimated? Press OK. c) Write down the estimated model. d) Superimpose the true model on the scatter plot (use the toolbar). What do you see?

31. Press the Replicate button. Set Replications to 1000 and run the experiment. a) Are the estimated parameters unbiased? b) Are the estimated parameters efficient and/or consistent? c) What does this mean about the estimated model? **Hint:** Superimpose the true model on the Estimates of Model graph.

32. Click on the Assumptions combo box and Select Zero Mean. Change the correlation between Z and X to 0. a) What s the mean of Z? b) What is its coefficient? Press OK. Press the Start Experiment button and then the Finish Experiment button. c) Is the estimator of the slope unbiased, normally distributed, efficient, and consistent? d) The mean for the intercept is 6.6 rather than 5. Why is the bias 1.6? **Hint:** The omitted variable Z is years until maturity (Y).

33. Click on the Assumptions combo box and Select Zero Mean. Change the correlation between Z and X to 1. Press OK. Press the Start Experiment button and then the Finish Experiment button. Why is the bias for the estimated slope 0.2? Why is the intercept still biased?

34. What does this tell us about the effect of omitting a variable from the model? **Hint:** 3 cases.

Stochastic Independent Variable

35. Press Exit Replication button, press Show Notebook, select the Omitted Variable scenario, and read it. a) Write down the true model. b) Which assumption is violated? c) Press OK. Change the sample size to 5 and press the Take New Sample button. Look at the scatter plot and press the Take New Sample button again. Do the X values change? d) Click on the Assumption combo box and select Independent of Fixed X. Select Fixed Independent Variable X and press OK. Press Yes when the message box appears. Press the Take New Sample button. Look at the scatter plot and press the Take New Sample button again. Do the X values change? e) What does stochastic X mean?

36. Press the Show Notebook button and OK to reload the scenario. Conduct a replication experiment to determine if the estimators are unbiased, normally distributed, consistent, and efficient?

37. Exit replication and return to the notebook. Select the Error Term and X Variable Correlated scenario and read it. a) Write down the true model. b) Which assumption is violated? c) Look at the scatter plot, probability plot, histogram, and bar graph. Repeatedly take new samples. Do you see any problems? d) Display the true model on the scatter plot. What do you see? e) Do you think the estimated parameters are unbiased?

38. Conduct a replication experiment to determine if the estimators are unbiased, normally distributed, consistent, and efficient.

Incorrect Functional Form

39. End replication and return to the notebook. Select the Incorrect Functional Form scenario and read it. a) What is the true model? b) What model will be estimated? c) Press OK. Look at the scatter plot, probability plot, histogram, and bar graph. Repeatedly take new samples. Do you see any problems? d) Display the true model on the scatter plot. What do you see? e) Do you think the estimated parameters are biased?

40. Press the Zoom Out icon on the toolbar for the scatter plot. a) What do you see? b) What does this suggest about using the wrong model for forecasting? c) Interpret the true model's slope. d) Interpret the estimated model's slope.

41. Press the Replicate button and conduct a replication experiment with 1000 replications. a) Are the parameters unbiased? b) What is the amount of the bias in the slope parameter and the intercept parameter? c) Given these results, are the estimators consistent or efficient?

42. Superimpose the true model and confidence interval on the Estimated Model graph by using their icons on the toolbar. What does this tell you about the effect of using the wrong functional form?

43. Examine the t-statistic histograms and the histogram of the variances. Superimpose the distribution of each statistic if the model had been estimated with the correct functional form. a) What do you see? b) Why do you think the histogram of the t-statistic of the slope is similar to its distribution if the model had been estimated with the correct functional form? c) Why is the distribution of the variance similar to its histogram?

44. Examine the F-statistic, R^2, and correlation coefficient. Since the wrong functional form is being used, why is the histogram for the F-statistic and the true distribution similar?

45. Using the histogram, write down approximate mean, minimum value, and maximum value of R^2.

46. Click on the Assumptions combo box and select Zero Mean. Change the estimated model to linear-log using the combo box. The model will now be correctly estimated. Press the OK button. Conduct an experiment to determine if the parameters are unbiased and normally distributed? How do you know?

47. Are the estimated parameters efficient? Explain.

48. Are the estimated parameters consistent? Explain.

49. What do your answers to questions 46 – 48 tell you about estimating nonlinear models?

50. Is the R^2 similar to the experiment you did earlier (see your answer to question 45)? Why?

Individual Learning Projects

Write a report on one of the two topics listed below. Use the cut-and-paste facilities of the module to place the appropriate graphs and tables in your report.

