

Association of Quantitative Variables

MATH 360 Homework for Pre-Service Teachers

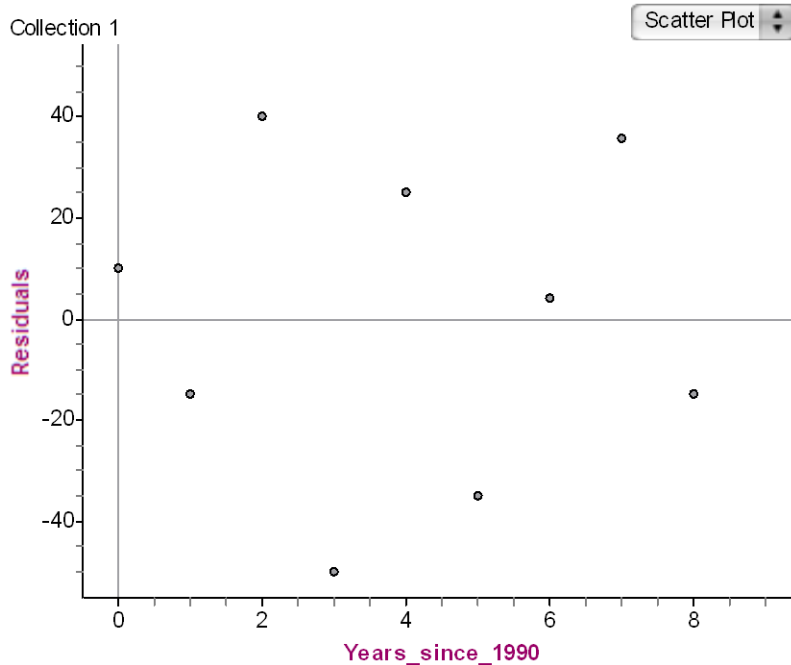
ASSIGNMENT ONE

1. The regression line _____ for the five points given in the table below was computed using the method of least squares.

| | | | | | |
|---|---|---|----|---|----|
| x | 1 | 2 | 3 | 4 | 5 |
| y | 6 | 3 | 10 | 7 | 15 |

- A. On graph paper, plot these five points and the regression line.
 - B. Using your work from part A, use pictures and words to demonstrate the meaning of the term “least squares”, i.e., why is it called the *least squares* regression line?
 - C. Why is the least squares criterion appropriate for a best fit line?
2. You are leading the class in analyzing the relationship between GPA and ACT scores for a data set. You have asked your students to use technology to find the correlation coefficient, coefficient of determination, and regression line for the data set. One pair of students asks for your help. They have done their work independently and are now comparing their answers. They have the same correlation coefficient & coefficient of determination, but their regression lines are different. They are asking you for help to understand how this could be possible. What is your response? Be specific in your description of your response, writing exactly what you would say and/or drawing anything you would use in your response.
 3. Adapted from College Board (1999). *AP Statistics Free-Response Questions*. New York, NY: Author.

Martha and Barry were searching the Internet to find information on air travel in the United States. They found data on the number of commercial aircraft flying in the United States during the years 1990-1998. The dates were recorded as years since 1990. Thus, the year 1990 was recorded as year 0. They fit a least squares regression line to the data. The graph of the residuals and part of the computer output for their regression are given below.



| <i>Predictor</i> | <i>Coef</i> | <i>Stdev</i> | <i>t-statistic</i> | <i>p</i> |
|------------------|-------------|--------------|--------------------|----------|
| <i>Constant</i> | 2939.93 | 20.55 | 143.09 | 0.000 |
| <i>Years</i> | 233.517 | 4.316 | 54.11 | 0.000 |
| s=33.43 | | | | |

- Is a line an appropriate model to use for this data set? What information tells you this?
- What is the value of the slope of the least squares regression line?
Interpret the slope in the context of this situation.
- What is the value of the y-intercept of the least squares regression line?
Interpret the y-intercept in the context of this situation.
- What was the actual number of commercial aircraft flying in 1992?
- Explain the meaning of the t-statistic value of 54.11 in the computer output, i.e. what does this numerical value of 54.11 tell you?

