

## Calculus Challenge Problem #4

Due, Wed., Nov. 12

### Solutions

1. Suppose we are given the data  $\{2, 3, 4, 7, 9\}$ .

a) Find the value of  $x$  that minimizes  $S$  for the data set  $\{2, 3, 4, 7, 9\}$ . Remember, the derivative of a sum is the sum of the derivatives.

We know that  $S(x) = (x-2)^2 + (x-3)^2 + (x-4)^2 + (x-7)^2 + (x-9)^2$ , so

$S'(x) = 2(x-2) + 2(x-3) + 2(x-4) + 2(x-7) + 2(x-9)$ . Setting this equal to zero, we notice the common factor of 2, so  $(x-2) + (x-3) + (x-4) + (x-7) + (x-9) = 0$ . This gives  $5x - 25 = 0$ , so  $x = 5$  is the least squares estimate.

Notice, this is the mean of the set  $\{2, 3, 4, 7, 9\}$ .

b) Generalize these results for  $n$  numbers in increasing order  $\{d_1, d_2, d_3, \dots, d_{n-1}, d_n\}$  to show that the mean is the least squares estimate for a set of data.

To minimize  $S(x) = \sum_{i=1}^n (x-d_i)^2$ , we differentiate and set the derivative equal to zero. So,

$$S'(x) = \sum_{i=1}^n 2(x-d_i) = 0. \text{ Then } 0 = \sum_{i=1}^n (x-d_i) = \sum_{i=1}^n (x) - \sum_{i=1}^n (d_i) = nx - \sum_{i=1}^n (d_i). \text{ So, } x = \frac{\sum_{i=1}^n (d_i)}{n}.$$

The mean is always a least squares estimate.

c) What value of  $x$  would minimize  $T(x) = \sum_{i=1}^5 |x-d_i|$ . Can you use calculus to help with this problem?

For the set  $\{2, 3, 4, 7, 9\}$ , we have

$$T(x) = \sum_{i=1}^5 |x-d_i| = |x-2| + |x-3| + |x-4| + |x-7| + |x-9|. \text{ Calculus can't help us here, since this}$$

function is not differentiable. But we know that  $|a| = \begin{cases} a & \text{when } a \geq 0 \\ -a & \text{when } a < 0 \end{cases}$ , so for  $x < 2$ , the

expressions in all of the absolute value signs are negative.

$$\text{This gives the line } T(x) = (-x+2) + (-x+3) + (-x+4) + (-x+7) + (-x+9) = -5x+25.$$

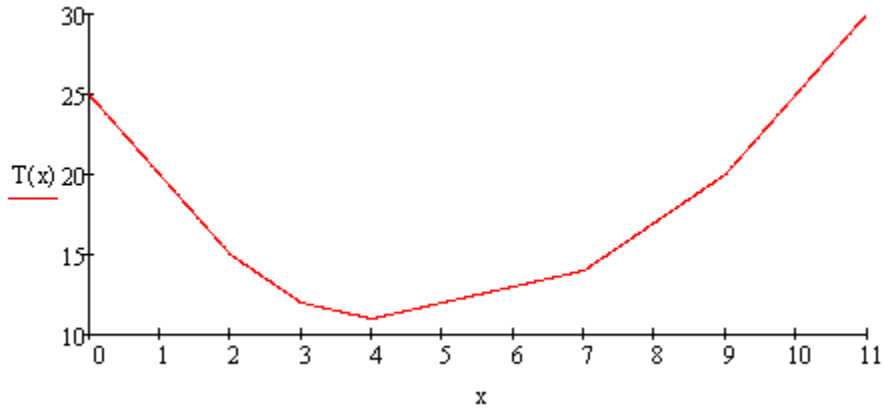
$$\text{Between 2 and 3, we have } T(x) = (x-2) + (-x+3) + (-x+4) + (-x+7) + (-x+9) = -3x+21.$$

$$\text{Between 3 and 4, we have } T(x) = (x-2) + (x-3) + (-x+4) + (-x+7) + (-x+9) = -x+15.$$

Between 4 and 7, we have  $T(x) = (x-2) + (-x+3) + (x-4) + (-x+7) + (-x+9) = x+7$ .

Between 4 and 7, we have  $T(x) = (x-2) + (-x+3) + (x-4) + (x-7) + (-x+9) = 3x-7$ .

Finally, for  $x > 9$ , we have  $T(x) = (x-2) + (-x+3) + (x-4) + (x-7) + (x-9) = 5x-25$ .



This piece-wise linear equation has its minimum when the slope turns from negative to positive. This is at  $x = 4$ . Notice this is the median of the set. The median minimizes the sum of the absolute values of the deviations.

2. Suppose we want to fit a line through the origin (that is  $y = kx$ ) to the ordered pairs  $\{(1,1), (2,3), (3,4)\}$  using the method of least squares. In this case, we can define

$$S(k) = \sum_{i=1}^3 (y_i - kx_i)^2.$$

a) Find the least squares estimate of  $k$  that minimizes  $S$ . Remember that  $k$  is the variable.