1. From the notebook, select one of the scenarios that have no violation of the assumptions. Using only this scenario, illustrate positive and negative autocorrelation, funnel-in and fan-out heteroskedasticity, and non-normality using a uniform and a Student's t distribution with 3 degrees of freedom. For each of the six cases, provide a scatter plot and indicate if the violation is apparent on the graph. Also, provide a residual display that would be useful in detecting the violation and indicate if the display reveals the violation or not. In studying autocorrelation, indicate the value of ρ you used. In studying heteroskedasticity, indicate the number of clicks on the scroll bar you used or, if you used the advanced options, the form of the heteroskedasticity and its parameters. The final paper should *explain and illustrate* the different forms of each of the 3 violations and how each may be detected visually.
2. Study either heteroskedasticity or autocorrelation. Select a scenario from the notebook with no violation of the assumptions. Using only this scenario, study one of the violations by creating different amounts and types of the violation. For autocorrelation, 6 different values of ρ should be used (both positive and negative). For heteroskedasticity, 6 different degrees of heteroskedasticity should be used (both funnel-in and fan-out). For each of the 6 cases, illustrate the problem with a scatter plot or one of the residual graphs. Also for each case, since the estimated parameters are inefficient, select a replication graph that illustrates the degree of inefficiency for that case. The final paper should *explain and illustrate* the different forms and degrees of the violation and the effect the violation has on efficiency.
3. Study the effect of estimating a linear model when it is *not* the correct model. Select a scenario from the notebook with no violation of the assumptions. Use this scenario to create a true model that is log-linear. Conduct a replication experiment. Illustrate and explain why the estimators are biased, inefficient, and inconsistent. Which t-statistic (slope or intercept) is approximately what it should have been and why? Since this is the wrong functional form, explain why R^2 is so large. What is the interpretation of the slope and intercept term in the linear model and in the log-linear model? Use the scatter plot and/or estimated model graph to explain the effect of incorrectly estimating a linear model. Repeat the process using a linear-log model and a log-log model.

Team Learning Projects

Select one of the three projects listed below. In each case, produce a team project that is suitable for an oral presentation using presentation software or large poster boards. Graphs should be large enough for your audience to see. Each team member should be responsible for producing some of the graphs. Ask your instructor if a written report is also expected.

1. This is a project for a team of 2 or 3. The purpose of the project is to study the ability to detect non-normality visually and the effect it has on the test statistics. The team should select a scenario to study throughout this project. Create a model that has a disturbance term that is uniformly distributed. This is a very platykurtic distribution. What effect did this distribution have on the t-statistics (copy the t-statistic histograms)? Is this violation evident on the probability plot (copy this diagram)? Repeat this process for a triangular distribution (less platykurtic), a normal distribution (a mesokurtic distribution), and a Student's t distribution with 24, 12, 6, and 3 degrees of freedom (increasingly leptokurtic). The presentation should focus on the ability to detect violations and the effect on the t-statistics.
2. This project is for a team of 2 or 3. The purpose is to study the effect of an omitted variable on the estimated included parameters. The team should select a scenario to study throughout this project and agree on the coefficient of the omitted variable, its mean, and its variance. Create a model that has an omitted variable that is not correlated with the included variable. Conduct a replication experiment to determine the bias in both the slope and intercept. Illustrate these biases. Repeat the process for correlations of 0.25, 0.50, 0.75, 1.00, –0.25, –0.50, –0.75, and –1.00. Derive the bias arithmetically in any situation where it is possible. The presentation should focus on the relationship between the correlation and the amount of bias in both estimators.
3. This project is for a team of 4. The purpose is to study the effect of estimating models with transformed variables. The team should select a scenario to study throughout this project. One team member should create a linear true model and estimate it with a linear-log model. Using a replication experiment show that both estimators are biased. Why are they also inefficient, and inconsistent? Why is R^2 so large, considering that the estimated model has the wrong functional form? Why is the distribution of the t-statistic for the slope about the same as it would have been if the model were estimated with a linear model? Use either the scatter plot or the Estimated Models graph to explain the cost of using the wrong functional form. Repeat the estimation process using a log-linear, a log-log, and a linear model. When the model is correctly estimated show that the estimators are unbiased, efficient, consistent, and normally distributed. The second team member should repeat the process with a log-linear model that is estimated with a linear, log-linear, linear-log, and log-log model. The third team member should repeat the process with a linear-log model that is estimated with a linear, log-linear, linear-log, and log-log model. The last team member should repeat the process with a log-log model that is estimated with a linear, log-linear, linear-log, and log-log model. The presentation should show: (a) that the parameters of any model correctly estimated will have all desirable properties, (b) that the parameters of any model estimated with the wrong functional form will have no desirable properties, (c) that models incorrectly estimated will still have large R^2s, and appropriate t-statistics, and (d) that regardless of the redeeming features, there are substantial costs of using the wrong functional form.

