SIMPLE LINEAR REGRESSION
EXCEL LAB #7

Overview

This lab is written for Excel 2010, which is available to students in the library. The notation => can be read as "go to" or "click on." This notation will most often be used when navigating the menu or toolbars in Excel. To indicate a command or icon that you might click on or search for in Excel, **bold** will be used. Likewise, anything that you are to type into Excel will be **bolded** in the instructions. Do not enter such text as bolded text unless the instructions ask you to do so.

Tutorial

1. Open Excel. Format columns A through L so that the column width is 10 and that the cells are center justified with Arial 9 point font.

2. In cell A1 enter **y**. In cell B1 enter **x**.

3. In cells A2 through A11 enter A2 = 50, A3 = 100, A4 = 30, A5 = 50, A6 = 10, A7 = 25, A8 = 85, A9 = 100, A10 = 70, A11 = 40.

4. In cells B2 through B11 enter B2 = 3, B3 = 10, B4 = 4, B5 = 9, B6 = 2, B7 = 4, B8 = 7, B9 = 9, B10 = 8, B11 = 4.

5. To calculate the regression statistics, we need to calculate several more columns of data. In cell A13 enter **=SUM(A2:A11)** and in cell A14 enter **=AVERAGE(A2:A11)**. Copy and paste these formulas into cells B13 and B14. Cell A14 shows that the average *y* value is 56, while cell B14 shows that the average *x* value is 6.

6. In cell C1 enter **y – ybar**, in cell D1 enter **x – xbar**, and then fill in these columns appropriately. That is, in cell C2 enter **=A2-$A$14** and in cell D2 enter **=B2-$B$14**. Notice how the dollar signs are used to keep the mean of each variable from rolling. These formulas can then be copied down the column through row 11. In cell C11, for example, the textbox should read **=A11-$A$14** and the value in the spreadsheet should be -16.

7. To check your work, copy and paste the formulas in cell A13 and A14 into cells C13 – D14. We know that the sum of deviations and the average deviation must be zero, and this is what we have in C13 and D13.

8. In cell E1 enter **(y-ybar)^2** and in F1 enter **(x-xbar)^2**. Then in cell E2 enter **=C2*C2**, and in cell F2 enter **=D2*D2**. Copy and paste these formulas down through row 11.

9. Copy cells C13 and D13, and paste them into cells E13 and F13. You should have that the sum of the squared deviations for y (in cell E13) is 8,990 and the sum of the squared deviations for x (in cell F13) is 76.

10. In cell G1 enter **(x-xbar)*(y-ybar)**. As this doesn't fit in the space of the cell very well, change column G's column width to 14.

11. In cell G2 enter **=C2*D2**. Copy this formula down through cell G11.

12. In cell G13 sum up the column: **=SUM(G2:G11)**. You should have that the sum of the multiplied deviations (in cell G13) is 695.

13. We are now ready to calculate the regression coefficient estimates. In cell A16 enter **b_1** and in cell A17 enter **b_0**.

14. In cell B16 we want to calculate the estimate of the slope coefficient β_1. As

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2},$$

in cell B16 enter **=G13/F13**.

15. In cell B17 we want to calculate the estimate of the constant coefficient β_0. We use the fact that the regression line must go through the average values to do this. Specifically, we know that
$$b_0 = \bar{y} - b_1\bar{x}.$$

Therefore, in cell B17 enter **= A14−B16*B14**.

16. Change the formats of cells B16 and B17 so that both estimates have 3 decimal places. To check your progress, you should have that b_1 = 9.145 and b_0 = 1.132

We need to calculate a standard error for the slope coefficient, but that is easiest to do after calculating the three sum of square ideas. So we turn to that now.

17. In cell H1 enter **y-hat**.

18. In cells H2 through H11, enter the predicted values of y. To do this, in cell H2 enter **=$B$17+$B$16*B2** (the dollar signs prevents the coefficient estimates from rolling) as
$$\hat{y}_i = b_0 + b_1\bar{x}.$$

Copy and paste this formula down the column. The predicted value for the first observation is 28.566.

