M442 Fall 2017 Assignment 2, due Friday Sept. 15

1. [10 pts] In developing our method of least-squares regression, we measured the distance between data points and the best-fit curve by vertical distance. In the case of a line, we could just as easily have measured this distance by horizontal distance from the line. For the following data, fit a line based on vertical distances and a second line based on horizontal distances, and draw both lines along with the data, all on the same plot. Give the slope and intercept for each line. (Note: Don't worry if a line isn't the best polynomial to fit through this data.)

Year (Fall)	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Tuition	2811	2975	3111	3247	3362	3508	3766	4098	4645	5132

Table 1: Average published tuition charge for public four-year schools.

2. [10 pts] Four sets of data of the form $\{(x_k, y_k)\}_{k=1}^{11}$ are defined in the M-file *anscombe.m*, available on the course web site. This data is taken from the paper "Graphs in Statistical Analysis," by F. J. Anscombe, in *The American Statistician* **27** (1973) 17-21.

a. For each data set, draw a scatter plot of the data, along with its least squares regression line, and give the slope and intercept associated with the fit. Describe the similarities and differences between the fits.

b. As we'll discuss in class, a reasonable estimate for standard deviation is s, where

$$s^2 = \frac{1}{N-m}\sum_{k=1}^{N}(y_k - f(x_k;\vec{p}))^2.$$

Here m is the number of parameters. Compute s for each of these data sets.

c. Another measure we'll discuss is the *coefficient of determination*, typically denoted R^2. We define

$$R^2 = 1 - \frac{E}{T},$$

where E is the usual SSR and T is the total sum of squares

$$T = \sum_{k=1}^{N}(y_k - \mu_y)^2; \quad \mu_y = \frac{1}{N}\sum_{k=1}^{N} y_k.$$

Compute R^2 for each of these fits.

3. [10 pts] Suppose a set of N data points $\{(x_k, y_k)\}_{k=1}^{N}$ appears to satisfy the relationship

$$y = ax + \frac{b}{x},$$

for some constants a and b. Find the least squares approximations for a and b.

4. [10 pts] Given a set of data points $\{x_k\}_{k=1}^{N}$, the mean (or average) is clearly

$$\mu_x := \frac{1}{N}\sum_{k=1}^{N} x_k,$$

and (a bit less clearly) the *sample variance* is

$$\text{Var }(\vec{x}) := \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \mu_x)^2.$$

Likewise, if we have an associated set of data points $\{y_k\}_{k=1}^N$, then the *sample covariance* is

$$\text{Cov }(\vec{x},\vec{y}) = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \mu_x)(y_k - \mu_y).$$

Show that when we use least squares regression to fit the data $\{(x_k, y_k)\}_{k=1}^N$ to the line $y = mx + b$ we obtain

$$m = \frac{\text{Cov }(\vec{x},\vec{y})}{\text{Var }(\vec{x})}$$
$$b = \mu_y - \mu_x \frac{\text{Cov }(\vec{x},\vec{y})}{\text{Var }(\vec{x})}.$$

5. [10 pts] Answer the following:

a. An alternative approach to the analysis we carried out in class for predicting a son's height based on the heights of his parents would be to use a multivariate fit with

$$S = p_1 + p_2 M + p_3 A,$$

where S denotes the son's height, M denotes mother's height, and A denotes father's height. (We avoid F here since it conflicts with our standard notation for the design matrix.) Use data stored in the M-file *heights.m* to find values for p_1, p_2, and p_3 for this fit. According to this model, which height is more significant for a son's height, the mother's or the father's? Write a MATLAB *anonymous* function for your model and evaluate it at $(M, A) = (60, 70)$; i.e., the case in which the mother is five feet tall and the father is five feet, ten inches. Estimate the standard deviation for your fit.

b. For the same data as in Part (a) find parameter values for a multidimensional polynomial fit of the form

$$S = p_1 + p_2 M + p_3 A + p_4 M^2 + p_5 AM + p_6 A^2.$$

Write a MATLAB *anonymous* function for your model and evaluate it at $(M, A) = (60, 70)$. Estimate the standard deviation for your fit, compare it with the standard deviation from Part (a), and discuss which model you find preferable.