Self-Evaluation Quiz

1. Which is *not* one of the five assumptions of regression?
   a. The disturbances are observable.
   b. The disturbances are independent of each other.
   c. The disturbances are normally distributed.
   d. The disturbances have a constant variance.
   e. The disturbances have zero mean.

2. If the disturbances $u_1, u_2, ..., u_n$ in a time series model exhibit autocorrelation, then
   a. u_t is independent of u_{t-1}.
   b. u_t is the same as u_{t-1}.
   c. u_t is related to u_{t-1}.
   d. u_t is related to X_t.
   e. u_t has non-zero mean.

3. Autocorrelation would most likely show up in
   a. the intercept estimate.
   b. the slope estimate.
   c. the R^2 statistic.
   d. the time plot of residuals.
   e. the histogram of residuals.

4. Which pattern of residual signs is most likely to indicate *positive* autocorrelation?
   a. + + + + + + – – – – – – – + + + + + + + + + + – – – – – – – – – – + + + + +.
   b. + – + – + – + – + – + – + – + – + – + – + – + – + – + – + – + + – + – +.
   c. – + + – + – + – + – + – + – + – – – + + – – + – – + + – + + – – + + – + – –.
   d. Any of the above indicates positive autocorrelation.
   e. Either a. or b. indicate positive autocorrelation.

5. If there is heteroskedasticity, it would be most evident in
   a. a plot of the residuals against X or $\hat{Y}$.
   b. a histogram of the residuals.
   c. the fitted regression equation.
   d. the probability plot of standardized residuals.
   e. the ANOVA table and R^2.

6. In a time-series regression model, which violation of an assumption is *most* likely?
   a. Heteroskedasticity.
   b. Non-zero mean.
   c. Non-stochastic X.
   d. Autocorrelation.
   e. Incorrect functional form.

7. Non-normality of errors
   a. can be detected from the residual probability plot.
   b. is often considered a relatively minor violation.
   c. will not bias the estimates of the slope and intercept.
   d. is likely to affect the t-tests for the slope and intercept.
   e. All of the above are correct.

8. If the disturbances are autocorrelated, then it is *incorrect* to say that
   a. the OLS estimators for the slope will be unbiased.
   b. the OLS estimators for the slope will be consistent.
   c. the OLS estimators for the slope will be efficient.
   d. the OLS estimator of the intercept is unbiased.
   e. the OLS formulas can still be used for the slope and intercept.

9. If the disturbances are heteroskedastic, the OLS estimators of β_0 and β_1 would no longer be
   a. unbiased.
   b. consistent.
   c. normally distributed.
   d. efficient.
   e. BLUE.

10. If a variable Z is inadvertently omitted from the regression model, then
   a. the slope estimate is biased if Z_i is correlated with X_i.
   b. the intercept estimate is biased.
   c. the slope estimate is not consistent if Z_i and X_i are correlated.
   d. the intercept estimate is not consistent.
   e. all of the above are correct.

11. If the functional form of the regression is misspecified, then
   a. the OLS estimators for the slope and intercept will be unbiased.
   b. the OLS estimators for the slope and intercept will be consistent.
   c. the OLS estimators for the slope and intercept will be efficient.
   d. the regression results will be completely invalid.
   e. the consequences may be considered relatively minor.

12. If the independent variable X is stochastic instead of being fixed then
   a. the consequences are minor unless X_i is correlated with u_i.
   b. the probability plot will reveal non-normality of the residuals.
   c. the distribution of the slope estimate is distinctly non-normal.
   d. the t-statistic for the slope is reliable if X and the error term are correlated.
   e. Two of the above are correct.

Glossary of Terms

ANOVA table Summary of decomposition of variance in a regression, showing total sum of squares and its sources (regression, error) along with degrees of freedom and mean squares. See **Error sum of squares** and **Regression sum of squares**.

Autocorrelation Non-independent errors in a regression model. Evidence of autocorrelation may be sought by examining the residuals in a regression. Runs of residuals with the same sign (e.g., + + + – – –) would suggest *positive* autocorrelation, while runs of residuals with alternating sign (e.g., + – + – + –) would suggest *negative* autocorrelation. Autocorrelation (particularly positive) is common in time series data, and is considered a major violation. See **Mutually independent disturbances**.

