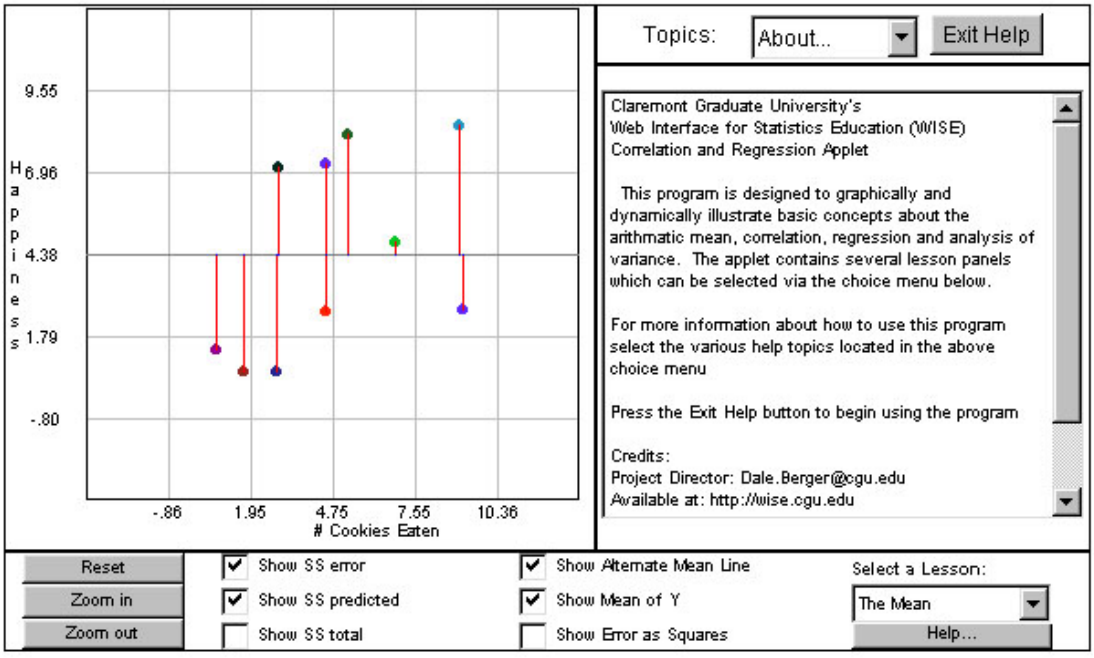
WISE Regression/Correlation Interactive Lab

Introduction to the WISE Correlation/Regression Applet

This tutorial focuses on the logic of regression analysis with special attention given to variance components. The tutorial provides a brief review and four interactive exercises. At the end of the tutorial there are some thought questions. If you have a hard copy of this handout, you do not need to print anything from the tutorial; everything that you need for recording results is in this handout.

Terminology:
Y variable: this is the 'Dependent' Variable, the variable you wish to predict.
X variable: this is the 'Independent' Variable, the variable used to predict of Y.
SS Regression = SS Predicted = SS Explained
SS Error = SS Residual = SS Unexplained



To begin the tutorial go to ***http://wise.cgu.edu*** **tutorials, choose '**correlation and regression.' For this assignment, you will complete Modules 1 through 3.

Using the WISE Correlation/Regression Tutorial

The applet screen will open with a description of the applet in the large box on the right. After reading this information, click the box marked 'Exit.' This box should now contain four columns with numbers. The first two columns are your X, Y pairs.

Change the box in the lower right hand corner from 'The Mean' to 'Regression.'

Near the bottom of the applet, there are several boxes. Some are checked (Show SS error, Show Regression Line, Show SS predicted, and Show Mean of Y) and some are not checked (Show SS total, Show Deviations as Squares). Modify these so that none of the boxes are checked.

For some exercises you will drag the data points to create relationships. To drag data points, click on the data point with your mouse and drag to another location. The correlation will appear in box on the right.

The remaining pages of the handout are the same as pages found in the tutorial. Use these to record your work. You will answer several on-screen multiple choice questions before you get to the sections that are on the pages that follow.

You may want to refer back to this page when you get to the part of the tutorial that uses the applet.

Module #1, Exercise #1

Introduction to the Regression Applet: Calculations and observations

In this exercise, you will learn how to use the WISE regression applet to deepen your understanding of regression. Using the very small set of data shown below, we will step through relevant regression values and see how they are calculated and how they are represented graphically. Answers are provided for all problems in Module 1 at the end of this handout.

Set up the applet: From the '*Select a Lesson*:' menu in the lower right hand corner of the applet, choose 'Regression.' Remove the checks from all boxes except for the box: *Show Regression Line*.

