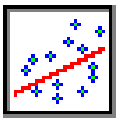
CHAPTER 15



Visualizing Simple Regression

CONCEPTS

- Estimated Model, True Model, Conditional Mean, Dependent Variable, Independent Variable, Ordinary Least Squares (OLS), Disturbance, Residual, Estimator, Parameter, Confidence Interval, Prediction Interval, Distribution of $\hat{\beta}_0$, Distribution of $\hat{\beta}_1$

OBJECTIVES

- Understand OLS terminology
- Understand how sample size, standard error, true parameters, and the range of the independent variable affect estimation accuracy
- Understand the sampling and population distribution of OLS estimators
- Understand the difference between confidence intervals for E(y|x) and prediction interval for y|x and the reason for their parabolic shape

Overview of Concepts

A simple regression model relates a **dependent variable** Y to an **independent variable** X. This model is used when the mean of each observation or value of y_i is conditional upon the value of x_i. This **conditional mean** is written as $E(y_i|x_i) = \beta_0 + \beta_1 x_i$, where β_0 is the intercept and β_1 is the slope. β_0 and β_1 are unknown **parameters**. The conditional mean is not observable. However, the actual values of y_i are observed and do not generally lie at this conditional mean. The deviations of the values of y_i from this conditional mean are called **disturbances**, or errors, and are denoted u_i. Combining the conditional mean with these disturbances results in the **true model** $y_i = E(y_i|x_i) + u_i = \beta_0 + \beta_1 x_i + u_i$. Since the conditional mean $E(y_i|x_i)$ is not known, the disturbances are not known or observed.

An example of this model occurs in macroeconomics. John Maynard Keynes theorized that the amount spent by all consumers (consumption) depends upon their total income. In statistical terms this is written as $E(\text{Consumption}|\text{Income}) = \beta_0 + \beta_1$ Income. This is the conditional mean. The true model is Consumption = $\beta_0 + \beta_1$ Income + u.

If data were collected on consumption and income, β_0 and β_1 could be estimated using the method of **Ordinary Least Squares (OLS)**. OLS is an estimation technique that is used to derive an **estimator** (an equation) for β_0 and β_1. It does this by minimizing the sum of the squared deviations between y_i and the estimated regression line. The OLS estimates obtained for β_0 and β_1 are generally denoted $\hat{\beta}_0$ and $\hat{\beta}_1$. Using this notation results in the **estimated model** $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. This is the equation of the estimated conditional mean or estimated regression line. The difference between each estimated value $\hat{y}_i$ and the corresponding observed value of y_i is called the **residual**. The residual $\hat{u}_i$ equals $y_i - \hat{y}_i$. The residuals sum to zero. The OLS technique guarantees unbiased estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize $\Sigma\hat{u}_i^2$.

Just as the sample mean $\bar{Y}$ is a point estimate of the population mean μ_y, $\hat{y}_i$ is a point estimate for the conditional mean $E(y_i|x_i)$. Similarly, just as an interval estimate for μ_y is based on a confidence interval using $\bar{Y}$, an interval estimate for $E(y_i|x_i)$ is based on a **confidence interval** using $\hat{y}_i$. However, using regression analysis we can also create a **prediction interval** for $y_i|x_i$. The confidence interval describes the location of the *conditional mean*, while the prediction interval describes the location of the *individual observation* y_i for a given value x_i.

Recall that the Central Limit Theorem states that the estimator $\bar{Y}$ has a normal distribution (if n is large enough) with a mean μ_y and a standard error $\sigma/\sqrt{n}$. Similar results have been derived for the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. Statisticians have derived the **distribution of $\hat{\beta}_0$** and the **distribution of $\hat{\beta}_1$** when the disturbances, or errors, u_i are independently and identically distributed with a normal distribution having a mean of zero and a variance of σ^2. In that case, $\hat{\beta}_0$ will be normally distributed with a mean β_0 and a standard error $\sigma\{\Sigma x_i^2 / [n\,\Sigma(x_i - \bar{X})^2]\}^{0.5}$, and $\hat{\beta}_1$ will be normally distributed with a mean β_1 and a standard error of $\sigma/[\Sigma(x_i - \bar{X})^2]^{0.5}$, where n is sample size and σ is the standard error of the true model.

Illustration of Concepts

Suppose we have a sample of income tax returns for 20 married couples (no dependents). We hypothesize the model Taxes = $\beta_0 + \beta_1$ Income + u, where Taxes are the **dependent variable**, Income is the **independent variable**, u is the **disturbance** (or error) term, and β_0 and β_1 are the unknown **parameters**. In this model, β_1 is the marginal tax rate on each dollar of additional income and β_0 is the average tax for a married couple without dependents with zero income.

