# *Teaching About Regression Before AP Statistics* **Ruth Wunderlich**

NCTM Philadelphia November 2016 RuthEllen@me.com
(717) 380-7023
Lititz, PA

This is a collaborative **graded assignment**.
You are expected to work on the problems together and hand in your own paper.

Clearly communicate your steps and explain your process. Do your work on a separate sheet of paper. You may include this sheet as a cover sheet.

1) Calculate the least square estimator of the following set of numbers: {4, 7, 8, 13, 27}
   a) Create a function $f(x)$ equal to the sum of the squares of the differences from x.
   b) Identify the value of x that minimizes the sum of the squares of the differences. Present an algebraic solution.

2) Generalize the least square estimator of the set of numbers: $\{x_1, x_2, x_3, \ldots x_n\}$ Name the solution, and write it using summation notation.

3) Find the least squares regression line in the form $y = kx$ for the following points: (1, 1), (1, 2), (3, 2) and (4, 5). Make a graph showing the points and your calculated regression line on the same coordinate grid.

4) Generalize the least squares regression line in the form $y = kx$ for n points: $(x_1, y_1)$, $(x_2, y_2)$, $(x_3, y_3)$ . . . $(x_n, y_n)$. Use summation notation to indicate a formula for *k*, the slope of the regression line.

5) Calculate the balance in an account when \$25 is invested at 100% annual interest, compounded 10,000 times every year, after 1 year, after 4 years, after 6 years and after 7 years. Use these points, and find a best-fit model for the balance in the account as a function of year using transformed data and linear regression.

6) Notes on the slope (b) of the actual regression line. $\hat{y} = a + bx$

(Ruth is looking for work. Gotta lead?)

1) Calculate the least square estimator of the following set of numbers: {4, 7, 8, 13, 27}

a) Create a function $f(x)$ equal to the sum of the squares of the differences from x.

b) Identify the value of x that minimizes the sum of the squares of the differences. Present an algebraic solution.

a)

$$(4-x)^2 = 16 - 8x + x^2$$
$$(7-x)^2 = 49 - 14x + x^2$$
$$(8-x)^2 = 64 - 16x + x^2$$
$$(13-x)^2 = 961 - 26x + x^2$$
$$(27-x)^2 = 729 - 54x + x^2$$

$$f(x) = 1819 - 118x + 5x^2$$

b) $f(x) = 5x^2 - 118x + 1819$

Minimize $f(x)$

$$x = \frac{118}{2(5)} = \boxed{11.8}$$

2) Generalize the least square estimator of the set of numbers: {$x_1$, $x_2$, $x_3$, . . . $x_n$ } Name the solution, and write it using summation notation.

$$\sum (X_i - x)^2 = \sum ( x_i^2 - (2 x_i) x + \sum x^2$$
$$= \sum x_i^2 - (2 \sum x_i) x + \sum x^2$$
$$= n x^2 - (2 \sum x_i) x + \sum x_i^2$$

minimize

$$x = \frac{2 \sum x_i}{2n} = \boxed{\frac{\sum x_i}{n}, \text{ the mean}}$$

3) Find the least squares regression line in the form $y = kx$ for the following points: (1, 1), (1, 2), (3, 2) and (4, 5). Make a graph showing the points and your calculated regression line on the same coordinate grid.

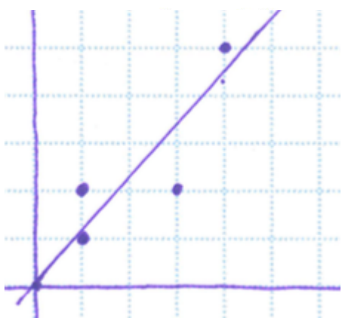
| x y | deviation $y - kx$ | |
|---|---|---|
| (1, 1) | $(1-k)^2$ | $= 1 - 2k + k^2$ |
| (1, 2) | $(2-k)^2$ | $= 4 - 4k + k^2$ |
| (3, 2) | $(2-3k)^2$ | $= 4 - 12k + 9k^2$ |
| (4, 5) | $(5-4k)^2$ | $= 25 - 40k + 16k^2$ |
| | | $34 - 58k + 27k^2$ |