Case	X	Y
1	1	2
2	3	5
3	5	7
4	7	6

a. **Correlation, Slope, and Y-intercept.** The applet provides these statistics, which are important for regression analysis. Find these terms in the applet and enter each below.

correlation (r) = _____

slope (b) = _____

intercept (a) = _____

b. **The regression equation**. The regression equation is a formula for the straight line that best fits the data. Later we will learn exactly how 'best fit' is defined. The regression equation can be used to predict the Y score (called Y′, or Y-prime) for each of our X values. The general form of the regression equation is $Y' = a + bX$.

In our example, a=2.2 and b=.700, so the regression equation is $Y' = 2.2 + .700X$. Our first X score is 1, which generates a predicted Y score of 2.9, from 2.2 + .7(1). Calculate the three remaining Y′ values by hand and enter them into the table below. If you get stuck, you may check your answers by clicking the links to answers in the on-line tutorial but make sure that you can do the calculations and interpretations yourself.

Case	X	Y	Y´
1	1	2	2.9
2	3	5	
3	5	7	
4	7	6	

SS Total (Total Variance)

SS total is the sum of squared deviations of observed Y scores from the mean of Y. This is an indication of the error we expect if we predict every Y score to be at the mean of Y. (If X is not available or if X is not useful, then the mean of Y is our best prediction of Y scores.)

c. To calculate SS Total, take each value of Y, subtract the mean, and square the result, then sum all of the values in the column. A general formula for SS Total is $\Sigma(Y - \overline{Y})^2$. For these data the mean of Y is 5. For the first case, the squared deviation from the mean is 9. Calculate the values for the last three cases, and sum the values for all four cases in the last column to get SS Total.

Case	X	Y	$(\mathbf{Y} - \overline{\mathbf{Y}})$	$(\mathbf{Y} - \overline{\mathbf{Y}})^2$
1	1	2	(2-5) = -3	$(-3)^2 = 9$
2	3	5		
3	5	7		
4	7	6		
Sum	16	20		

d. *Now in the applet, place a check mark in the boxes titled Show SS Total and Show Mean of Y and remove all other checks.* The [note that there are only three, because one is zero] vertical black lines represent the deviations of each case from the mean of Y. Verify the correspondence of the length of these lines with the values in the table for the column $(\mathbf{Y} - \overline{\mathbf{Y}})$. Which case has the largest deviation from the mean?

The largest deviation from the mean is _____ for Case ___.

Hint: Look at the graph in the applet and at your calculations in the table.

e. *Now check the box labeled Show Error as Squares*. The sizes of the black squares correspond to the squared deviations from the mean, and the sum of the areas of these squares corresponds to SS Total. Notice how the deviations from the mean for the first and fourth cases are -3 and +1, while the squared deviations are 9 and 1. This shows how

points farther from the mean contribute much more to SS Total than points closer to the mean. What is the contribution of the second case to SS Total? Why?

The contribution to SS Total for Case 2 is _____ because

f. Now calculate the sum of the squared deviations from the mean $\Sigma(Y - \overline{Y})^2$. You can do this by adding the values in the column headed $(\mathbf{Y} - \overline{\mathbf{Y}})^2$.

$\Sigma(Y - \overline{Y})^2$ = SS Total = _____. In the applet, SS for Total = _____.

g. Explain what SS Total means. How would the plot differ if SS Total was much smaller, say 2.00? What if SS Total was much larger, say 100?

SS Error

SS Error is the sum of squared deviations of observed Y scores from the predicted Y scores when we use information on X to predict Y scores with a regression equation. SS Error is the part of SS Total that CANNOT be explained by the regression.

h. Calculations. Complete the calculations below using the predicted scores (**Y´**) calculated in question 1b. The sample mean is 5.0 for every case.

Case	**X**	**Y**	**Y´**	**(Y - Y´)**	**$(Y - Y')^2$**
1	1	2	2.9	2-2.9 = -0.9	$(-0.9)^2 = 0.81$
2	3	5			
3	5	7			
4	7	6			
Sum	16	20			

i. Now *place check marks in the boxes titled Show Regression Line* and *Show SS error,* and remove checks from all other boxes. Deviations of the observed points from their predicted values on the regression line are shown in red.

The largest deviation is for Case ____, and the size of the deviation is _____.

The smallest deviation is for Case ____, and the size of the deviation is _____.

j. Now *check the box titled Show Errors as Squares*. The sizes of the red squares correspond to the squared deviations. In the table for part h, compare the squared deviations shown in the last column for Cases 2 and 3. Observe how the red boxes for Cases 2 and 3 correspond to these values. The sum of the squared deviations is the sum of the last column in the table.

Record your calculated value here _____. This is the Sum of Squares Error.

In the applet under Analysis of Variance find the value for SS Error _____

k. Explain in simple English what SS Error means. What would the plot look like if SS Error was very small compared to SS Total? What would the plot look like if SS Error is about as large as SS Total?

