

Statistics Exploration

How Well Does a Regression Line Fit a Set of Data?

Name: _____ Class: _____ Date: _____

PURPOSE: This exploration allows you to explore how to determine how well a least-squares regression line fits a set of bivariate data.

MATERIALS NEEDED: TI-83 Calculator
 REGRFIT program for TI-83
 This worksheet

GOAL: Upon completion of this exploration, you should be able to understand, explain and interpret how well a least-squares regression line fits a set of bivariate data.

PART ONE

1. Complete the following table of calculations using the data in columns L1 and L2 (provided below).

Value of \bar{y} (mean of response variable (L2)): _____ (Round to 1 decimal place.)

L1 (x)	L2 (y)	\bar{y} 's Prediction of y	$y - \bar{y}$	$(y - \bar{y})^2$
1	7			
3	15			
4	14			
7	20			
8	29			

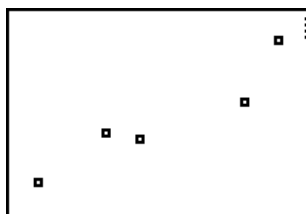
Sum of $(y - \bar{y})^2$: _____

2. Have your teacher check the values that you have entered into the table. When your teacher is satisfied with your entries, he or she will let you download the program REGRFIT to your calculator (TI-83).
3. The program will overwrite (erase) data in equation 1 (Y1) and pictures 8 and 9 (PIC8, PIC 9). If you have unsaved data in those locations that you wish to save, please do so now.
4. Run REGRFIT. Press enter and read each screen until you get to the dataset menu choice screen. (If you need to quit at any point in the program, press ON.)

- The first time that you run REGRFIT, use the program's data so that everyone in your group or class is using the same data. The values that REGRFIT uses by default in List1 (L1) and List 2 (L2) are those that are given in the table in step 1. If you have unsaved data in L1 and L2 that you wish to save, please do so before you choose to use the program's data because it will overwrite whatever is in L1 and L2. In Part Three, you will load your own data into L1 and L2 before running the program.



- You now have a graph of the data in L1 and L2. The data in L1 is the explanatory (independent or x) variable for the graph. The data in L2 is the response (dependent or y) variable for the graph.



- QUESTION 1:** What type of graph is on the screen?

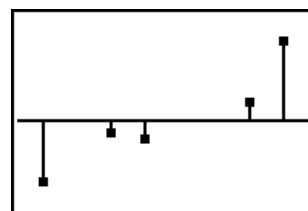
- Press ENTER

- The horizontal line that has been added to the graph is the mean (\bar{y}) line for the response variable.



- Press ENTER.

- QUESTION 2:** Vertical lines have been added to each point. What "statistical term" is used to refer to these lines?



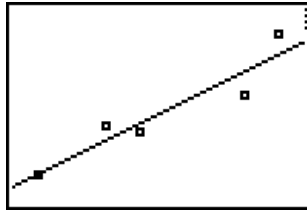
12. **QUESTION 3:** Refer back to the table in step 1 and the picture in step 11. What do the values for $y - \bar{y}$ represent? What in the picture represents these values?

13. Press ENTER.

14. A number has been added to the screen. Does it match your value for the sum of $(y - \bar{y})^2$ above? If it does not, go back and check your calculations above.

15. Press ENTER.

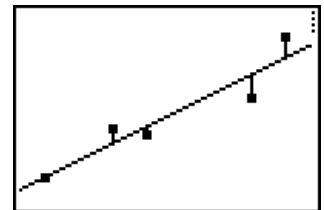
16. The original graph has been redrawn and a new line has been graphed. This new line is the least squares regression line (also called the linear regression line).



17. **QUESTION 4:** What is the purpose of the linear regression line (equation) of a set of bivariate data; that is, what is a linear regression line (equation) used for?

18. Press ENTER.

19. **QUESTION 5:** Vertical lines have been added to each point. What “statistical term” is used to refer to these lines?