$27k^2 - 58k + 34$

minimize $\cup$

$$k = \frac{56}{2(27)} = \frac{29}{27}$$

$$\boxed{y = \frac{29}{27}x}$$



4) Generalize the least squares regression line in the form $y = kx$ for n points: $(x_1, y_1)$, $(x_2, y_2)$, $(x_3, y_3)$ . . . $(x_n, y_n)$. Use summation notation to indicate a formula for *k*, the slope of the regression line.

$$\sum (y_i - kx_i)^2 = \sum (y_i^2 - 2x_i y_i k + x_i^2 k^2)$$

$$\sum y_i^2 - 2k \sum x_i y_i + k^2 \sum x_i^2$$

$$\sum x_i^2 \, k^2 - 2 \sum x_i y_i \cdot k + \sum y_i^2$$

minimize $\cup$

$$k = \frac{2 \sum x_i y_i}{2 \sum x_i^2} = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$\boxed{y = \frac{\sum x_i y_i}{\sum x_i^2} x}$$

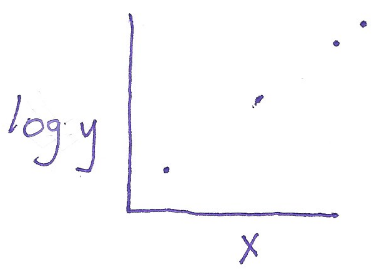
5) Calculate the balance in an account when \$25 is invested at 100% annual interest, compounded 10,000 times every year, after 1 year, after 4 years, after 6 years and after 7 years. Use these points, and find a best-fit model for the balance in the account as a function of year using transformed data and linear regression.

$$\$25\left(1+\frac{1}{10{,}000}\right)^{10{,}000} = \$67.^{95} \qquad (1, 67.95)$$

$$\$25\left(1+\frac{1}{10{,}000}\right)^{10000\cdot 4} = \$1{,}364.^{70} \qquad (4, 1364.7)$$

$$\$25\left(1+\frac{1}{10{,}000}\right)^{10000\cdot 6} = \$10{,}083.^{00} \qquad (6, 10083)$$

$$\$25\left(1+\frac{1}{10{,}000}\right)^{10000\cdot 7} = \$27{,}406.^{00} \qquad (7, 27406)$$




$$10^{\log y} = 10^{(.434277x + 1.397919)}$$

$$y = \left(10^{.434277}\right)^{x} \cdot 10^{1.397919}$$

$$y = 2.71817^{x} \cdot 24.99879$$

$$y = \$25\, e^{x} \;!$$

$$\lim_{n\to\infty} P\left(1+\frac{r}{n}\right)^{nt} = Pe^{rt}$$

6) The slope of the regression line, $\hat{y} = a + bx$, can be written

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad \text{and} \qquad b = r\frac{s_y}{s_x}$$

Can you see $m = \frac{\Delta y}{\Delta x}$ in these formulas?

*r* is the correlation coefficient.

*r* measures the strength and direction of the <u>linear relationship</u> between two variables.

$s_y$ is the standard deviation of the y values.

$s_x$ is the standard deviation of the x values.

The LSRL contains the point, $(\bar{x}, \bar{y})$. If x is average we predict y to be average.

The slope of the regression line is the ratio of the standard deviations of the x and y values tempered by the strength of the linear relationship between x and y as measured by r, the correlation coefficient.

$$b = r\frac{s_y}{s_x}$$

This illustrates the statistical phenomenon of "regression towards the mean." If the x differs from the mean, we expect y to differ from it's mean but not by as much. Things are more often ordinary than they are extraordinary!