

[Return to Cover Page](#)

LESSON 16 –CORRELATION AND REGRESSION

In this lesson we will learn to find the equation of the regression line and to plot it. We will also find the coefficient of determination, the linear correlation coefficient, and we will learn to predict values of Y for given values of X. We will use Problem 25 page 510 and Problem 15 on page 518 as an example. Notice that these two problems involve the same data, but they ask different questions about that data. Type the data into C1 and C2 in the data window, and label the columns; we will call C1 "Hour" and C2 "Score." Clear the session window below the time/date stamp then type your name, Lesson 16, and Example. Now display the data.

REGRESSION ANALYSIS

To find the equation of the regression line and the coefficient of determination click on Stat > Regression > Regression. Select C2 Score into the "Response:" box and C1 Hour into the "Predictors:" box, then click "OK". The two pieces of information we want, plus a great deal of other information we don't care about will be added to our session window. The equation of the regression line is given at the top, the item R-Sq = 85.1% is what our text calls the coefficient of determination. The figure below has the unwanted information highlighted. It should be deleted. When the excess information has been removed, the session window should look like the first figure on the next page.

Regression Analysis: Score versus Hour						
The regression equation is						
Score = 34.6 + 7.35 Hour						
Predictor	Coef	SE Coef	T	P		
Constant	34.617	4.778	7.24	0.000		
Hour	7.3497	0.9262	7.94	0.000		
S = 7.69806 R-Sq = 85.1% R-Sq(adj) = 83.8%						
Analysis of Variance						
Source	DF	SS	MS	F	P	
Regression	1	3731.4	3731.4	62.97	0.000	
Residual Error	11	651.9	59.3			
Total	12	4383.2				
Unusual Observations						
Obs	Hour	Score	Fit	SE Fit	Residual	St Resid
4	4.00	48.00	64.02	2.21	-16.02	-2.17R
R denotes an observation with a large standardized residual.						

