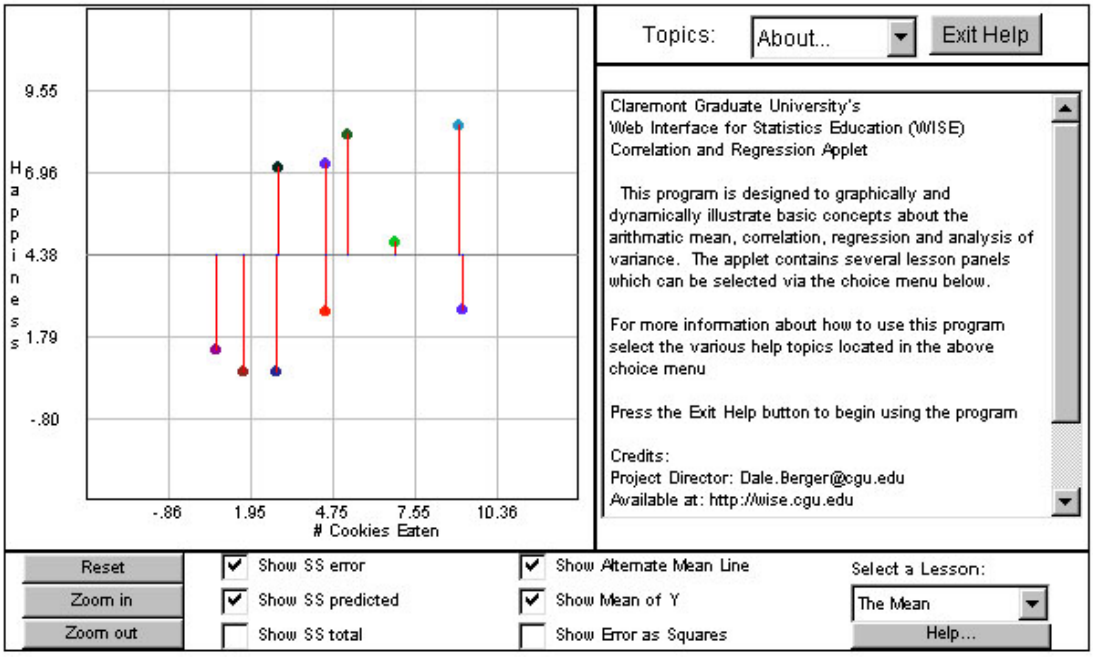
WISE ANOVA and Regression Lab

Introduction to the WISE Correlation/Regression and ANOVA Applet

This module focuses on the logic of ANOVA with special attention given to variance components and the relationship between ANOVA and regression.

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' then choose Module 4: ANOVA Variance Components

Using the WISE Correlation/Regression/ANOVA 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 may want to refer back to this page when you get to the part of the tutorial that uses the applet.

Module #4. Calculations and observations based on a small dataset.

Module #4, Interactive Exercise #1 Regression with a small dataset

In this module, you will learn how to use the WISE regression applet to deepen your understanding of the relationship between regression and ANOVA. 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. If you completed Modules #1 and #2 this will be review. Answers are provided for most problems at the end of this handout. In Exercise 2, you will compare regression calculations to ANOVA calculations for these same data.

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*.

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 the formula for a straight line that best fits the data. The regression equation can be used to predict scores (called Y′, or Y-prime). For each of our X-values, we can predict Y. Below you will complete several calculations, deriving y-prime for each value of X. First, you will need the regression equation. The general form of this equation is $Y' = a + bx$

In our example, a=2.0 and b=2.0, so the regression equation is $Y' = 2.0 + 2.0X$. Our first X score is 1, which generates a predicted Y score of 4.0, from 2.0 + 2.0(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 referring to the final pages of this handout but make sure that you can do the calculations and interpretations yourself.

Case	**X**	**Y**	**Y′**
1	1	2	4
2	1	6	
3	2	5	
4	2	7	
Sum			

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 $\sum (Y - \overline{Y})^2$. 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**	**Y´**	$(Y - \overline{Y})$	$(Y - \overline{Y})^2$
1	1	2	4	2-5 = -3	$(-3)^2 = 9$
2	1	6			
3	2	5			
4	2	7			
Sum					Σ =

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 $(Y - \overline{Y})$. 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 +2, while the squared deviations are 9 and 4. 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? Why?