Suppose this model is estimated using **ordinary least squares (OLS)** with 20 pairs of observations on taxes and income to obtain the **estimated model** Predicted Taxes = 580.4 + 0.035 Income. The estimated intercept (580.4) and slope (0.035) are the OLS estimates of β_0 and β_1. The OLS **estimators** of these parameters are the equations that are used to obtain the estimates, not the estimates themselves. The **residual** is the difference between the predicted tax and actual tax each family paid to the state. On a graph, a residual is the vertical distance between a data point and the estimated model. Figure 1 shows a scatter plot with the estimated model and two residuals labeled.

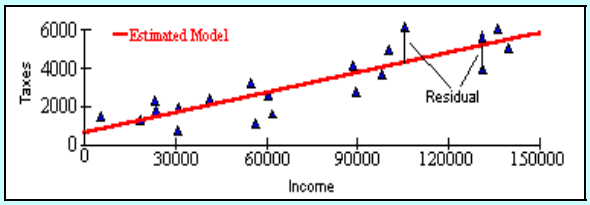


Figure 1: Estimated Model and Residuals

For the estimated intercept, the t-value for the estimated intercept is 1.52, suggesting that the true intercept is possibly close to zero. For the estimated slope, the t-value is 7.71, suggesting that the true slope is positive.

Although the **true model** is not known, based on the structure of the tax law, a financial economist in the state treasurer's office believes that the **conditional mean** is E(Taxes|Income) = –135 + 0.045 Income. If this conditional mean is correct, the true model becomes Taxes = –135 + 0.045 Income + u. The u in this model is the disturbance, or error, term. If the economist is correct, and if the disturbances are normally distributed with a mean of zero and a constant variance, then the **distribution of $\hat{\beta}_0$** is normal with a mean of 135 and the **distribution of $\hat{\beta}_1$** is normal with a mean of 0.045.

Using the estimated model, a **confidence interval** for the conditional mean and a **prediction interval** for y can be constructed (assuming that the disturbances are normally distributed). A 99% confidence interval is shown in Figure 2 along with the line of conditional means. Note that the confidence interval contains this conditional mean. Figure 3 shows a 90% prediction interval. Note that, as expected, one point and maybe two points are outside the interval's boundary.