SS Predicted

SS Predicted is the part of SS Total that CAN be predicted from the regression. This corresponds to the sum of squared deviations of predicted values of Y from the mean of Y.

L. Calculations. Complete the calculations below using the predicted scores (**Y´**) calculated for each case in part 1b and the mean of Y (5).

Case	**X**	**Y**	**Y´**	$(\mathbf{Y}' - \overline{\mathbf{Y}})$	$(\mathbf{Y}' - \overline{\mathbf{Y}})^2$
1	1	2	2.9	$2.9 - 5.0 = -2.1$	$(-2.1)^2 = 4.41$
2	3	5			
3	5	7			
4	7	6			
Sum	16	20	20.0		

m. Now *click the boxes marked Show Mean of Y and Show Regression Line* and remove the checks from all other boxes. Check *Show SS Predicted* to see deviations of regression line from the mean, shown in blue. The blue lines represent the differences between the mean and predicted scores. If X were not useful in predicting Y, then the best prediction of Y would be simply the mean of Y for any value of X, and the blue lines would be zero

in length. If X is useful in predicting Y, then the predicted values differ from the mean. The blue lines give an indication of how well X predicts Y.

Click the box marked Show Error as Squares, to see the squared deviations of predicted scores from means. Compare these to the red squares for SS Error. (You can click Show SS Error if you would like to be reminded of the size of the red squares.) Is X useful for predicting Y in this plot? How do you know?

n. The sum of the squared deviations of the predicted scores from the mean is the sum of the last column in the table in part L.

Record the calculated value here _____. This is the Sum of Squares Predicted.

In the applet under Analysis of Variance find the value for SS Predicted _____

o. Explain what SS Predicted means. What would the plot look like if SS Predicted was very small relative to SS Total?

r-squared as proportion of variance explained

p. Note that SS Total = SS Predicted + SS Error. (14.000 = 9.800 + 4.200). Thus, with the regression model, we split SS Total into two parts, SS Predicted and SS Error. We can compute the proportion of SS Total that is in SS Predicted. In terms of sums of squares, this is the ratio of SS Predicted to SS Total.

Calculate [SS Predicted/ SS Total] = _____ / _____ = _____.

SS Total is the numerator of the variance of Y (i.e., $\Sigma(Y - \overline{Y})^2$), so the calculated ratio can be interpreted as the proportion of variance in Y that can be predicted from X using the regression model. A useful fact in regression is that this ratio is equal to the correlation squared (r-squared). Thus, the correlation squared (r-squared) represents the proportion of variance in Y that can be explained by X, using the regression model.

What does the applet report for the correlation r and r-squared?

r = _____; r squared = _____

Summarize the relationship between X and Y for this set of data in simple English.

Go to Module #1, Exercise #2. Be sure to following the link at the bottom of the applet screen.

Regression Module #1, Exercise #2

Sums of Squares: Computation and interpretation

You will learn about the meaning of Sums of Squares and how they can be represented and understood graphically. You will be asked to make some calculations using a very small data set. These data, along with many calculated values, are represented in the applet. Proceed to the applet for Module 1, Exercise 2 (follow the link on the web page). Again, answers for this exercise are found at the end of this handout.

Case	**X**	**Y**	**Y´**	$(\mathbf{Y} - \overline{\mathbf{Y}})$	$(\mathbf{Y} - \overline{\mathbf{Y}})^2$	**(Y - Y´)**	**(Y - Y´)²**	$(\mathbf{Y'} - \overline{\mathbf{Y}})$	$(\mathbf{Y'} - \overline{\mathbf{Y}})^2$
1	1	4	3.40	.75	.5625	.60	.3600	.15	.0225
2	3	3							
3	5	2							
4	7	4							
Sum	**16**	**13**	**13.00**	**0.00**	**2.7500**	**0.00**	**2.7000**	**0.00**	**.0500**

a. *Look at the plot of the data*. Does it appear that the regression model will explain a large portion of the variance in Y?

b. *Calculate values for all of the empty cells in the table*. Values for Case 1 and for the Sum are shown so that you can check your work. Some useful information can be found in the applet.

Hints: You can calculate Y′ with the formula shown in the applet: Y′ = -.050X + 3.450.
Calculate the mean of Y by dividing the Sum of Y by n. (You should get 3.25.)

c. *Now place check marks only in the boxes titled Show SS Total and Show Mean of Y*. The four vertical black lines represent the deviations of each case from the mean of Y. Check the correspondence of the length of these lines with the values in the table that you calculated for the column $(\mathbf{Y} - \overline{\mathbf{Y}})$. Which case has the largest deviation from the mean?

The largest deviation from the mean is _____ for Case ___.