The contribution to SS Total for Case 3 is _____ because

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

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

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.

g. Calculations. Complete the calculations below using the predicted scores (**Y´**) calculated in question 1b.

Case	**X**	**Y**	**Y´**	**(Y - Y´)**	**$(Y - Y')^2$**
1	1	2	4	2-4 = -2	$(-2)^2 = 4$
2	1	6			
3	2	5			
4	2	7			
Sum					Σ =

h. 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 deviations are for Cases ____, and the size of the deviation is _____.

The smallest deviations are for Cases ____, and the size of the deviation is _____.

i. 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 _____

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.

j. 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	4	4-5 = -1	$(-1)^2 = 1$
2	1	6			
3	2	5			
4	2	7			
Sum					Σ =

k. 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?

l. 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 j.

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 _____

m. 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

o. Note that SS Total = SS Predicted + SS Error. (14.0 = 4.0 + 10.0). 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 = _____

Module #4, Interactive Exercise #2 ANOVA with the same data

This problem uses the data below (same data as in Exercise 1). These data are for a two-group example where our goal is to compare the means between Group 1 and Group 2. For this exercise, you will complete several calculations and observe how the values you calculate relate to the values provided by the applet. The paper and pencil section below steps you through calculation of each value. Answers are provided for all problems 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. Do not click any of the points on the applet or move points – this will create a distribution that does not match your calculations.

X = 1 represents scores for group #1, X = 2 represents scores for group #2.

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

a**. Means.** For the data calculate the means for Group 1 (the mean of the two Y scores for cases with X = 1), Group 2 (the mean of the two Y scores for cases with X = 2), and an overall mean for Y (mean of all four Y scores).

Mean for Group 1($\overline{Y}_1$) _____

Mean for Group 2 ($\overline{Y}_2$) _____

Grand Mean for Y ($\overline{Y}$) _____

SS Total (Total Variance)
SS total is the sum of squared deviations of observed Y scores from the mean of Y.

b. To calculate SS Total, take each value of Y, subtract the overall 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	1	6		
3	2	5		
4	2	7		
				Σ =

c. *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.

d. 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 = _____.

Do these values match? If not, check your calculations against the answers found in the final section of this handout.

e. Compare your SS Total calculated in the ANOVA section to the SS Total calculated in the regression section? Are they the same value?

SS Between Groups

SS Between Group is the part of the SS total that CAN be explained by group differences. This corresponds to the sum of squared deviations of the group mean from the overall mean.

f. Calculations. Complete the calculations below using the group means from above and the overall mean (5).

Case	**X**	**Y**	$\overline{Y}_j$	$\overline{Y}_j - \overline{Y}$	$(\overline{Y}_j - \overline{Y})^2$
1	1	2	4	4-5 = -1	$(1)^2 = 1$
2	1	6			
3	2	5			
4	2	7			
					$\Sigma =$

g. Sum the values in the column headed $(\overline{Y}_j - \overline{Y})^2$. Which SS value from the ANOVA section of the applet corresponds to your calculated value, is it SS Predicted or SS Error? (Note: If neither, check your calculations).

$\sum(\overline{Y}_j - \overline{Y})^2$ = SS Between = _____.

Which SS value does this correspond to from Exercise #1 (SS Total, SS Regression, or SS Error)? _____

h. Now *place a check mark in the boxes titled Show SS Predicted and Show Regression Line,* and remove checks from all other boxes. Deviations predicted values from the mean of Y on the regression line are shown in blue.

Where does the regression line intersect X = 1? Another way to think of this is what is the predicted value for Y when X = 1. (Hint for Group 1 it should be a whole number between 2 and 6).

Enter the value here _____

Which Y' value from Exercise #1 does this correspond to? Y' = _____ for X = _____

Where does the regression line intersect X = 2?

Enter the value here _____

Which Y' value from Exercise #1 does this correspond to? Y' = _____ for X = _____

What is the relationship between group means in ANOVA ($\overline{Y}_j$) and predicted scores in Regression (Y')?