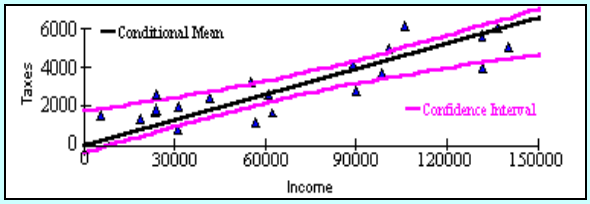


Figure 2: 90% Confidence Interval and Conditional Mean

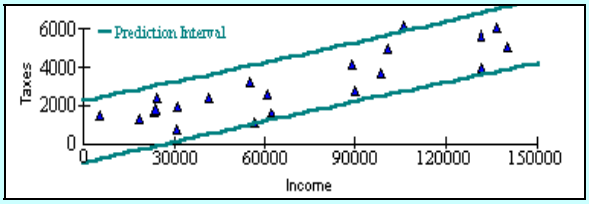


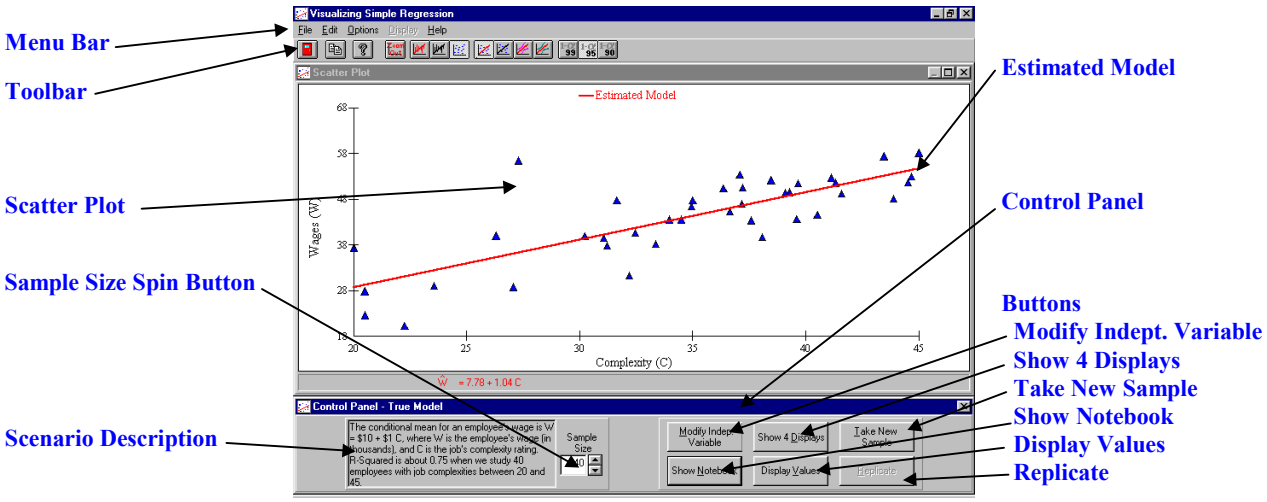
Figure 3: 90% Prediction Interval

Orientation to Basic Features

This module generates data from a true regression model that you create and then estimates the model. 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, and correlation coefficients.

1. Opening Screen

Start the module by clicking on the module's icon, title, or chapter number in the *Visual Statistics* menu and pressing the Run Module button. When the module is loaded, you will be on the introduction page of the Notebook. Read the questions and then click the Concepts tab to see the concepts that you will learn. Click on the Scenarios tab. Click on Firms, select a scenario, read it, and press OK. A scatter plot appears with a Control Panel at the bottom. Click on the scatter plot. A toolbar appears at the top of the screen.



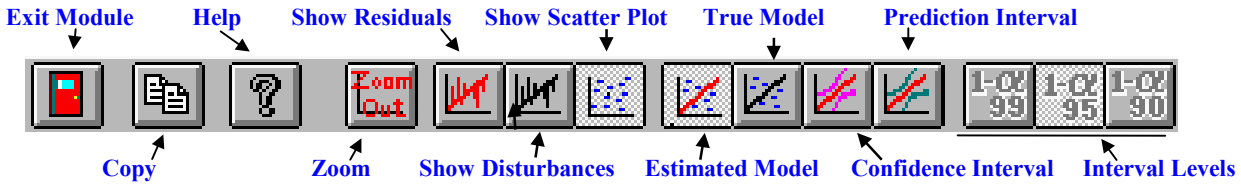
2. Control Panel

The Control Panel contains a short description of the scenario, a Sample Size spin button, and six buttons. Press the Modify Indept. Variable button to bring up a new window that controls the independent variable's minimum value, maximum value, type of variable (integer or decimal), and whether the sample contains the x end points. 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.

3. Scatter Plot

Click on the scatter plot graph. Its header changes color indicating that the display is active. The red line is the estimated model. The estimated model equation is shown below the graph. New toolbar buttons appear on the toolbar to the right of the ? button. Move the cursor over a button and its description appears in a tooltip. The ? button and the ones to its left are always visible. The buttons to its right vary depending on the display that is active. The Exit Module button ends the program, the Copy button copies the active display, and the ? button activates Help. The remaining toolbar buttons control the scatter plot display. The Zoom button changes

the scale of the X and Y axes. This lets you show the X origin or the minimum value of X. The Show Residuals button displays the residuals. The Show Disturbances button displays the disturbances. The Scatter Plot button displays the scatter plot. Only one of these three buttons is active at a time. When the Scatter Plot button is active, all of the toolbar buttons to its right are active. The Estimated Model button displays or removes the estimated model line from the scatter plot. The True Model button displays or removes the true model from the scatter plot (its equation is at the bottom of the display). The Confidence Interval button displays the confidence interval for E(y|x), and the Prediction Interval button displays the prediction interval for y|x. The Interval Level buttons control 1 – α (99%, 95% is the default, 90%).

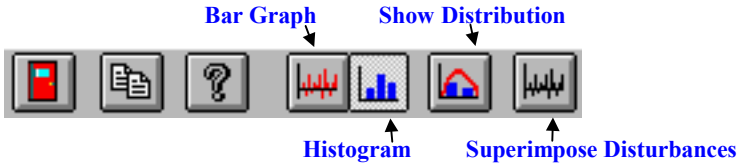


4. **Additional Displays**

Press the Show 4 Displays button to display four quadrants. The scatter plot is in the upper left quadrant, an ANOVA table is in the lower left quadrant, a histogram of the standardized residuals is in the lower right quadrant, and a Control Panel is in the upper right quadrant. Two buttons on the panel have changed. The Show 4 Displays button has been renamed Main Display (returns you to the scatter plot and Control Panel display), and the Replication button has become active.