19. In cell I1 enter **(y–ybar)^2**, in cell J1 enter **(y-yhat)^2**, and in cell K1 enter **(yhat–ybar)^2**. In cell I2 enter **=C2*C2** (as y-bar already been calculated), in cell J2 enter **=(A2–H2)^2**, and in cell K2 enter **=(H2-$A$14)^2**. Copy and paste these formulas down through the 11th row.

20. In cells I13 – K13 include the sum of each column (in cell I13 enter **=SUM(I2:I11)**, in cell J13 enter **=SUM(J2:J11)**, and in cell K13 enter **=SUM(K2:K11)**). Again the sum of the squared $y - y$bar deviations is 8,990. You should also have that

$$USS = \sum (y_i - \hat{y}_i)^2 = 2{,}634.4 \text{ and } ESS = \sum (\hat{y}_i - \bar{y})^2 = 6{,}355.6 .$$

21. Make columns J and K report all numbers up to 1 decimal points.

22. The three totals in cells I13, J13, and K13 correspond to the three sum of squared terms. So that these can be referred to easily, in cell I14 enter **TSS**, in cell J14 enter **USS**, and in cell K14 enter **ESS**.

23. In cell A19 enter **MSE**. In cell A20 enter **SER**. In cell A21 enter **R-sq**. Proceed to enter each of these values in cells B19, B20, and B21. In particular, in cell B19 enter **=J13/(10-2)**; in cell B20 enter **=sqrt(B19)**; and in cell B21 enter **=K13/I13**. Let MSE have 1 decimal point, but have SER and R-sq both have 3 decimal points. You should have that SER in cell B20 equals 18.147.

24. Now we want to calculate the standard error for the slope coefficient. In cell C16 enter **se**. In cell D16, calculate the standard error by entering **=B20/sqrt(F13)** as

$$s_{b_1} = \frac{SER}{\sqrt{\sum (x - \bar{x})^2}} .$$

You should have that the standard error of the slope coefficient is 2.082.

25. In cell c17 enter **t-stat** and in cell c18 enter **p-value**. Then in cell D17 enter the t-statistic associated with the slope coefficient by entering **=B16/D16**. And enter the p-value in cell D18 by entering **=TDIST(ABS(D17),10-2,2)**.
   - Recall that this is the better function to use (compared to NORMSDIST), because this function takes the small sample into account. Also notice that we wouldn't need to include **ABS(D17)** as the t-stat is already positive, but it is a good habit to get into using **ABS()** for whenever the t-stat is negative. The **10-2** is the degrees of freedom (sample size of 10, estimating 2 parameters b_0 and b_1). And the final **2** in the above formula indicates that we, as always, want a two-sided p-value.
   - Let cells D16 and D17 report up to 3 decimals, but have cell D18 report up to 4 decimals as this is the convention. As the t-stat is 4.393, we know that the p-value is extremely small. It would be tiny (less than 0.0001) if we had a large sample. Because the sample size is only 10, however, the p-value comes in at 0.0023.

26. ***Regression Analysis Using the Data Analysis Toolpak.***

At this point we have generated all of the necessary regression results. All of this could have been accomplished much more easily in Excel, however, by using the Data Analysis Toolpak.

1. Load Excel's Data Analysis ToolPak: **File** tab => **Options** => **Ad Ins** => **highlight Data Analysis ToolPak VBA** => **Go** => **check Data Analysis ToolPak** => **OK**.

2. Execute the regression: **Data** tab => **Analysis** box => **Data Analysis** => **Regression =>** enter A1:A11 for the y range and B1:B11 for the x range => click on labels => click on Output Range and enter cell A24 => **OK**.

All of the regression results are then provided by Excel in cells A24 through I41. Most importantly, notice how all of the results are the same:
- the coefficient estimates are identical as B40 = B17 and B41 = B16
- the standard errors are identical as C41 = D16
- the t-stats are identical as D41 = D17
- the *p*-values are identical as E41 = D18
- The sum of squares are identical as C37 = I13 (TSS), C36 = J13 (USS), and C35 = K13 (ESS)
- the mean square errors are identical as D36 = B19
- the standard errors of the regression are identical as B30 = B20
- the R-squares are identical as B28 = B21.