Hint: Look at the graph in the applet and at your calculations in the table.

d. *Now check the box labeled Show Error as Squares*. The black squares correspond to the squared deviations from the mean, and the sum of the areas of these squares

corresponds to SS Total. Notice how the deviations from the mean for the first and second cases are .75 and -.25 (a ratio of 3:1), while the squared deviations are .5625 and .0625 (a ratio of 9:1). This shows how points farther from the mean contribute much more to SS total than points closer to the mean. What is the contribution of the third case to SS Total? What is the ratio of this contribution compared to the contribution of the second case?

The ratio of the contribution to SS total for Case 3 vs Case 2 is _____:_____.

e. *Now remove checks from all of the boxes, and check the boxes labeled Show Regression Line and Show SS error*. The regression line allows us to find the predicted value of Y for any value of X. The vertical red lines correspond to the deviations of the observed values for Y from the predicted values on the regression line (Y′). Which case has the largest deviation from its predicted value of Y (Y′)?

The largest deviation from the predicted value is _____ for Case ___.

Hint: Look at the graph and your table.

f. *Now check the box labeled Show Error as Squares*. The red squares correspond to the squared deviations of Y from the predicted values (Y′). The sum of these areas corresponds to SS Error. You can compare SS Error with SS Total by also checking the box labeled Show SS Total. How does SS Error compare to SS Total? Do you think SS Error is much smaller than SS Total? Now check your table. What are the values for SS Error and SS Total?

SS Error = _____ SS Total = _____

g. *Now remove checks from the boxes Show SS Error, Show SS Total, and Show Error as Squares, and check the boxes labeled Show Mean of Y and Show SS Predicted.* (Show Regression Line is still checked.) The blue vertical lines show the deviations between the predicted value of Y (Y′) and the mean of Y for each case.

Now check Show Error as Squares. The sum of the areas of these squares corresponds to SS Predicted. If the regression line is near to the mean, that tells us that the regression model does not predict Y much better than the mean does. Do you think SS Predicted is much smaller than SS Total? Now check your table. What are the values for SS Predicted and SS Total?

SS Predicted = _____ SS Total = _____

h. Verify that SS Total = SS Predicted + SS Error.

_____	=	_____	+	_____
SS Total	=	SS Predicted	+	SS Error

i. We can calculate r squared from our SS values.

What portion of SS Total is accounted for by SS Predicted?

[SS Predicted / SS Total] = [_____ / _____] = _____

What is the value for r squared as reported in the applet? _____

Thus, we can say "r squared is the proportion of variance in Y that is explained by X." The total variability in Y is measured by the sum of the squared deviation of Y scores around the mean of Y, which is SS Total. When we use linear regression, information on X is used to generate the regression line and the predicted value of Y. The part of SS Total that is 'explained' by the model is SS Predicted. The proportion explained is SS Predicted divided by SS Total.

Answers for Module 1, Exercise 1 and 2:

Module 1, Exercise 1a:

The correlation (r) is .837, the slope of the line (b) is .700, and the Intercept (a) is 2.200, taken from the right panel in the applet for Regression Module 1, Exercise 1.

Module 1, Exercise 1b:

4.3, 5.7, and 7.1. Calculation for the last value is 2.2 + .7x7 = 2.2 + 4.9 = 7.1.

Module 1, Exercise 1c:

SS Total = 14. The squared deviations from the mean for the four cases are 9, 0, 4, and 1, respectively.

Module 1, Exercise 1d:

The largest deviation from the mean is -3, for Case 1.

Module 1, Exercise 1e:

The contribution of the second case to SS Total is zero, because the Y value of 5 is exactly equal to the mean.

Module 1, Exercise 1f:

SS Total = 14. You can find this as the sum of the last column in your table, and this value is also shown in the applet in the SS column in the Analysis of Variance section.

Module 1, Exercise 1g:

SS Total is the sum of the squared deviations of Y scores from the mean of Y. If SS Total was much smaller, then all of the Y values must be close to the mean. SS Total could be much larger for several reasons: many of the Y values could be somewhat farther from the mean, a few values, or even one value, could be very far from the mean, or we could simply have many more Y values. Note that a single Y value that differed from the mean by 10 points would contribute 100 to SS Total.

There is a close relationship between SS Total and variance. An estimate of the population variance taken from a sample is calculated at the sum of the squared

deviations from the mean divided by the degrees of freedom, which is (SS Total) / (n-1) for a single sample. In our example, this is 14/3 = 4.667. The standard deviation is the square root of variance = 2.16, the value shown in the applet as the std dev for the DV.

Module 1, Exercise 1h:

For the second case, Y' = 4.3, Y-Y' = (5 – 4.3) = .7, and $(Y - Y')^2$ = .49. For the third case, Y' = 5.7, (Y-Y')=(7 – 5.7) = 1.3, and $(Y - Y')^2$ = 1.69.

Module 1, Exercise 1i:

The largest deviation is for Case __3__, and the size of the deviation is _1.3_.

The smallest deviation is for Case __2__, and the size of the deviation is __.7__.