5. **Standardized Residuals Display**

Activate the standardized residuals histogram by clicking on the display. Four new toolbar buttons appear to the right of the ? button. The first two determine whether a bar graph or a histogram of the standardized residuals is shown. The third button (active only if the histogram is chosen) superimposes a standard normal distribution ($\mu = 0$, $\sigma = 1$) for comparison. The last button superimposes the disturbances on the display.



6. **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.

7. **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.

8. **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

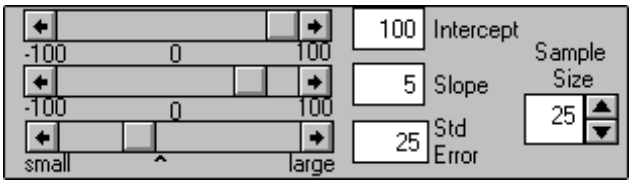
1. **Enlarging a Graph**
   The scatter plot or standardized residuals display can be maximized (or minimized) with the windows buttons in the upper right corner. When the scatter plot is maximized, a Take New Sample button appears.

2. **Replication**
   a. Press the Replicate button. Read the Hint that appears. A Control Panel enables you to start an experiment, set the number of replications (50, 100, 200, 500, or 1000), or exit replication mode. Press the Start Experiment button. Press the Pause Experiment button to halt the experiment and Continue Experiment button to resume sampling. Press the Finish Experiment button and the experiment is completed without building the displays one replication at a time. The other quadrants contain a graph of all estimated models, a histogram of the estimated slopes, and a histogram of the estimated intercepts.
   b. Click on the Estimates of Model display (upper left quadrant). Five new icons appear on the toolbar. The first will draw the True Model on the graph and the second will draw an Interval about the True Mean (the percentage is indicated by the last three icons) to illustrate why regression confidence intervals have a parabolic shape.

   c. Click on the histogram of estimated slopes or intercepts to add two new toolbar buttons. The first superimposes a normal curve on the histogram and the second brings up a table that compares the observed and expected frequencies in each interval.

   d. Right-click on any quadrant. A menu of other displays appears. Select one to replace the display you clicked on. You can superimpose a chi-square distribution on the histogram of estimated variances, but there is no distribution for the correlation coefficients or R^2.
   e. Press the Exit Replication Mode button to return to nonreplication mode.

3. **Do-It-Yourself Controls**
   Click on the Show Notebook button and select the Do-It-Yourself tab. Click OK. A scroll bar for the intercept, slope, and standard error replace the scenario description on the Control Panel. If you prefer to control the error by setting the desired R^2 rather than the standard error, click on Options on the menu bar and select Set Error Using and then Desired R-Squared.



Basic Learning Exercises Name _____

The Estimated Model

1. If you are displaying the four quadrants, press the Main Display button. Only the estimated model should be displayed on the scatter plot (use the toolbar buttons to control this feature). Press the Show Notebook button, select the Scenarios tab, click on Firms, and select the Wage vs. Job Complexity scenario. Read the scenario and click OK. What is the estimated model? **Hint:** The estimated model is displayed below the scatter plot.

2. Interpret this estimated model. **Hint:** Press the Zoom Out toolbar button.

3. What are the advantages and disadvantages of seeing the entire horizontal axis?

4. Use the toolbar button to either show or hide the origin. Press the Show Residuals toolbar button. a) What are residuals? b) How are they calculated. c) How many are there? d) Why do the residuals sum to zero? **Hint:** See Ordinary Least Squares in Help or the Glossary.

5. Press the Show 4 Displays button. A histogram of the standardized residuals is being displayed. a) Why have the residuals been standardized? b) How are the residuals standardized? c) Click on the Standardized Residuals window (lower right quadrant) to activate the display. Use the toolbar to change the display to a bar graph of the standardized residuals. What is the difference between the bar graph display and the histogram? **Hint:** Use the Glossary.

6. What is the relationship between this display and the scatter plot of the residuals?

7. What is the R^2 and standard error of your estimated model? What is the standard error and t-value of your estimated intercept and slope? Is the estimated intercept statistically different from zero? Is the slope statistically different from zero?

   R^2 = _____ Standard Error of Model _____
   β_0: Standard Error _____ t-value _____ Decision _____
   β_1: Standard Error _____ t-value _____ Decision _____

8. Interpret the estimated standard error of the model and the R^2?

The True Model

9. The true model in this scenario is Wage = $\beta_0 + \beta_1$ Complexity + u. a) Identify the parameters of this model. b) What does β_1 represent? c) Why are the parameters of the true model rarely known? d) What does the u represent?