3. In cell F16 enter **Interpretation:** and right justify the cell. In cell G16 write the interpretation of the estimated slope coefficient, and left justify this cell. In cell G16, therefore, write, "As the *x*-variable increases by 1 unit, the *y* value is expected to increase by 9.145 units."

27. Name this worksheet **Tutorial**, and save the entire file as YourName_Lab7.xlsx.

Exercises

1. In the same file as above, select a new worksheet and name it **Exercises**.

2. Center justify columns A through L, and make all of the columns have a width of 10. Set these columns to have Arial 9 point font as well.

3. In cell A1 enter **y**. In cell B1 enter **x**.

4. In cells A2 through A16 enter A2 = 40, A3 = 10, A4 = 30, A5 = 40, A6 = 60, A7 = 125, A8 = 90, A9 = 110, A10 = 50, A11 = 45, A12 = 50, A13 = 10, A14 = 25, A15 = 30, A16 = 35.

5. In cells B2 through B16 enter B2 = 12, B3 = 8, B4 = 8, B5 = 7, B6 = 4, B7 = 8, B8 = 8, B9 = 4, B10 = 10, B11 = 10, B12 = 8, B13 = 15, B14 = 10, B15 = 12, B16 = 11.

6. Enter the sum of each column in row 18 and the average of each column in row 19.

7. In cell C1 enter **y − ybar** and in cell D1 enter **x − xbar**. Fill in these columns appropriately.

8. In cell E1 enter **(y-ybar)^2** and in F1 enter **(x-xbar)^2**. Fill in these columns appropriately.

9. In cell G1 enter **(x-xbar)*(y-ybar)**. As this doesn't fit well, change column G's width to 14. Then fill in this column appropriately.

10. Enter the sum of each column C through G in row 18.

11. In cell A21 enter **b_1** and in cell A22 enter **b_0**. Fill in cells B21 and B22 appropriately.

12. In cell H1 enter **y-hat**, and fill in this column appropriately.

13. In cell I1 enter **(y−ybar)^2**, in cell J1 enter **(y-yhat)^2**, and in cell K1 enter **(yhat−ybar)^2**. Fill in these three columns appropriately, and calculate their sums in row 18. In cell I19 enter **TSS**, in cell J19 enter **USS**, and in cell K19 enter **ESS**.

14. In cell A24 enter **MSE**, in cell A25 enter **SER**, in cell A26 enter **R-sq**. Proceed to enter each of these values in cells B24, B25, and B26.

15. In cell C21 enter **se**. In cell D21 calculate the standard error.

16. In cell c22 enter **t-stat** and in cell c23 enter **p-value**. In cell D22 enter the t-statistic associated with the slope coefficient, and enter the *p*-value in cell D23 (be careful: how many degrees of freedom are there for this problem?)

17. To double-check your answers, have Excel perform the analysis using the Regression feature in the Data Analysis Toolpak. Have Excel produce the results in the same worksheet, starting in cell A29.

18. In cell F21 enter **Interpretation:** and right justify the cell. In cell G21, write an interpretation of the estimated slope coefficient, and left justify this cell.

19. If you haven't already done so, name this worksheet **Exercises** and then save the entire file as YourName_Lab7.xlsx. Use your results from this worksheet to fill in the required information on the answer sheet.

Turning in your work

Email YourName_Lab7.xlsx to your professor as a file attachment to an email with the subject heading Excel Lab 7: Your Name. Turn in your filled-in answer sheet during class.

Answer Sheet for Lab #7: Simple Linear Regression

Name: _____

Use your results from the Exercise portion of the lab to answer the following questions.

1. What is the estimated regression equation: $\hat{y}_i = b_0 + b_1 x_i$?

2. What is the interpretation of b_1?

3. What is the standard error associated with the estimated slope coefficient, b_1?

4. What is the t-stat and p-value associated with the estimated slope coefficient, b_1? Is the x variable a statistically significant predictor of y at the 5% significance level? At the 1% significance level?

5. What is the total sum of squares, the explained sum of squares, the unexplained sum of squares, and the R-squared?

     TSS =

     ESS =

     USS =

     R-sq =