Module 1, Exercise 1j:

The calculated value for the Sum of Squares Error = SS Error = 4.200.

Module 1, Exercise 1k:

SS Error is the sum of the squared deviations of observed scores from the predicted scores. If SS Error is very small, every observed score is close to the predicted score, so the plot of every observed score is close to the regression line.

If SS Error is much smaller than SS Total, then the sum of deviations around the regression line is much smaller than the sum of deviations around the mean. Thus, the regression equation gives much more accurate predictions of scores than simply using the mean as the prediction for all scores. The plot would show a strong linear relationship between X and Y.

If SS Error is about the same size as SS Total, then the regression equation has not improved our prediction of Y scores. The regression line would be close to horizontal at the mean. The plot would not show any indication of a linear relationship between X and Y.

Module 1, Exercise 1L:

For the second case, the predicted score is 4.3, which is .7 below the mean of 5.0, so the squared deviation of the predicted score from the mean is .49. For the third case, the

deviation is +.7, and for the fourth case the deviation is +2.1. The sum of the squared deviations is 9.80.

Module 1, Exercise 1m:

Yes, it appears that X is useful in predicting Y in our plot. The blue lines, which indicate predictive ability, are substantial. They are relatively long, compared to the red lines we observed for error deviations, and the blue squares are relatively large compared to the red squares. Thus, it appears that the SS Predicted is substantial.

Module 1, Exercise 1n:

The Sum of Squares Predicted from the Analysis of Variance table in the applet is 9.800, which is also the sum of the last column in the table in part L.

Module 1, Exercise 1o:

SS Predicted is the sum of the squared deviations of predicted scores from the mean. If the regression model is not at all useful, then the predicted score will be the mean for each case, and SS Predicted will be zero. If the regression model is only slightly helpful, then the predicted scores will be only slightly different from the mean, and SS Predicted will be small relative to SS Total. This plot would show virtually no linear relationship between X and Y, and the regression line would be close to the horizontal line for the mean of Y.

If there is a strong linear relationship in the data, SS Predicted is large relative to SS Error, and the observed data fall close to the regression line.

Module 1, Exercise 1p:

[SS Predicted / SS Total] = 9.800 / 14.000 = .700.

The applet reports r = .837 and r squared = .700.

This sample data shows a strong linear relationship, as measured by r=.837. The plot shows this strong positive relationship, with larger values of X generally associated with larger values of Y. In this sample, 70% of the variance in Y can be explained by the linear relationship with X.

We should note that this is an extremely small sample, and that we would not be able to generalize to the relationship in a population of X and Y values, even if there four cases are a random sample from that population.

Answers for Module 1, Exercise 2:

Module 1, Exercise 2a:
Look at the plot of the data. Does it appear that the regression model will explain a large portion of the variance in Y?

The plot does not show a clear relationship between values of X and Y. Thus, we do not expect that a linear regression model will account for much of the variance in Y.

Module 1, Exercise 2b:
Add your calculated values for each column to check against the sum that is shown in the table. If you get a different answer, check your calculations for the first row to make sure that you used the right values. Then use the same procedure for each of the remaining rows. Here are answers for the second case.

Case	**X**	**Y**	**Y´**	$y-\bar{y}$	$(y-\bar{y})^2$	$y-y'$	$(y-y')^2$	$y'-\bar{y}$	$(y'-\bar{y})^2$
2	3	3	3.30	-0.25	0.0625	-0.30	0.09	0.05	0.0025

Module 1, Exercise 2c:
The largest deviation from the mean is -1.25 for Case 3.

Module 1, Exercise 2d:
The ratio of the contribution to the SS total for Case 3 compared to Case 2 is 1.5625:.0625, which simplifies to 25:1. Notice that the deviation from the mean is five times greater for Case 3 compared to Case 2 (-1.25 vs. -.25), so the squared contribution is 25 times greater.

Module 1, Exercise 2e:
The largest deviation from the mean is -1.20 for Case 3.

Module 1, Exercise 2f:
The red boxes for SS Error are very similar to the black boxes for SS Total, so it appears that SS Error is nearly as large as SS Total. The computed values confirm this conclusion. SS Error = 2.70, while SS Total = 2.75.

Module 1, Exercise 2g:
The red boxes for SS Predicted are much smaller than the black boxes for SS Total, so it appears that SS Predicted is only a small fraction of SS Total. The computed values confirm this conclusion. SS Predicted = .05, while SS Total = 2.75.

Module 1, Exercise 2h:

SS Total = 2.75
SS Predicted = .05
SS Error = 2.70

SS Total = SS Predicted + SS Error
2.75 = .05 + 2.70

Module 1, Exercise 2i:

[SS Predicted / SS Total] = [.05_ / 2.75] = .018

Go to Module #2. Be sure to following the link at the bottom of the applet screen.