10. In this module we know the conditional mean of the true model. This enables us to see the relationship between the true model and the estimated model. Press the Show Notebook button and reread the scenario. In this hypothetical situation what are the values of β_0 and β_1? What is the true model? Press the Cancel button to return to your previous screen.

11. Press the Show Scatter Plot toolbar button to bring back the scatter plot with the estimated model displayed. Press the Show True Model toolbar button. Why isn't the estimated model exactly equal to the true model?

12. Press the Show Disturbances toolbar button. What are disturbances? Can they be calculated?

13. Click on the Standardized Residuals window to activate it. The bar graph of the residuals should be displayed. Press the Superimpose Disturbances toolbar button. The disturbances are shown in black. What is the relationship between the black bars and the scatter plot displaying the disturbances? Compare the residuals (shown in red) and the disturbances.

14. Press the Histogram of Residuals toolbar button. A histogram of the disturbances is superimposed in green on top of the histogram of residuals. Compare the two histograms. In this model the disturbances are normally distributed with a mean of 0 and a standard deviation of 5. Press the Show Distribution toolbar button to superimpose this distribution on the histogram. Press the Take New Sample button 10 times. In general, are the disturbances and residuals approximately distributed as expected?

15. Activate the Standardized Residuals window. Press the Bar Graph of Residuals toolbar button. The true and estimated models should be displayed on the scatter plot. Press the Take New Sample button 10 times. Each time, record each R^2 value and compare the true and estimated models, and the disturbances and residuals. In this case, is the estimated model a good predictor of the true model? Are its residuals good predictors of the disturbances?

    R^2: _____ _____ _____ _____ _____ _____ _____ _____ _____ _____

16. Press the Show Notebook button and select the Test Scores vs. Job Performance scenario. Read the scenario. Click OK. Repeat exercise 15 using this scenario. Did having a low R^2 change your answers appreciably?

    R^2: _____ _____ _____ _____ _____ _____ _____ _____ _____ _____

Confidence Intervals and Prediction Intervals

17. Activate the scatter plot, click on the Confidence Interval for E(y|x) toolbar button, and push the 90% Confidence Level toolbar button. Click the maximize button in the upper right corner of the scatter plot display. Only the estimated model and confidence interval should be displayed. Why is the estimated model in the middle of the confidence interval?

18. Remove the estimated model and display the true model on the scatter plot by using the toolbar buttons. Press the Take New Sample button in the lower right corner of the computer screen. Note the relationship between the confidence interval and the true model. Repeat this 20 times. What did you observe about the confidence interval and the true model?

19. Press the 95% Confidence Level toolbar button. Repeat exercise 18.

20. Interpret a 95% confidence interval.

21. Press the Prediction Interval for y|x toolbar button. Remove the true model from the display. What is the relationship between the confidence interval and prediction interval?

22. Record the number of data points that lie outside the prediction interval. Repeat this process for nine more samples. Of the 500 observations, what percentage were outside the interval?

Trial Number	1	2	3	4	5	6	7	8	9	10
Number outside interval	___	___	___	___	___	___	___	___	___	___

_____%

23. Describe what a 95% prediction interval shows.

Intermediate Learning Exercises Name _____

Replication Experiment

24. Press the Show Notebook button, select the Scenarios tab, click on Firms, and select the Wage vs. Job Complexity scenario. Read the scenario and click OK. What is the true model in this scenario?

25. Press the Replicate button. Read the hint if it appears. Set Number of Replications to 500 (use 1,000 if you have a fast computer or 100 if you have a slow computer). Press the Start Experiment button. After the experiment has finished, activate the Estimates of the Model window (upper left quadrant). This display shows every model you estimated during the experiment. Describe the shape of the shaded area. Press the Conditional Mean toolbar button. Where is the conditional mean located in relation to the shaded area? Press the Interval about the Conditional Mean toolbar button. Describe the interval's shape. Compare the interval's shape with the shape of the shaded area.

26. Why do the various OLS estimates of the model form an hourglass shape? **Hint:** Use 50 replications and pay special attention to estimated models that are very flat or steep.

27. Reset replications and start the experiment. Describe the shape of the histogram of estimated slopes. What is its mean and standard error? How does this show that the estimated slope is an unbiased estimator β_1 (read exercise 24)? **Hint:** Read the bottom labels on the histogram.

28. Assume that you are testing H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$. Using two standard deviations (approximate α-level is 0.05) as the critical values, what percentage of the estimated slopes would reject H_o? Why is this the power of the test? Activate the histogram window and use the toolbar buttons to show the true distribution and the table of frequencies. Why are the expected and observed results similar? **Hint:** Consider the regression assumptions.