Regression Analysis: Score versus Hour

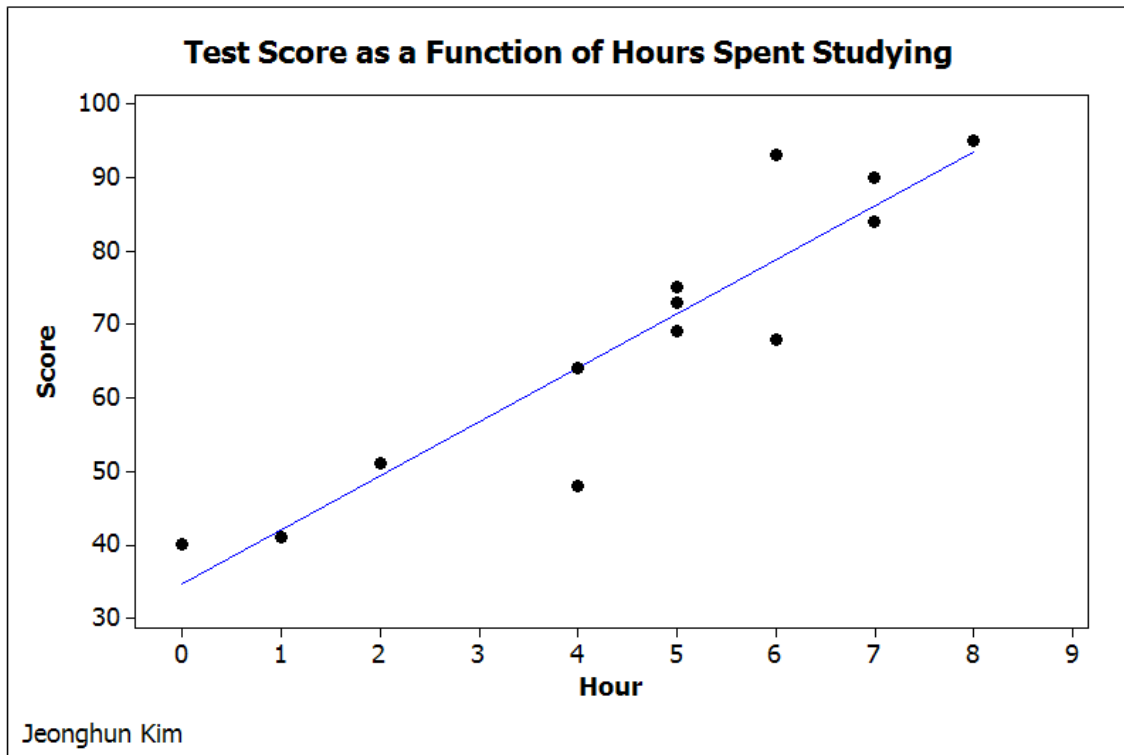
The regression equation is

$$\text{Score} = 34.6 + 7.35 \text{ Hour}$$

R-Sq = 85.1%

GRAPH

To create a scatter plot and graph of the regression line, click on Graph > Scatterplot. Click on "With Regression" then "OK" in the first dialog box. In the second, select Score into the first box in the Y column, and Hour into the first box in the X column. Now click on "Labels" to add a title and your name. Let's type "Test Score as a Function of Hours Spent Studying" for the title. Then click "OK" and "OK". After making everything black and white as in our past graphs, it should look like the figure below.



Notice that this graph violates two of our rules for graphs with two axes, the 3/4 rule and the rule to scale the y-axis from zero. It is not uncommon to violate both of these rules for scatter plots.

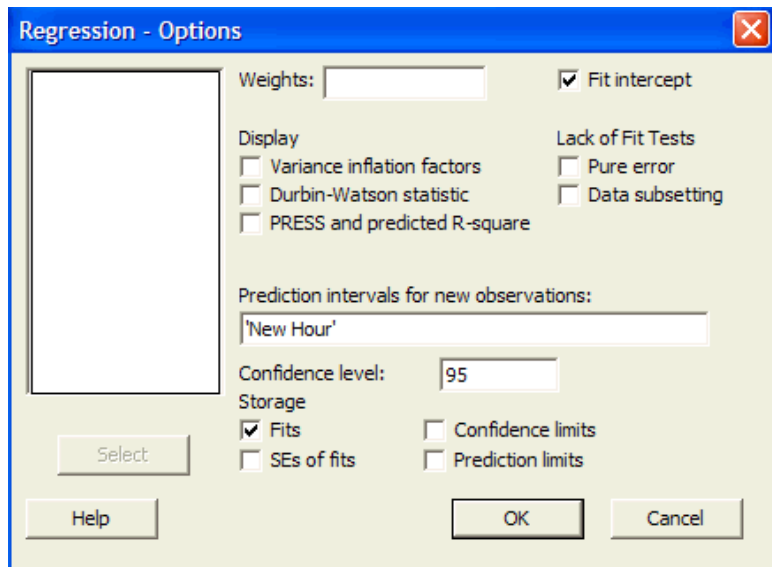
We can answer the questions (a) ~ (d) of Problem 15 on page 518 to estimate Test score using the equation of the regression line above. For instance, for (a), we can estimate the test score as 56.65 when the hour spent studying is 3 since $34.6 + 7.35 * 3 = 56.65$. This completes what is requested for the question but we will add one more requirement that is frequently expected when dealing with this type of problem. "Find the coefficient of determination and interpret its value." NOTE: The coefficient of determination was found along with the equation of the regression line. Write out the answers to the questions asked in the problems on a separate sheet of paper. See the Sample Write-up at the end of the lesson.

CORRELATION

To do Problem 25 on page 510 we need to compute the linear correlation coefficient. Click on Stat > Basic Statistics > Correlation then select Hour and Score into the "Variables:" box and click "OK." The linear correlation coefficient, called the Pearson correlation in Minitab, (0.923 in this case) will be printed in the session window. Notice that the P-Value is also printed, so the problem can be finished by either the classical method or the P-Value method. We will use the classical approach. On a separate sheet of paper write out the 5 step classical method for this hypothesis testing problem. In the "For our sample" step, just write down the value of r provided by Minitab as in the sample write-up.