Now we have  $S(k) = (1-k)^2 + (3-2k)^2 + (4-3k)^2$ , so

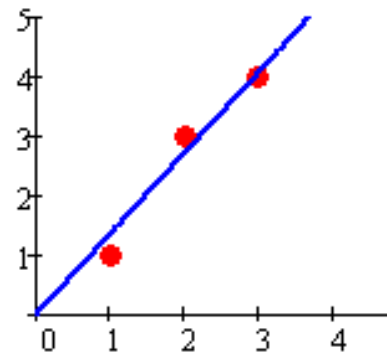
$$S'(k) = 2(1-k)(-1) + 2(3-2k)(-2) + 2(4-3k)(-3) = 0.$$

Simplifying, we have

$0 = -19 + 14k$ . Our least squares estimate for the slope of

a line through the origin is  $k = \frac{19}{14}$ . Look at the graph of

$y = \frac{19}{14}x$  graphed against the data.



b) Generalize these results for  $n$  ordered pairs  $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$  to show that

$$k = \frac{\sum xy}{\sum x^2}.$$

Now we have  $S(k) = \sum_{i=1}^n (y_i - kx_i)^2$ , so  $S'(k) = \sum_{i=1}^n 2(y_i - kx_i)(-x_i) = 0$ . Simplifying, we find

that  $0 = \sum_{i=1}^n (-x_i y_i) + k \sum_{i=1}^n (x_i^2)$ . Solving for  $k$ , we find  $k = \frac{\sum xy}{\sum x^2}$ . We see in part a) that

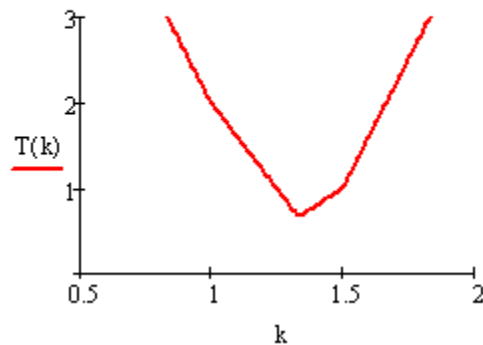
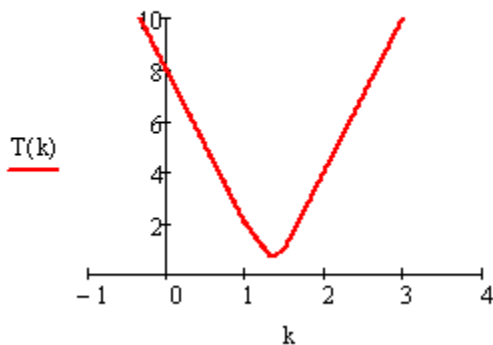
$$1 \cdot 1 + 2 \cdot 3 + 3 \cdot 4 = 17 \text{ and that } 1^2 + 2^2 + 3^2 = 14.$$

c) Find a value of  $k$  that minimizes  $T(x) = \sum_{i=1}^3 |y_i - kx_i|$  (minimize the sum of the absolute values of the deviations). Can you generalize this result?

So, now we consider  $T(k) = \sum_{i=1}^3 |y_i - kx_i|$  for the ordered pairs  $\{(1,1), (2,3), (3,4)\}$ .

$T(k) = |1 - k| + |3 - 2k| + |4 - 3k|$ . Again, we have a piece-wise linear function.

$$T(k) := |1 - k| + |3 - 2k| + |4 - 3k|$$



Using  $|a| = \begin{cases} a & \text{when } a \geq 0 \\ -a & \text{when } a < 0 \end{cases}$ , the intervals on which the lines are defined depend upon the zeros

for each absolute value. So, we have  $(-\infty, 1), (1, \frac{4}{3}), (\frac{4}{3}, \frac{3}{2}), (\frac{3}{2}, \infty)$ .

$$\text{So, } T(k) = \begin{cases} -6k + 8 & \text{for } x < 1 \\ -4k + 6 & \text{for } 1 \leq x < \frac{4}{3} \\ 2k + 2 & \text{for } \frac{4}{3} \leq x < \frac{3}{2} \\ 6k - 8 & \text{for } x < 1 \end{cases}$$

The minimum is at  $k = \frac{4}{3}$ . Notice that this is the median value of  $\frac{y_i}{x_i}$ .

3. Suppose we want to fit a function  $y = x^k$  to the ordered pairs  $\{(1,1), (2,4), (3,10)\}$ .

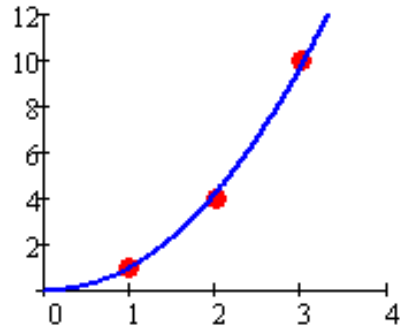
There are two ways to go about this. The traditional approach is to perform a log-log transformation and fit a line. If  $y = x^k$ , then  $\ln y = k \ln x$ . So, use the result in 2 b) to fit a line of the form  $Y = kX$  with  $Y = \ln(y)$  and  $X = \ln(x)$ .

a) Do a linear least squares estimate for  $k$  on the ordered pairs  $(\ln x, \ln y)$ . How does the model fit the original data?