Module #2, Interactive Exercise #1

For this section you will create relationships. Remember that the statistical tests provided by regression analyses are valid only when relationships are linear. Be sure that all relationships you create follow a straight-line pattern. For this section, we do not provide answers.

For this problem, distribute the data points so that you have a correlation of about +.90. (Place the curson over a point, hold the left mouse button, and slide the point to a desired location.) After getting the correct correlation (or close to the correct value) *place a check mark in the box titled Show mean of Y*. (Note: Y is the dependent variable).

a. What does the plot of scores look like? Answer in terms of general pattern of the scores (Are the points all close to a line? Is there a positive or a negative relationship? Is the relationship strong or weak?).

b. Do the points deviate a lot from the mean on Y? (Check 'Show SS total' to see the deviations, and check 'Show Error as Squares' to see how much the deviation of each data point from the mean contributes to SS Total.)

c. Record the numeric value for SS total here _____ (The SS values are shown in the applet.)

d. Now place a check mark in the box titled Show Regression Line. Do your points seem to deviate a lot from the regression line? (Remove the check from the 'Show SS total' box and check 'Show SS error' to see deviations of observed points from the regression line, shown in red.).

e. Explain what SS error means. What would the plot look like if SS error was even smaller?

f. Record the numeric value for SS error here _____ .

g. Now *click the box marked Show SS predicted and remove the check from the box Show SS error*. The blue lines that appear represent the difference between the mean and predicted scores. How are these scores distributed? Do the predicted scores deviate a lot from the mean?

h. Record the correct numeric value for SS Predicted here _____ .

i. If you were predicting y (y – prime) from X = 10? Would you expect this to be an accurate prediction? Why or why not?

j. Calculate r squared from your SS values.

r squared = [SS predicted/ SS total] = _____/ _____ = _____.

Check: r = _____; r squared = _____

Follow the link at the bottom of the applet that to Go to Module #2, Interactive Exercise #2

Module #2, Interactive Exercise #2

Distribute the data points so that you have a correlation of about +.30.
Move the points around until you have a correlation of about .30. Note: this example works better if you move several of the points a little rather than just moving one to an extreme value on the distribution. After getting the correct correlation (or close to the correct value) *place a check mark in the box titled Show mean of Y*. (Note: Y is the dependent variable).

a. What does the distribution of scores look like now, how does this compare to the data in the previous problem (r = .90)?

b. Do the points deviate a lot from the mean on Y? (Check 'Show SS total' to see the deviations, and check 'Show Error as Squares' to see how much the deviation of each data point from the mean contributes to SS Total.)

c. Record the numeric value for SS total here _____ (The SS values are shown in the applet.)

d. Now *place a check mark in the box titled Show Regression Line*. Do your points seem to deviate a lot from the regression line? (Remove the check from the 'Show SS total' box and check 'Show SS error' to see deviations of observed points from the regression line, shown in red.).

e. Record the SS error here _____

f. Compare the error results to the results from the previous problem (r = .90). How do these compare? Compare proportion of error variance for each by taking the SS_{error} / SS_{total}.

g. Now *click the box marked Show SS predicted and remove the check from the box Show SS error*. The blue lines that appear represent the difference between the mean and predicted scores. How are these scores distributed? Do the predicted scores deviate a lot from the mean?

h. Record the correct numeric value for SS Predicted here _____ .

i. How does the difference between predicted scores and the mean differ from the previous problem (r = .90)? Focus on the proportion of predicted out of total ($SS_{Predicted}$ / SS_{total}) rather than the SS value itself.

j. Why is the distribution of predicted vs. mean scores so different between the situations where r = .90 versus r = .30?

Follow the link at the bottom of the applet that to Go to Module #2, Interactive Exercise #3

Module #2, Interactive Exercise #3

Now, move the data points so that you have a correlation of about .00. After getting the correct correlation (or close to the correct value) *place a check mark in the box titled Show mean of Y*. (Note: Y is the dependent variable).

a. What does the distribution of scores look like? Compare to r = .90 and r = .30.

b. Record the numeric value for SS total here _____ (The SS values are shown in the applet.)

c. Now *place a check mark in the box titled Show Regression Line*. Do your points seem to deviate a lot from the regression line? (Remove the check from the 'Show SS total' box and check 'Show SS error' to see deviations of observed points from the regression line, shown in red.).

d. Record the SS error here _____

e. Compare the error results to the results from the previous problems (r = .30 and r = .90). How do these compare? Compare proportion of error variance for each by taking the SS_{error} / SS_{total}.

f. Record SS predicted here _____

g. Compare the predicted variance results to the results from the first problem (r = .90). How do these compare? Compare proportion of predicted variance for each by taking the $SS_{predicted}$ / SS_{total}.

h. If we don't know anything about X for a new case, our best prediction of Y is the mean of Y. Does our regression line provide better predictions of Y?

i. Why, when the correlation is .00, does the predicted score match the mean?

j. Does knowing about the relationship between X and Y help us to predict Y in this case?

