Calculus Challenge Problem #4 Due, Wed., Nov. 12

Least Squares Estimates

Often, when developing mathematical models in science or engineering applications, we need to use data to estimate the value of a parameter. The most common use of least squares estimates is to find estimate for the slope and intercept for the least squares line. That is, given a set of ordered pairs, find the equation of the best fitting line $y = a + bx$. Finding the values of *a* and *b* requires the use of partial derivatives, a concept beyond the first course in calculus.

However, there are many examples of one parameter estimates that you can do with the calculus you know right now.

1. Suppose we are given the data {2, 3, 4, 7, 9}, so $d_1 = 2$, $d_2 = 3$, $d_3 = 4$, $d_4 = 7$, $d_5 = 9$..

We want a single number that "best" represents these values (your teacher does this every time she takes all your test grades and puts a single grade on your report card). We call this number an "average". Typically, we define an average by minimizing some measure of distance (called the deviations) from each element of the set. A least squares estimate for the data set above would be the value of *x* that minimizes *S*, the sum of the squares of the deviations from *x*. So,

$$S(x) = \sum_{i=1}^{5} (x - d_i)^2 .$$

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.

b) Generalize these results for *n* numbers in increasing order $\{d_1, d_2, d_3, \cdots, d_{n-1}, d_n\}$. What value is this?

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?

2. Suppose we want to fit a line through the origin, of the form $y = kx$, to the ordered pairs $\{(1,1), (2,3), (3,4)\}$ using the method of least squares. A least squares estimate for *k* can be found by minimizing $S(k) = \sum_{i=1}^{3} (y_i - kx_i)^2$. (Notice that *S* is a function of *k*.)

a) Find the least squares estimate of *k* that minimizes *S* for these ordered pairs.

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

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

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

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 $\ln y = k\ln x$.

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?

b) The more direct, but difficult method is to try to minimize $S(k)=\sum_{i=1}^{3}\left(y_i - x_i^k\right)^2$.

Approximate the value of k that minimizes S. What problems do you run into trying this approach?