29. Describe the shape of the histogram of estimated intercepts. What is its mean and standard error? Is it an unbiased estimator? Why, or why not?

30. Assume that you are testing H_0: $\beta_0 = 0$ versus H_a: $\beta_0 \neq 0$. Using two standard deviations (approximate α-level is .05) as a critical value, what percentage of the estimated slopes would reject H_o? Why is the power so low? What lesson can you learn from this result?

31. Right-click on the Histogram of Intercepts and select Histogram – R-Squared. Describe the shape of the histogram. Decrease the sample size to 5. Press Start Experiment. Describe the shape of the histogram. Increase the sample size to 100. Press Start Experiment. Describe the shape of the histogram. Why don't we use R^2 as a test statistic?

Advanced Learning Exercises Name _____

Effect of Sample Size

32. Press the Show Notebook button and select the Number of Bar Scanning Errors scenario. Read it and press OK. Set sample size to n = 15. Run an experiment with 1000 replications and use the histograms to visually estimate the mean and range of the estimated intercept, slope, variance, R^2, and correlation coefficient (r). Double the sample size (to 30) and rerun the replication experiment. Double the sample size again (to 60) and rerun the experiment.

n	Statistic	Intercept	Slope	Variance	R^2	r
15	Mean					
	Range					
30	Mean					
	Range					
60	Mean					
	Range					

33. What is the effect of increasing the sample size on the mean value of the estimated intercept, estimated slope, estimated variance, R^2, and r? What is the effect on the range of the estimated intercept, estimated slope, standard error of the model, R^2, and r? Explain these results.

Effect of Range of X

34. Press the Show Notebook button and select a scenario. Read it and press OK. Conduct an experiment with 1000 replications and use the histograms to estimate the mean and range of the estimated intercept, slope, variance of the model, R^2, and r. Press the Modify Indept. Var. button. Increase the maximum value and decrease the minimum value so that the range is doubled, and repeat the experiment. Again, double the range and repeat the experiment.

X Range	Statistic	Intercept	Slope	Variance	R^2	r
Small	Mean					
	Range					
Medium	Mean					
	Range					
Large	Mean					
	Range					

35. What is the effect of increasing the range of the independent variable on the mean value of the estimated intercept, slope, variance of the model, R^2, and r? What is its effect on the range of the estimated intercept, slope, standard error of the model, R^2, and r?

36. Explain the results in exercise 35.

Distribution of Estimated Variance

37. Run a replication experiment. What is the distribution of the estimated variance of the model? Why does it have this distribution? Why isn't the histogram of the standard error used instead of the variance?

Distribution of the Correlation Coefficient

38. Run a replication experiment. Examine the shape of the histogram of correlation coefficients. Use several different sample sizes and rerun the replication experiment. Each time, reexamine the shape of the histogram of correlation coefficients. Is the histogram normally distributed?

Individual Learning Projects

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

1. Explain and illustrate the difference between the confidence interval for E(y|x) and the prediction interval for y|x. What does each mean? What is the importance of the confidence level $1 - \alpha$? Why are both intervals parabolic in shape? **Hint:** The replication feature may help you answer this question.
2. Explain and illustrate how the slope, intercept, standard error of the model, sample size, and range of the independent variable (take care not to change its midrange) affect the distribution of an estimator. Investigate *either* the estimated slope, estimated intercept, or estimated variance of the model. Use the Do-It-Yourself controls to create a true model. Run a replication experiment using this model. The histogram of each estimator (either slope or intercept) is your benchmark. Change one of the factors you are to investigate (larger changes are generally better). Run another replication experiment. Examine the estimator's histogram. Return the factor to its original setting and change another factor. Continue the process for all five factors you are investigating. What effect did each factor have on the distribution of the estimator? Which factor affected the distribution the most? The least?
3. Explain and illustrate how the slope, intercept, standard error, sample size, and range of the independent variable affect the R^2 statistic. Use the Do-It-Yourself controls to create a true model. Run a replication experiment using this model. The histogram of R^2 is your benchmark. Change one of the factors you are to investigate (larger changes are generally better). Run another replication experiment. Examine the histogram of R^2. Return the factor to its original setting and change another factor. Continue the process for all five factors you are investigating. What effect did each factor have on the distribution of R^2? Did it change its mean, its standard error, or its shape? Which factor affected the distribution the most? The least?