Module 3. The Impact of Outliers

This problem is a little different than the others. For this problem, first distribute the data points so that you have a correlation of about +.90.

Imagine that you added an extreme point to the upper left-hand corner (don't actually add the point yet). How do you think this would change your correlation?

Would you expect the correlation to become larger or smaller? _____

If you had to guess, what do you think your correlation value would be after the addition of this single point? _____

Now add a point to the upper left-hand corner (to add a point, double click the spot where you want to add the point). This new point will represent a single point that deviates from your strong pattern of correlations.

What is your new correlation? _____

Does this differ from your expectation? How? Comment on the impact of the addition of an extreme score. (It may be interesting to look at the squares for errors – Check "Show SS error" and "Show Error as squares.")

Imagine that you had a set of scores where one value was incorrectly entered (e.g., X = 100 entered when x was actually 10). How might this type of error impact your results?

Regression Tutorial Follow-Up Questions:

1. A misguided friend says, "I took statistics. My book said that for a normal distribution of scores, the mean is the same value as the mode. Since the mode is the most common score, wouldn't our 'best guess' of the value of our DV be the mean (or the mode)? Why can't we just use the mean as our predicted value for y? See, you don't need any of this silly regression stuff." How would you respond to this?

2. For a class project, your statistics professor provides you with data from his previous classes on final grade percentage and number of classes sessions missed. You calculate a correlation between the two variables and find r = -.85. Those people who miss class tend to do poorly in the class.

Next, your misguided friend from problem #1 presents some data. His analysis is on class grades and the number of psychology courses that the student has previously taken. He finds a correlation of r = +.10. Not only does your friend conclude that it is a good idea to take a lot of psychology classes before you take statistics but that attendance is much less important than prior experience with psychology. His argument goes like this: "My correlation is +.10, that is a lot bigger than -.85. The smallest correlation is –1.0 and your correlation is pretty close to that!"

Again, enlighten your friend as to the error(s) in his thinking.

3. You performed a statistical analysis and found with 25 pairs of scores the correlation between two variables was +.25. In examining a scatterplot of this relationship you find that most of the pairs of scores exhibit a strong positive relationship. One score however follows a different pattern than the others. You verify the accuracy of the score and find that it is not the result of a data entry error. You delete the pair and the correlation between the variables rises to +.82.

a. How would you interpret the first correlation value (.25)?

b. How would you interpret the second correlation value (.82)?

c. Comment on the deletion of the single outlying score – is it appropriate to delete this score? Why or why not?

d. Suggest a strategy for handling outlying scores such as this.

e. Suggest a strategy for reporting the results of the tests above (i.e., What value should you report?).

Answers for Module 1, Exercise 1 and 2:

Module 1, Exercise 1a:

The correlation (r) is .837, the slope of the line (b) is .700, and the Intercept (a) is 2.200, taken from the right panel in the applet for Regression Module 1, Exercise 1.

Module 1, Exercise 1b:

4.3, 5.7, and 7.1. Calculation for the last value is 2.2 + .7x7 = 2.2 + 4.9 = 7.1.

Module 1, Exercise 1c:

SS Total = 14. The squared deviations from the mean for the four cases are 9, 0, 4, and 1, respectively.

Module 1, Exercise 1d:

The largest deviation from the mean is -3, for Case 1.

Module 1, Exercise 1e:

The contribution of the second case to SS Total is zero, because the Y value of 5 is exactly equal to the mean.

Module 1, Exercise 1f:

SS Total = 14. You can find this as the sum of the last column in your table, and this value is also shown in the applet in the SS column in the Analysis of Variance section.

Module 1, Exercise 1g:

SS Total is the sum of the squared deviations of Y scores from the mean of Y. If SS Total was much smaller, then all of the Y values must be close to the mean. SS Total could be much larger for several reasons: many of the Y values could be somewhat farther from the mean, a few values, or even one value, could be very far from the mean, or we could simply have many more Y values. Note that a single Y value that differed from the mean by 10 points would contribute 100 to SS Total.

There is a close relationship between SS Total and variance. An estimate of the population variance taken from a sample is calculated at the sum of the squared

deviations from the mean divided by the degrees of freedom, which is (SS Total) / (n-1) for a single sample. In our example, this is 14/3 = 4.667. The standard deviation is the square root of variance = 2.16, the value shown in the applet as the std dev for the DV.

Module 1, Exercise 1h:

For the second case, Y' = 4.3, Y-Y' = (5 – 4.3) = .7, and (Y – Y')2 = .49. For the third case, Y' = 5.7, (Y-Y')=(7 – 5.7) = 1.3, and (Y – Y')2 = 1.69.

Module 1, Exercise 1i:

The largest deviation is for Case __3__, and the size of the deviation is _1.3_.