SS Within Groups

The sum of squares within groups is the sum of squared deviations of observed Y scores from the mean of their group. This is an indication of the deviations from the group average. This is called the within group variance as it refers to the amount of deviation *within* each group (deviation from group's mean). SS Within Groups is the part of SS Total that CANNOT be explained by ANOVA.

i. Calculations. First, calculate the group mean for each group (often noted as $\overline{Y}_j$ with $\overline{Y}_1$ used to represent the first group $\overline{Y}_2$ used to represent the second group). To get the group mean, take

the average of scores for each group (i.e., for the X = 1 group and X = 2 group separately). Complete the calculations below.

Case	**X**	**Y**	$\overline{Y}_j$ (Group Mean)	$(Y - \overline{Y}_j)$	$(Y - \overline{Y}_j)^2$
1	1	2	4	2-4 = -2	$(-2)^2 = 4$
2	1	6			
3	2	5	6	(5-6) = -1	$(-1)^2 = 1$
4	2	7			
					$\Sigma =$

j. Now *place a check mark in the boxes titled 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 deviations are for Cases ____, and the size of the deviation is _____.

The smallest deviations are for Cases ____, and the size of the deviation is _____.

k. Sum the values in the column marked $(Y - \overline{Y}_j)^2$**.** Which SS value from the ANOVA section of the applet corresponds to your calculated value, is it SS Predicted or SS Error? (Note: If neither check your calculations).

$\sum (Y - \overline{Y}_j)^2$ = SS Within = _____.
Which SS value does this correspond to from Exercise #1 (SS Total, SS Regression, or SS Error)? _____

Calculation of F

l. Calculate F using your SS values.

First, you will need to calculation the Degrees of Freedom Between and Within

DF Between Groups = # Groups – 1 = _____
DF Within Groups = # People (or cases) - # of Groups = _____

F = [SS Between Groups/DF Between Groups] / [SS Within Groups/DF Within Groups] = _____/ _____ = _____. (Note: this is often expressed as MS Between / MS Within)

m. What does the applet report for F? _____

Note that SS Total = SS Between Groups + SS Within Groups. (14 = 4 + 10). Thus, with ANOVA, we split SS Total into two parts, SS Between Groups and SS Within Groups. We can compute the proportion of SS Total that is in SS Between. In terms of sums of squares, this is the ratio of SS Between to SS Total. This value is termed eta-squared (η^2).

n. Calculation of Eta-Squared (not given in applet) = [SS Between / SS Total] _____

Which value from Exercise #1 does η^2 correspond to? _____

Module #4, Exercise #3: Follow-Up Questions

a. Match the regression statistics to the ANOVA statistics

Regression	ANOVA
SS Total _____	a. SS Within Groups
SS Regression _____	b. η^2
SS Error _____	c. SS Between Groups
Y´_____	d. $\overline{Y}_j$
r^2 _____	e. SS Total
	f. F

b. Imagine we had a third group of two scores of 8 and 10. What would the predicted value (y´) be for the score of 8? _____

What would the predicted value be for the score of 10? _____

Answers to Selected Questions

Interactive Exercise #1 Regression with a small dataset

a. **Correlation, Slope, and Y-intercept**

r = .535, b = 2.0, a = 2.0

b. **The regression equation** is y′ = a +bx . a = y-intercept, b = Slope, y′ = predicted score for x.

Case	**X**	**Y**	**Y′**
1	1	2	2.0+2.0(1) = 4
2	1	6	2.0+2.0(1) = 4
3	2	5	2.0+2.0(2) = 6
4	2	7	2.0+2.0(2) = 6
Sum			

c. **SS Total Calculations**

Case	**X**	**Y**	**Y′**	$(\mathbf{Y} - \overline{\mathbf{Y}})$	$(\mathbf{Y} - \overline{\mathbf{Y}})^2$
1	1	2	4	2-5 = -3	$(-3)^2 = 9$
2	1	6	4	6-5 = 1	$(1)^2 = 1$
3	2	5	6	5-5 = 0	$(0)^2 = 0$
4	2	7	6	7-5 = 2	$(2)^2 = 4$
Sum					$\Sigma = 14$

d. Largest deviation from mean is –3 for Case 1.