20. Complete the following table of calculations using the data in L1 and L2 (provided below).

L1 (x)	L2 (y)	Regression Prediction of y	y – Predicted y	(y – Predicted y)²
1	7	7.3494		
3	15	12.7108		
4	14	15.3916		
7	20	23.4337		
8	29	26.1145		

Sum of $(y - \text{Predicted } y)^2$: _____

21. **QUESTION 6:** Refer back to the table in step 20 and the picture in step 19. What do the values for $(y - \text{Predicted } y)$ represent? What in the picture represents these values?
22. Press ENTER.
23. A number has again been added to the screen. Does it match your value for the sum of $(y - \text{Predicted } y)^2$ above? If it does not, go back and check your calculations above.
24. **QUESTION 7:** Compare the sums that you got for question 1 and question 21. Are they the same? If yes, explain why they should be the same. If no, explain why they should not be the same.

Statistics Exploration

How Well Does a Regression Line Fit a Set of Data?

Name: _____ Class: _____ Date: _____

PURPOSE: This exploration allows you to explore how to determine how well a least-squares regression line fits a set of bivariate data.

MATERIALS NEEDED: TI-83 Calculator
REGRFIT program for TI-83
This worksheet

PART TWO

1. To get the sums in question 1 and question 20 of Part One, you squared residual values for the data pairs and summed the squared values.

2. In question 1 of Part One, you squared the residual (error) differences between the response values in L2 and the sample mean, \bar{y} , of the response values in L2 (y). You then summed these squared values.

3. Such a sum is called the **total variation in y** or the **Sum of Squares T**otal; let's denote it as **SSTO**. It represents the amount of the variation in y that is explained by the mean value line.

4. Based on your Part One work, $SSTO =$ _____.

5. In question 20 of Part One, you squared the residual (error) differences between the response values in L2 (y) and the predicted values of the response values based on the least squares regression line for the response values in L2 (y). You then summed these squared values.

6. Such a sum is called the **Sum of Squared Errors**; let's denote it as **SSE**. It represents the amount of the variation in y that is explained by the linear regression line.

7. Based on your Part One work, $SSE =$ _____.

8. Like SSE, SSTO also represents an error value based on predictions of the response values where each prediction of y is the estimated mean value of all y 's, \bar{y} .

9. **QUESTION 1:** In your work above, which produced the smaller sum: making predictions of the response variable based on the mean value (SSTO) **or** making predictions of the response variable based on the least squares regression line (SSE)?
10. **How much better is the better of the two predictions?** Let's create an expression that will answer this question.

The difference in the two sums that you have calculated, SSTO and SSE, is a measure of the reduction (improvement) in the squared prediction error:

$$\text{reduction in squared prediction error} = \text{SSTO} - \text{SSE}$$

The ratio of the reduction in the squared prediction error to the total variation in y (SSTO) will serve as a measure of the percentage of reduction (improvement) in the squared prediction error as a result of using the least-squares regression line instead of the mean line.

$$\frac{\text{reduction in squared prediction error}}{\text{total variation}}$$

We can write this more symbolically as follow:

$$\frac{\text{SSTO} - \text{SSE}}{\text{SSTO}}$$

11. Use your values for SSTO and SSE to calculate the value of this new expression; express your answer as a decimal to four places.

$$\frac{\text{SSTO} - \text{SSE}}{\text{SSTO}} = \frac{\quad - \quad}{\quad} = \quad = \quad$$

12. What does this value represent?? Let's go back to the linear regression's correlation coefficient for a moment.

13. Use the initial values from the REGRFIT program in L1 and L2 of your calculator; these values may still be stored in your calculator.
14. Make certain that ***Diagnostics On*** has been turned on in your calculator. If you are not sure, it doesn't hurt to turn it on again.
15. Now, use your calculator to calculate the values for the linear regression of the data in L1 and L2.
16. Notice that two values are shown at the bottom of the list of values: r^2 and r . Write down the values of each:

$$r^2 = \underline{\hspace{10em}} \qquad r = \underline{\hspace{10em}}$$

17. The r -value is the correlation coefficient (more formally, the Pearson product-moment correlation).
18. **QUESTION 2:** What is the meaning of the correlation coefficient (r)?
19. Enter the value of r into the Home Screen of your calculator and square it. Does the answer equal the value of r^2 recorded above? It should.
20. Does the value for r^2 equal the value that you got for $\frac{SSTO - SSE}{SSTO}$ in question 11 above? It should, or at least be the same to four or more decimal places. (The regression prediction values used in question 21 of Part One were rounded to four decimal places.)