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. Use presentation software or large poster boards to display your results. 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 project is for a team of three. Investigate the effect sample size has on the distribution of five statistics: the estimated slope, intercept, variance of the model, R^2, and correlation coefficient. The team should decide on a true model (its slope, intercept, standard error of the model, minimum value of X, and maximum value of X). Each team member will run a replication experiment using two different sample sizes. For each experiment the distribution of the five statistics will be examined. The team should cover the range of sample sizes from 2 to 100. What effect did increasing the sample size have on each statistic? Did it affect its mean, its standard error, or its shape? What statistic was affected the most? The least? If your team was hired to estimate a model, what do these results suggest about the sample size you should use?
2. This project is for a team of three to four. Investigate the effect the independent variable has on the distribution of five statistics: the estimated slope, intercept, variance of the model, R^2, and correlation coefficient. The team should decide on a true model (its slope, intercept, standard error of the model), sample size, and the midrange of the independent variable (between 50 and 100). Each team member will use a different range for the independent variable (the team should cover the range 10 to 100). The team member will modify the independent variable so that it has the correct range and midrange and will then run a replication experiment. The team member will then run a second replication experiment after modifying the independent variable so that its midrange is *five* times the first midrange but has the same range. For each experiment, the distribution of the five statistics will be examined. What effect did increasing the midrange of the independent variable have on each statistic? What effect did increasing the range of the independent variable have on each statistic? Did it affect its mean, its standard error, or its shape? What statistic was affected the most? The least? If your team was hired to estimate a model, do these results suggest that you should use an independent variable with a large variance or a small variance?
3. This project is for a team of two. Investigate the distribution of the correlation coefficient r. The team should decide on the initial true model (intercept, slope, and standard error), minimum value of X, maximum value of X and sample size. Use the Do-It-Yourself controls to set up this model. Run a replication experiment. The histogram of r from this experiment is your benchmark. Each team member should select another value of the six factors (large changes generally work better). A replication experiment should be run using each one of these new factor values (always return to the initial true model). How did each factor affect the distribution of r? Did it affect its mean, its standard error, or its shape? Did it affect the probability that r would be judged to be different from zero? **Hint:** Remember the rule of thumb that an estimate that is within 2 standard errors of zero is statistically equal to zero.

Self-Evaluation Quiz

1. In the simple regression model, the disturbance is *not* assumed to
   a. be observable.
   b. be normally distributed.
   c. have zero mean.
   d. have constant variance.
   e. be uncorrelated with X.

2. The conditional mean of y is written as
   a. E(y|x).
   b. y|x.
   c. μ.
   d. x|y.
   e. E(x|y).

3. Residuals have which characteristic?
   a. They give clues about unobservable disturbances.
   b. If their sum is non-zero, a mistake has been made in the OLS calculations.
   c. They are used to calculate R^2.
   d. All of the above.
   e. Only b and c.

4. The residuals can be thought of as predictors of
   a. the slope.
   b. the the true model.
   c. the disturbances.
   d. the intercept.
   e. none of the above.

5. In a simple regression, which does *not* suggest a relationship between X and Y?
   a. Small Error Sum of Squares and large Regression Sum of Squares in the ANOVA table.
   b. Large F statistic in the ANOVA table.
   c. Small p-value for the F statistic.
   d. Large p-value for the estimated slope.
   e. All of the above suggest a relationship.

6. Which is indicative of an *inverse* relationship between X and Y?
   a. A negative coefficient of determination.
   b. A negative estimated intercept.
   c. A negative p-value for the slope.
   d. A negative F statistic for ANOVA table.
   e. A negative correlation coefficient.

7. When x = 100 the confidence interval for E(y|x)
   a. is narrower than the prediction interval for y|x.
   b. is wider than the prediction interval for y|x.
   c. is equal to 50 for the lower interval and 150 for the upper interval.
   d. is an equal distance above and below the true model.
   e. None of the above.

8. A 95% prediction interval for y|x
   a. is inside a 90% confidence interval for E(y|x).
   b. will always contain 95% of your observations.
   c. is outside a 90% prediction interval for y|x.
   d. will on average contain the true model 95% or less of the time.
   e. has none of the above characteristics.

9. Which is *not* correct of the estimated slope of the regression line?
   a. It is divided by its standard error to get its t-value.
   b. It shows the change in y for a unit change in x.
   c. It is measured in the units of y per unit change in x (e.g., miles per gallon).
   d. It may effectively be regarded as 0 if its t-value is 8.
   e. It is an unbiased estimator of the true slope.

10. A 95% confidence interval for the E(y|x) means, on average, that
    a. 95% of the observations will be contained within the confidence interval.
    b. the confidence interval will contain the true mean 95% of the time.
    c. the true mean will lie within that interval 95% of the time.
    d. out of 100 confidence intervals the true mean will be contained in 95 of them.
    e. none of the above is correct.