4. To start a lesson on lines of best fit, you are following the curriculum guide and have presented your students the following data set about the pounds of beans eaten by families of different sizes when traveling on the Overland Trail:

| | | | | | | | | | | | | | | | |
|------------------|----|----|----|----|-----|-----|----|-----|-----|----|-----|-----|-----|-----|-----|
| Number of people | 5 | 8 | 6 | 7 | 11 | 10 | 5 | 7 | 10 | 5 | 8 | 7 | 9 | 12 | 10 |
| Pounds of beans | 61 | 95 | 56 | 75 | 125 | 135 | 80 | 100 | 103 | 75 | 100 | 105 | 125 | 150 | 125 |

Note: This is a data set used in a mathematics curriculum, but is it unknown whether it is real data.

The first question in the activity asks the students to discuss ways in which you could use the information in the table to decide how many pounds of beans a family with 20 people would actually eat.

- A. Mary responds first, stating “You could look at people with like 10 people in their families and just double that amount that they use.”

How would you respond to this student? Be specific in your description of your response, writing exactly what you would say and/or drawing anything you would use in your response.

- B. The class is convinced that Mary’s strategy is the best one to take. However, the curriculum writers designed this question to be an introduction to the line of best fit, which is to be the focus of the day’s lesson. What would you do in this situation? Things to address include a way to convince the students that Mary’s strategy is not the best one; how to transition from this “scaling” method to the “best fit line” method; how to help the students look at the data from an aggregate view (looking at data set as a whole) as opposed to a case view (looking at one data point at a time).

5. A. Go to the following website. Practice placing the line of best fit on the scatterplot by eye (without any buttons activated), then compare it to a best fit line. Do this for multiple scatterplots until you feel confident in your ability to accurately place a line of best fit by eye.

<http://hspm.sph.sc.edu/courses/J716/demos/LeastSquares/LeastSquaresDemo.html>

Now that you’re an expert at fitting a line by eye, why is a least squares regression line still a useful tool?

B. Students in your class have been doing an activity where they place the line of best fit on scatterplots by eye then come up with their own criteria for determining where to place the line. Below are some of the student responses. For each response, determine if it is generally a good criteria (works in all reasonable cases), create counter examples of scatterplots where that criteria would create a poor line of best fit (if possible), and provide your detailed response to the student.

JOE: Goes through the most points

SUSAN: Equal number of points above and below the line

JONATHAN: Connect the first and last points

(optional) Anticipate other possible student criteria & formulate responses to those students.

ASSIGNMENT TWO

1. A student in your class raises his hand and asks “What’s the difference between correlation, association, and regression?”. How would you respond to this student? Be specific in your description of your response, writing exactly what you would say and/or drawing anything you would use in your response.
2. Create three distinctly differently patterned scatterplots which all have a correlation coefficient of approximately zero. Explain why the correlation coefficient is near zero for each of the cases.
3. Page 273 of the textbook presents some common transformations used to linearize data. Pick one of them and show why that transformation works to linearize the data. For example, if the original data set has an exponential relationship between x and y , why is the relationship between x and $\ln(y)$ linear?
4. You have been teaching your class about finding the best model for a data set. A student says “So I just try all of these different transformations on the data set and whichever one gives me the biggest value of r is the best one, right?”. How would you respond to this student? Be specific in your description of your response, writing exactly what you would say and/or drawing anything you would use in your response.
5. One way to calculate the correlation coefficient r is through the following formula:

- A. Define what $\sum(x - \bar{x})(y - \bar{y})$ and $\sum(x - \bar{x})^2$ represent in this formula & how to find each.
- B. Explain why the correlation coefficient is calculated this way, keeping in mind that the purpose of this statistic is to measure the strength and direction of the linear association between the x - and y -variables (How is it calculating the strength? How is it calculating the direction?). Include reasoning for every aspect of this formula, including the z -scores, the product of the z -scores, the summation of the product of the z -scores, and division by $n-1$.
- C. Use this formula to explain why the correlation coefficient is unaffected by linear transformations of the data (e.g., if all of the x -value have 5 added to them, the value of r doesn’t change. Why?)
- D. Use this formula to explain why the correlation coefficient is unit-less (which gives it the advantage to be compared across various data sets!)
- E. Use this formula to explain why the correlation coefficient of y on x is the same as the correlation coefficient of x on y .
- F. Use this formula to explain why the correlation coefficient is sensitive to influential points.