[NOTE: r^2 is also referred to as the **coefficient of determination**.]

21. What does all of this mean? Write a short paragraph about what you think these explorations have revealed to you so far.

Statistics Exploration

How Well Does a Regression Line Fit a Set of Data?

Name: _____ Class: _____ Date: _____

PURPOSE: This exploration allows you to explore how to determine how well a least-squares regression line fits a set of bivariate data.

MATERIALS NEEDED: TI-83 Calculator
REGRFIT program for TI-83
Additional data sets
This worksheet

PART THREE: Now you are ready to explore further on your own with new data.

RUN A

1. Choose a set of bivariate data (or use one assigned by your teacher) and enter it into your calculator (L1 and L2)
2. Run REGRFIT and choose “*YOUR OWN*” when you get to the dataset menu choice screen.
3. Write in the following values from your exploration:
 - a. Sum of Squares Total (SSTO) = _____
 - b. Sum of Squared Errors (SSE) = _____
 - c. Make certain that *Diagnostics On* has been turned on in your calculator. Use your calculator to calculate the values for the linear regression of the data in L1 and L2
 - d. $r^2 =$ _____ $r =$ _____
 - e. $\frac{SSTO - SSE}{SSTO} =$ _____ $=$ _____
4. Does $r^2 = \frac{SSTO - SSE}{SSTO}$?
5. Write a short paragraph about what you think this exploration has revealed to you.

RUN B

1. Choose another set of bivariate data (or use one assigned by your teacher) and enter it into your calculator (L1 and L2)
2. Run REGFIT and choose “YOUR OWN” when you get to the dataset menu choice screen.
3. Write in the following values from your exploration:
 - a. Sum of Squares Total (SSTO) = _____
 - b. Sum of Squared Errors (SSE) = _____
 - c. Make certain that *Diagnostics On* has been turned on in your calculator. Use your calculator to calculate the values for the linear regression of the data in L1 and L2
 - d. $r^2 =$ _____ $r =$ _____
 - e. $\frac{SSTO - SSE}{SSTO} = \frac{\quad - \quad}{\quad} =$
4. Does $r^2 = \frac{SSTO - SSE}{SSTO}$?
5. Write a short paragraph about what you think this exploration has revealed to you.

RUN C

1. Choose another set of bivariate data (or use one assigned by your teacher) and enter it into your calculator (L1 and L2)
2. Run REGFIT and choose “YOUR OWN” when you get to the dataset menu choice screen.
3. Write in the following values from your exploration:
 - a. Sum of Squares Total (SSTO) = _____
 - b. Sum of Squared Errors (SSE) = _____
 - c. Make certain that *Diagnostics On* has been turned on in your calculator. Use your calculator to calculate the values for the linear regression of the data in L1 and L2
 - d. $r^2 =$ _____ $r =$ _____
 - e. $\frac{SSTO - SSE}{SSTO} = \frac{\quad - \quad}{\quad} =$
4. Does $r^2 = \frac{SSTO - SSE}{SSTO}$?
5. Write a short paragraph about what you think this exploration has revealed to you.

Statistics Exploration

How Well Does a Regression Line Fit a Set of Data?

(AKA: Exploration of r-squared: the square of the correlation coefficient)

PURPOSE: This exploration allows you to explore how to determine how well a least-squares regression line fits a set of bivariate data.