The smallest deviation is for Case __2__, and the size of the deviation is __.7__.

Module 1, Exercise 1j:

The calculated value for the Sum of Squares Error = SS Error = 4.200.

Module 1, Exercise 1k:

SS Error is the sum of the squared deviations of observed scores from the predicted scores. If SS Error is very small, every observed score is close to the predicted score, so the plot of every observed score is close to the regression line.

If SS Error is much smaller than SS Total, then the sum of deviations around the regression line is much smaller than the sum of deviations around the mean. Thus, the regression equation gives much more accurate predictions of scores than simply using the mean as the prediction for all scores. The plot would show a strong linear relationship between X and Y.

If SS Error is about the same size as SS Total, then the regression equation has not improved our prediction of Y scores. The regression line would be close to horizontal at the mean. The plot would not show any indication of a linear relationship between X and Y.

Module 1, Exercise 1L:

For the second case, the predicted score is 4.3, which is .7 below the mean of 5.0, so the squared deviation of the predicted score from the mean is .49. For the third case, the

deviation is +.7, and for the fourth case the deviation is +2.1. The sum of the squared deviations is 9.80.

Module 1, Exercise 1m:

Yes, it appears that X is useful in predicting Y in our plot. The blue lines, which indicate predictive ability, are substantial. They are relatively long, compared to the red lines we observed for error deviations, and the blue squares are relatively large compared to the red squares. Thus, it appears that the SS Predicted is substantial.

Module 1, Exercise 1n:

The Sum of Squares Predicted from the Analysis of Variance table in the applet is 9.800, which is also the sum of the last column in the table in part L.

Module 1, Exercise 1o:

SS Predicted is the sum of the squared deviations of predicted scores from the mean. If the regression model is not at all useful, then the predicted score will be the mean for each case, and SS Predicted will be zero. If the regression model is only slightly helpful, then the predicted scores will be only slightly different from the mean, and SS Predicted will be small relative to SS Total. This plot would show virtually no linear relationship between X and Y, and the regression line would be close to the horizontal line for the mean of Y.

If there is a strong linear relationship in the data, SS Predicted is large relative to SS Error, and the observed data fall close to the regression line.

Module 1, Exercise 1p:

[SS Predicted / SS Total] = 9.800 / 14.000 = .700.

The applet reports r = .837 and r squared = .700.

This sample data shows a strong linear relationship, as measured by r=.837. The plot shows this strong positive relationship, with larger values of X generally associated with larger values of Y. In this sample, 70% of the variance in Y can be explained by the linear relationship with X.

We should note that this is an extremely small sample, and that we would not be able to generalize to the relationship in a population of X and Y values, even if there four cases are a random sample from that population.

Answers for Module 1, Exercise 2:

Module 1, Exercise 2a:
Look at the plot of the data. Does it appear that the regression model will explain a large portion of the variance in Y?

The plot does not show a clear relationship between values of X and Y. Thus, we do not expect that a linear regression model will account for much of the variance in Y.

Module 1, Exercise 2b:
Add your calculated values for each column to check against the sum that is shown in the table. If you get a different answer, check your calculations for the first row to make sure that you used the right values. Then use the same procedure for each of the remaining rows. Here are answers for the second case.

Case	**X**	**Y**	**Y´**	$y-\bar{y}$	$(y-\bar{y})^2$	$y-y'$	$(y-y')^2$	$y'-\bar{y}$	$(y'-\bar{y})^2$
2	3	3	3.30	-0.25	0.0625	-0.30	0.09	0.05	0.0025

Module 1, Exercise 2c:
The largest deviation from the mean is -1.25 for Case 3.

Module 1, Exercise 2d:
The ratio of the contribution to the SS total for Case 3 compared to Case 2 is 1.5625:.0625, which simplifies to 25:1. Notice that the deviation from the mean is five times greater for Case 3 compared to Case 2 (-1.25 vs. -.25), so the squared contribution is 25 times greater.

Module 1, Exercise 2e:
The largest deviation from the mean is -1.20 for Case 3.

Module 1, Exercise 2f:
The red boxes for SS Error are very similar to the black boxes for SS Total, so it appears that SS Error is nearly as large as SS Total. The computed values confirm this conclusion. SS Error = 2.70, while SS Total = 2.75.

Module 1, Exercise 2g:
The red boxes for SS Predicted are much smaller than the black boxes for SS Total, so it appears that SS Predicted is only a small fraction of SS Total. The computed values confirm this conclusion. SS Predicted = .05, while SS Total = 2.75.

Module 1, Exercise 2h:

SS Total = 2.75
SS Predicted = .05
SS Error = 2.70

SS Total = SS Predicted + SS Error
2.75 = .05 + 2.70

Module 1, Exercise 2i:

[SS Predicted / SS Total] = [.05_ / 2.75] = .018