11. If the assumptions about the disturbance term are correct then the Ordinary Least Squares method guarantees
    a. unbiased estimators.
    b. a large R^2.
    c. that the slope and intercept are minimized.
    d. all of the above.
    e. two of a, b, or c.

12. Which is *not* normally distributed for regression with normally distributed disturbances?
    a. R^2.
    b. Estimated slope.
    c. Estimated intercept.
    d. Residuals.
    e. All of the above are normally distributed.

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.

Coefficient Estimated value of a regression parameter (slope or intercept) based on sample data.

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

Conditional mean In a regression, the expected value of the dependent variable *given* the observed value(s) of the independent variable(s).

Confidence interval In a regression, the range of Y values that is expected to enclose the true *conditional mean* of Y. For a 95% confidence interval, on average, 95 out of 100 such intervals will contain the conditional mean. See **Confidence level**.

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

Correlation Measure of fit in a bivariate regression (–1 indicates a perfect inverse relationship, 0 indicates no relationship, and +1 indicates a perfect direct relationship). It is the square root of R^2 in the simple regression model. More generally, it is the sample covariance divided by the product of the sample standard deviations of X and Y.

Degrees of freedom In a regression ANOVA table, *total* degrees of freedom are $n - 1$, *error* degrees of freedom are $n - k - 1$, and the *regression* degrees of freedom are k, where n is the sample size and k is the number of predictors in the model ($k = 1$ for a bivariate 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 (denoted X). On the scatter plot, the dependent variable is customarily shown on the vertical axis.

Disturbance An unobservable random error. It is the difference between the conditional mean and the observed value of y (the dependent variable). Disturbances are assumed to be independent and normally distributed with zero mean and constant variance.

Error See **Disturbance**.

Estimated model Bivariate regression equation whose slope and intercept are the coefficients that are estimated using the ordinary least squares method from sample data.

Estimator In a simple regression model, the equations for $\hat{\beta}_0$ and $\hat{\beta}_1$ that are used with sample data to estimate the unknown parameters β_0 and β_1.

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

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 (denoted Y). On a scatter plot, the independent variable is usually shown on the horizontal axis.

Intercept Value of the dependent variable when $x_i = 0$ in the regression model $y_i = \beta_0 + \beta_1 x_i$. On a graph, the intercept β_0 is the point where the regression line intersects the Y-axis.

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

Parameters Values that define a particular distribution. For example, in a simple regression model, β_0 and β_1 are parameters that describe the conditional mean of the distribution of y.

Prediction interval In a regression, range of y_i values that would enclose the true *individual* y_i values a given percentage of the time (typically 90%, 95%, or 99%).

P-value Probability of Type I error if we reject the null hypothesis (e.g., that the true slope is zero). For example, a small p-value (such as 0.01) for the estimated slope would incline us to reject the hypothesis that the true slope is zero.

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

R-squared 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.

Scatter plot Visual display in which each observed pair of data values (x_i,y_i) is plotted as a symbol (e.g., a dot) at the correct coordinate on the graph. It allows visual assessment of "fit" of an estimated regression line to the observed data.

Slope The change in Y for a unit change in X in the bivariate model $Y = \beta_0 + \beta_1 X$. On a graph, the slope β_1 is the rise divided by the run.

Standard error Estimate of the standard deviation of the disturbances, calculated as the square root of the sum of the squared residuals divided by $n - k - 1$. See **Degrees of freedom**.

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

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.

True model An unobservable equation assumed to underlie the observed bivariate data.

T-value Ratio of an estimated coefficient in a regression model to its standard error (distributed as Student's t if the parameter is zero). A large t-value suggests that the parameter is not zero.

Solutions to Self-Evaluation Quiz

1. a Do Exercises 9–12. Read the Overview of Concepts.
2. a Read the Overview of Concepts and Illustration of Concepts.
3. d Do Exercises 4, 5, and 12–15.
4. c Do Exercises 12-15. Read the Overview of Concepts.
5. d Do Exercises 1–7.
6. e Do Exercises 1, 2, and 9.
7. a Do Exercises 17–23. Read the Illustration of Concepts.
8. c Do Exercises 21–23. Read both the Overview and the Illustration of Concepts.
9. d Do Exercises 1–9 and 24–27.
10. d Do Exercises 17–26. Read both the Overview and the Illustration of Concepts.
11. a Do Exercises 15, 16, and 24–27.
12. a Do Exercises 14 and 27–31. Read the Overview of Concepts.