GOAL: Upon completion of this exploration, you should be able to

1. explain how well a least-squares regression line fits a set of bivariate data.
2. understand the computations that lie behind r-squared.
3. interpret what r-squared means

WRAP-UP:

- ✓ **r-squared is a measure of how well a least squares regression line fits a set of bivariate data points.**
- ✓ It can assume values between 0 and 1.
 - A value of 1 implies a perfect fit to the data.
 - A value of zero means that there is no correlation between the explanatory (independent) variable and the response (dependent) variable.
- ✓ Expressed differently, **r-squared is the percent of variation in the response variable that is explained by the explanatory variable.**
- ✓ In the first run of REGRFIT using the programs data, you obtained the following values

$$SSTO = 266$$

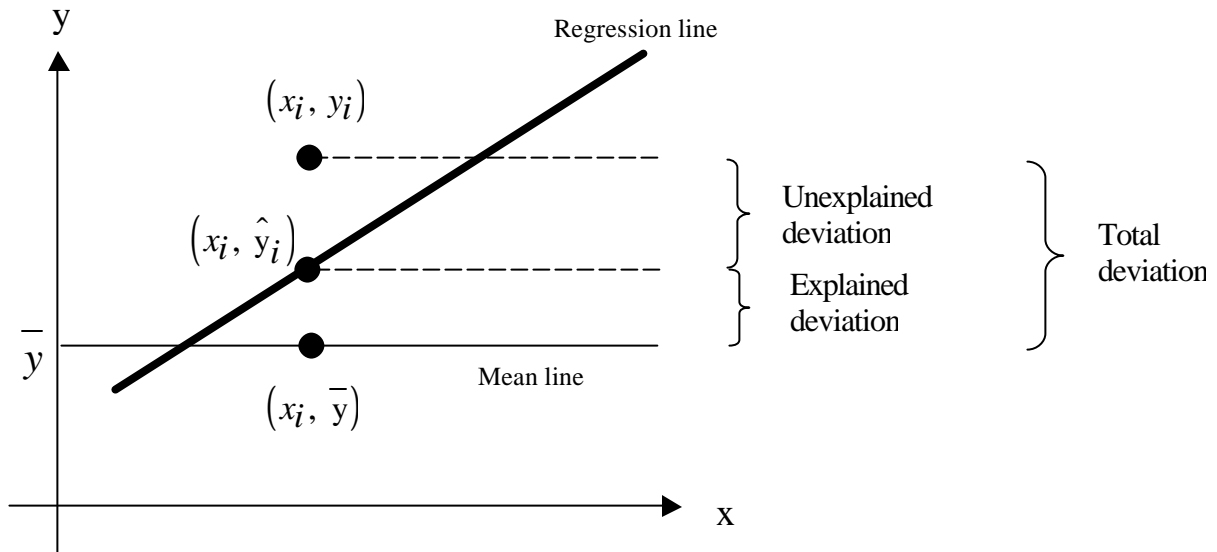
$$SSE = 27.41566$$

- ✓ You then calculated r^2 using both the linear regression capability of your calculator as well as the following expression: $\frac{SSTO - SSE}{SSTO}$. Both values were the same.

$$r^2 = \frac{SSTO - SSE}{SSTO} = \frac{238.58434}{266} \approx .8969$$

- ✓ This result means that approximately 89.7% of the variation in the response (dependent) variable is explained by the relationship between the explanatory (independent) variable and the response variable. The remaining 10.3% of the variation is unexplained and is due to other factors such as chance or sampling error.

See the diagram on the next page.



References/Credits

- REGRFIT program for the TI-83 calculator originally written as RSQUARED by John Lieb (The Roxbury Latin School, West Roxbury, MA) and modified by Joe Joyner (Norview High School)
- *Elementary Statistics: Picturing the World*; Larson, Ron and Farber, Becky; Prentice Hall Publishing, 2000, pg 442-443
- *Mind on Statistics*; Utts, Jessica and Heckard, Robert; Thompson Learning, Inc., 2002, pg 126-127