Coefficient of determination See **R-squared**.

Confidence interval Upper and lower limits that are expected to enclose the true model. For a 95% confidence interval, on average, 95 out of 100 such intervals will contain the true model in repeated sampling. See **Confidence level** and **Prediction interval**.

Confidence level Desired probability of enclosing an unknown parameter, equal to $1 - \alpha$. Typical confidence levels are 90%, 95%, and 99%.

Correlation coefficient Measure of association between two variables, equal to the sample covariance divided by the product of the sample standard deviations of X and Y. A correlation of –1 indicates a perfect inverse relationship, 0 indicates no relationship, and +1 indicates a perfect direct relationship. The formula for the correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Degrees of freedom In a regression ANOVA table, *total* degrees of freedom is $n - 1$, *error* degrees of freedom is $n - k - 1$, and the *regression* degrees of freedom is k, where n is the sample size and k is the number of independent predictors in the model.

Dependent variable In a regression, the variable (denoted Y) that is placed on the left-hand side of the equation and is assumed to be affected by the independent variable X.

Error sum of squares In a regression ANOVA table, the error sum of squares is the portion of the total sum of squares that is not explained by the model.

Estimated coefficient Sample statistic used to estimate a parameter of the regression model. The estimated regression coefficients are denoted $\hat{\beta}_0$, $\hat{\beta}_1$. See **Ordinary least squares**.

Fixed X In a regression, it is assumed that the independent variable X is non-stochastic. That is, X does not have a distribution, but is assumed to have predetermined values.

F statistic In a regression ANOVA table, the ratio of the *regression* mean square to the *error* mean square.

Heteroskedasticity Non-constant variance of errors in a regression. If the underlying constant variance is denoted σ^2, some commonly assumed patterns are $k X_i^c \sigma^2$ (variance proportional to a power of X_i), $k E(Y)_i^c \sigma^2$ (variance proportional to expected value of Y), $k Z_i^c \sigma^2$ (variance proportional to a power of another variable), or $k i^c \sigma^2$ (variance proportional to a power of the observation order). If the disturbances are heteroskedastic, the OLS estimators are still unbiased, consistent, and normally distributed, but are not efficient. Also, in a given sample, the predictive validity of a confidence interval for the expected value of Y is in doubt over some regions of the X range. Heteroskedasticity is considered a major violation.

Homoskedasticity Constant variance of errors in a regression model. If the errors are homoskedastic, there should be no discernible pattern in the residuals on a scatter plot that displays the fitted regression. Heteroskedasticity would be suggested by a "fan-out" pattern of residuals (increasing variance as we move to the right) or a "funnel-in" pattern of residuals (decreasing variance as we move to the right).

Independent of fixed X Assumption about the disturbances of a regression, which states that the errors are unrelated to the values of the independent variable X, where X is assumed to be a non-stochastic variable. See **Fixed X**.

Independent variable In a regression, the variable (denoted X) that appears on the right-hand side of the equation and is thought to cause variation in the dependent variable.

Intercept Value of the dependent variable when the independent variable in the regression model is zero. However, zero values may have little or no meaning for some predictors. Although it is often included by default, an intercept is not required in a regression model.

Misspecified model Omission of a relevant predictor or the incorrect functional form (e.g., assuming linearity when the relationship is non-linear). See **Zero mean**.

Mutually independent disturbances Absence of a relationship between errors in a regression. A violation of this assumption is called *autocorrelation*. First-order autocorrelation is the most common form: $u_t = \rho u_{t-1} + v_t$ where ρ is a constant such that $-1 \le \rho \le +1$ and v_t is a well-behaved disturbance (normally distributed, mutually independent, homoskedastic, zero mean). If $\rho = -1$ then a disturbance in period $t - 1$ yields an identical disturbance in period t but opposite in sign (negative autocorrelation) plus a random disturbance. If $\rho = 0$ then there is no carryover from period $t - 1$ to period t (non-autocorrelated). If $\rho = +1$ then a disturbance in period $t - 1$ yields the same disturbance in period t (positive autocorrelation) plus a random disturbance. If this violation exists, the OLS estimates are still unbiased, consistent, and normally distributed, but no longer efficient. Moreover, because of the increased variance, a high degree of positive autocorrelation may make it impossible to obtain reasonable estimates of your parameters except in very large samples. See **Autocorrelation**.