e. Contribution of SS Total for Case 3 is 0 because Y = 5 is equal to the mean.

f. See answer for c above.

g. **SS Error Calculations**

Case	**X**	**Y**	**Y′**	**(Y - Y′)**	**$(Y - Y')^2$**
1	1	2	4	2-4 = -2	$(-2)^2 = 4$
2	1	6	4	6-4 = 2	$(2)^2 = 4$
3	2	5	6	5-6 = -1	$(-1)^2 = 1$
4	2	7	6	7-6 = 1	$(1)^2 = 1$
Sum					$\Sigma = 10$

h. The largest deviations are for Cases 1 and 2, and the size of the deviation is 2.The smallest deviations are for Cases 3 and 4, and the size of the deviation is 1.

i. See g above. Both values should be 10

j. **SS Predicted Calculations**

Case	**X**	**Y**	**Y´**	$(\mathbf{Y}' - \overline{\mathbf{Y}})$	$(\mathbf{Y}' - \overline{\mathbf{Y}})^2$
1	1	2	4	4-5 = -1	$(-1)^2 = 1$
2	1	6	4	4-5 = -1	$(-1)^2 = 1$
3	2	5	6	6-5 = 1	$(1)^2 = 1$
4	2	7	6	6-5 = 1	$(1)^2 = 1$
Sum					$\Sigma = 4$

l. See j above

o. **r-squared as proportion of variance explained**

[SS Predicted/ SS Total] = 4 / 14 = .286.
Applet reports r = .535; r squared = .286

Interactive Exercise #2 ANOVA with a small dataset

a**. Means**
Mean for Group 1($\overline{Y}_1$) = (2 + 6) / 2 = 4
Mean for Group 2 ($\overline{Y}_2$) = (5 + 7) / 2 = 6
Grand Mean for Y ($\overline{Y}$) = (2 + 6 + 5 + 7) / 4 = 5

b. **SS Total Calculations**

Case	**X**	**Y**	$(\mathbf{Y} - \overline{\mathbf{Y}})$	$(\mathbf{Y} - \overline{\mathbf{Y}})^2$
1	1	2	2-5 = -3	$(-3)^2 = 9$
2	1	6	6-5 = 1	$(1)^2 = 1$
3	2	5	5-5 = 0	$(0)^2 = 0$
4	2	7	7-5 = 2	$(2)^2 = 4$
				$\Sigma = 14$

d. See b above. Both values should be 14

f. **SS Between Groups Calculations**

Case	**X**	**Y**	$\overline{Y}_j$	$\overline{Y}_j - \overline{Y}$	$(\overline{Y}_j - \overline{Y})^2$
1	1	2	4	4-5 = -1	$(-1)^2 = 1$
2	1	6	4	4-5 = -1	$(-1)^2 = 1$
3	2	5	6	6-5 = 1	$(1)^2 = 1$
4	2	7	6	6-5 = 1	$(1)^2 = 1$
					$\Sigma = 4$

g. See f above for SS. Should correspond to SS Regression.

h. Regression line intersects X = 1 at Y = 4.
Which Y' values from Exercise #1 does this correspond to? Y' = 4 for X = 1
Regression line intersects X = 2 at Y = 6.
Which Y' values from Exercise #1 does this correspond to? Y' = 6 for X = 1

i. **SS Within Groups Calculations**

Case	**X**	**Y**	$\overline{Y}_j$(Group Mean)	$(Y - \overline{Y}_j)$	$(Y - \overline{Y}_j)^2$
1	1	2	4	2-4 = -2	$(-2)^2 = 4$
2	1	6	4	6-4 = 2	$(2)^2 = 4$
3	2	5	6	5-6 = -1	$(-1)^2 = 1$
4	2	7	6	7-6 = 1	$(1)^2 = 1$
					$\Sigma = 10$

k. See i for SS. Corresponds to SS Error.

l. **Calculation of F**

F = [4 / 1] / [10 / 2] = 4 / 5 = 0.8.

n. **Calculation of Eta-Squared**

4 / 14 = .286. Corresponds to r^2.