We know that

$$k = \frac{\sum XY}{\sum X^2} = \frac{\sum \ln(x) \ln(y)}{\sum (\ln x)^2} \approx \frac{3.49055}{1.6874} \approx 2.069.$$

So, our model is  $y = x^{2.069}$ .



b) The more sophisticated method is to try to minimize  $S = \sum_{i=1}^3 (y_i - x_i^k)^2$ . Approximate the value of  $k$  that minimizes  $S$ . What problems do you run into trying this approach?

Compare the graph to the result of the log-log transformation (you will need to use this technique on a Challenge Problem later in the year).

In this case, we have  $S = \sum_{i=1}^3 (y_i - x_i^k)^2$ , so  $\frac{dS}{dk} = \sum_{i=1}^3 2(y_i - x_i^k)(-x_i^k) \ln(x_i) = 0$ .

So, we need to solve the equation

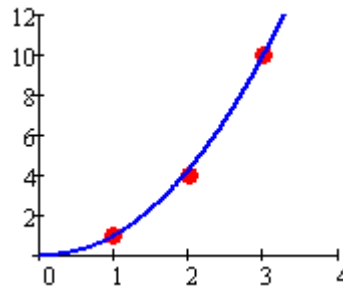
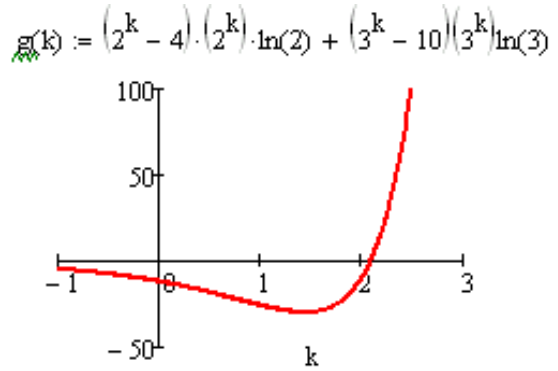
$(1-1^k)(-1^k) \ln(1) + (4-2^k)(-2^k) \ln(2) + (10-3^k)(-3^k) \ln(3) = 0$ . Unfortunately, this is not possible. There is no algebraic technique that will allow us to solve this equation. However, we can approximate the solution.

The first term is just 0, so we have  
 $(2^k - 4)(2^k) \ln(2) + (3^k - 10)(3^k) \ln(3) = 0$ .

The best we can do is to approximate the solution using Newton's method or some other numerical technique. In this case,  $k \approx 2.09$ .

This value and that given by the log-log re-expression will differ. This technique will always yield a smaller sum of squared residuals.

The graph of  $y = x^{2.09}$  is shown at right.



Extra Question that did not appear in the challenge:

4. Suppose we want to fit a function  $y = kx^2$  to the ordered pairs  $\{(1,1), (2,4), (3,10)\}$  using the method of least squares. In this case, we define  $S(k) = \sum_{i=1}^3 (y_i - kx_i^2)^2$ .

a) Find the least squares estimate of  $k$  that minimizes  $S$ .

Now we have  $S(k) = (1-k)^2 + (4-4k)^2 + (10-9k)^2$ , so

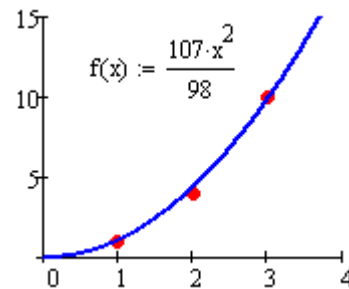
$$S'(k) = 2(1-k)(-1) + 2(4-4k)(-4) + 2(10-9k)(-9) = 0.$$

Simplifying, we have

$$0 = -107 + 98k. \text{ Our least squares estimate for the slope of a}$$

line through the origin is  $k = \frac{107}{98}$ . Look at the graph of

$$y = \frac{107}{98} x^2 \text{ graphed against the data.}$$



b) Generalize these results for  $n$  ordered pairs  $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ .

Now we have  $S(k) = \sum_{i=1}^n (y_i - kx_i^2)^2$ , so  $S'(k) = \sum_{i=1}^n 2(y_i - kx_i^2)(-x_i^2) = 0$ . Simplifying, we find

that  $0 = \sum_{i=1}^n (-x_i^2 y_i) + k \sum_{i=1}^n (x_i^4)$ . Solving for  $k$ , we find  $k = \frac{\sum x^2 y}{\sum x^4}$ . We see in part a) that  $1^2 \cdot 1 + 2^2 \cdot 4 + 3^2 \cdot 10 = 107$  and that  $1^4 + 2^4 + 3^4 = 98$ .