Non-normal errors Violation of a basic regression assumption that may affect confidence intervals and hypothesis tests. Evidence may be found in the residuals from a fitted regression. If the disturbances are not normally distributed, the OLS estimators are unbiased and consistent, but are not efficient. However, they are asymptotically efficient, and in large samples the confidence intervals for $E(Y|X)$ may be reasonable. Non-normal disturbances are not usually considered a major violation, since the OLS method is robust to considerable non-normality. Evidence of non-normality may be sought by examining the histogram of residuals or the normal probability plot.

Ordinary Least Squares (OLS) Method of estimating a regression that guarantees the smallest possible sum of squared residuals. The residuals sum to 0 using the OLS method.

Parameter Numerical constant needed to define a particular model or distribution. In a regression model, the parameters are the intercept and the coefficient of the independent variable. They are denoted β_0, β_1. See **Estimated coefficient**.

Prediction interval Upper and lower limits that are expected to enclose the observed data points. For a 95% prediction interval, on average, 95 out of 100 of the observations will lie within the interval. See **Confidence level** and **Confidence interval**.

Predictor An independent variable in a regression model. See **Binary variable**.

Probability plot Comparison of each observed residual with the value that would be expected assuming that it came from a normal distribution. To construct a probability plot, calculate the inverse of the hypothesized normal distribution function for the i^{th} residual, and plot it against the observed i^{th} residual. This is done for all n residuals to produce a scatter plot. If the hypothesized normal distribution is correct, the scatter plot should be roughly linear along the diagonal. This is a simple, powerful visual test for normality of the sample residuals.

P-value Probability (usually two-tailed) of type II error if we reject the null hypothesis of a zero parameter. Thus, a small p-value (such as 0.01) would incline us to reject the hypothesis that the true parameter is zero.

Regression sum of squares In a regression ANOVA table, the regression sum of squares is the portion of the total sum of squares that is explained by the model.

Residual Difference between an actual and estimated value of the dependent variable.

Residual plot Scatter plot of the residuals against a predictor, used to check the residuals for evidence of a violation known as *heteroskedasticity* (non-constant residual variance). For k predictors we get k residual plots. To simplify matters, statisticians sometimes just look at a plot of the residuals against the fitted Y, though this method reveals less than the k plots. If the residuals are *homoskedastic*, there should be no discernible pattern. See **Heteroskedasticity**.

R-squared Also called the coefficient of determination, it is the ratio of the *regression* sum of squares to the *total* sum of squares. R^2 near 0 indicates the fit is poor while R^2 near 1 indicates the fit is good.

Standard error Estimate of the standard deviation of the stochastic disturbances, using the square root of the sum of the squared residuals, divided by $n - k - 1$. It is often called the *standard error of the estimate* to distinguish it from the standard error of each regression coefficient. See **Degrees of freedom**.

Standardized residual For each observation, the residual divided by the estimated standard error of the estimate.

Sum of squares In a regression ANOVA table, the total sum of squares is decomposed into two parts: *regression* sum of squares and *error* sum of squares.

Truncated normal Normal distribution whose tails are cut off. For example, a standard normal that is truncated at ± 2.5 would never have any outliers and would have reduced variation.

t-value Ratio of an estimated coefficient in a regression model to its standard error, used to test the null hypothesis that the parameter is zero. This ratio is distributed as Student's t if the parameter is zero. A large t-value would suggest that the true parameter is not zero.

Zero mean The mean of the disturbances may be thought of as the "center" of the errors around the true regression line. If this mean is non-zero, or if it is non-constant, the OLS estimates will lose some of their desirable properties. Non-zero mean may be an annoyance or a major violation, depending on its form. A constant non-zero mean is a trivial problem, since it will merely be reflected in the intercept estimate and may not be noticeable. An omitted variable Z is a much worse manifestation that can make the OLS slope estimate biased and inconsistent (although if Z and X are uncorrelated it is unbiased and may have other desirable asymptotic properties). Incorrect functional form for the equation can also lead to serious consequences, since the wrong model will be estimated. Bias due to incorrect functional form can be severe.

Solutions to Self-Evaluation Quiz

1. a Read the Overview of Concepts.
2. c Do Exercise 5. Read the Overview and Illustration of Concepts.
3. d Do Exercises 5–7. Read the Overview and Illustration of Concepts.
4. a Do Exercises 5–7. Read the Overview and Illustration of Concepts.
5. a Do Exercises 1–4. Read the Illustration of Concepts.
6. d Do Exercise 5. Read the Illustration of Concepts.
7. e Do Exercises 8–11, and 26–29.
8. c Do Exercises 21–25.
9. d Do Exercises 16–20.
10. e Do Exercises 30–34.
11. d Do Exercise 39–44.
12. a Do Exercises 35–38.