PREDICTION

We could predict test score by hand as above when the hour spent studying is 3. We can also make Minitab compute it. To do this we must give C3 a reasonable label; then list in it all the values of hours for which we want predictions. Go to the data window and give C3 the title "New Hour" and type 3, 6.5, and 4.5 in the column. (Note that the number 13 is outside of the range of the original data. So it will not be reasonable to predict the test score in this case.) Now click on Stat > Regression > Regression. Select Score into the "Response:" box and Hour into the "Predictors:" box. Now go to "Options" and select "New Hour" into the box called "Prediction intervals for new observations:" then click on the "Fits" box under "Storage" to check it. The Options dialog box should now look like the figure on the right. Now click "OK" to get back to the Regression dialog box and click on "Results" then click on the top button,



"Display nothing." Now click "OK" on the Results dialog box and on the Regression dialog box. Notice that C4 has now been given the name PFIT1 and the values 56.666, 82.390, 130.163, and 67.690 have been added. This is the result we are seeking, but the name is rather strange. Change the label of C4 to "New Score." Return to the session window and do a Display Data for C3 and C4. The complete session window for this example is shown below.

```

_____ 7/20/2007 3:28:19 PM _____

Jeonghun Kim
Lesson 16
Example

Data Display

Row  Hour  Score
  1     0    40
  2     1    41
  3     2    51
  4     4    48
  5     4    64
  6     5    69
  7     5    73
  8     5    75
  9     6    68
 10     6    93
 11     7    84
 12     7    90
 13     8    95

Regression Analysis: Score versus Hour

The regression equation is
Score = 34.6 + 7.35 Hour

R-Sq = 85.1%

Correlations: Hour, Score

Pearson correlation of Hour and Score = 0.923
P-Value = 0.000

Data Display

Row  New Hour  New Score
  1     3.0    56.666
  2     6.5    82.390
  3     4.5    67.690

```

The next page is a complete write-up for this example. Your write-ups for the assignment at the end of this lesson should be similar to this, except that you need not type it.

SAMPLE WRITE-UP

Jeonghun Kim

July 20, 2007

Lesson 16

Example

x = hours spent studying

y = test score

Problem 15 page 518:

1. Line of best fit is given by $y = 34.6 + 7.35x$.
2. Scatter plot with regression line. See attached graph.
3.
 - (a) If $x = 3$, then the test score should be about 56.666.
 - (b) If $x = 6.5$, then the test score should be about 82.390.
 - (c) It is not reasonable to predict the test score in this case since $x = 13$ is outside the range of the original data.
 - (d) If $x = 4.5$, then the test score should be about 67.690.

Problem 25 page 510:

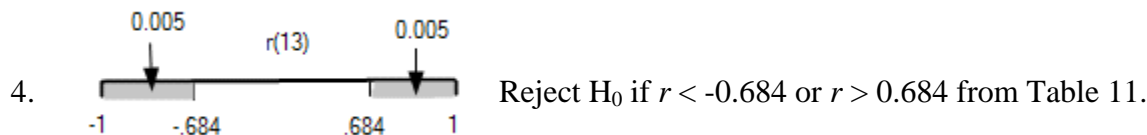
1. $H_0: \rho = 0$

$H_a: \rho \neq 0$

2. $\alpha = 0.01$

3. Assume H_0 is true.
$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}} \sim r(n)$$

For our sample $r = 0.923$



5. Decision: Since $r = 0.923 > 0.684$, reject H_0 .

Conclusion: There is a significant linear correlation between the number of hours spent studying and the test score at the 0.01 level of significance.

MINITAB ASSIGNMENT 16

See instructions on page 8.

For each of the problems below enter and display the data then let Minitab do all the computations. On a separate sheet of paper write out the answers to the questions posed by the problem in the text and those added to the problem below. NOTE: These may be different from the questions posed in the sample problem in the lesson above. In each case have Minitab produce a scatter plot with the regression line included as we did in the example above. When a prediction is required, be sure to change the name of the predicted value(s) to something more reasonable than the default name given by Minitab.

1. Do Problem 14 on page 544 with the following parts added.
 - (a) Test the claim, at the 5% level of significance, that there is a linear correlation between the engine displacement and the fuel efficiency.
 - (b) Does this prove that engine displacement affects fuel efficiency? Explain.

2. Do Problem 16 on page 518 with the following part added.
 - (e) Find and interpret the coefficient of determination.NOTE: No hypothesis testing is required for this problem.

[Return to Cover Page](#)