M647, Spring 2011, Assignment 2, due Friday Feb. 4

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. (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. For each data set, draw a scatter plot of the data, along with its least squares regression line, and give the slope, intercept and standard deviation (estimate) associated with the fit. Describe the similarities and differences between the fits. In each case, use MATLAB's *polyval* command to predict y for $x = 15$, along with a standard deviation error.

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

$$s^2 = \frac{1}{N-q}\sum_{k=1}^{N}(y_k - mx_k - b)^2.$$

Here q is the number of parameters.

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] The goals of this problem are: (1) to review some important concepts from multidimensional calculus; and (2) to use them to carry out a vector form of the least squares minimization calculation.

First, recall that for a function $\vec{f}(\vec{x})$, with $\vec{x} \in \mathbb{R}^n$ and $\vec{f} \in \mathbb{R}^m$, the Jacobian matrix is

$$D_x\vec{f}(\vec{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}.$$

(There are several different common choices of notation for the Jacobian; here, we follow our PDE reference Evans.)

(4a) Show that if A is any $n \times n$ matrix and $\vec{x} \in \mathbb{R}^n$ is a column vector, then

$$D_x(A\vec{x}) = A,$$

and likewise

$$D_x(\vec{x}^{tr}A^{tr}) = A$$

(4b) Show that if $\vec{f} : \mathbb{R}^n \to \mathbb{R}^m$ and $\vec{g} : \mathbb{R}^n \to \mathbb{R}^m$ are regarded as row vectors then

$$D_x(\vec{f}(\vec{x}) \cdot \vec{g}(\vec{x})) = \vec{g}(\vec{x})D_x\vec{f}(\vec{x}) + \vec{f}(\vec{x})D_x\vec{g}(\vec{x}),$$

and in particular

$$D_x|\vec{f}(\vec{x})|^2 = 2\vec{f}(\vec{x})D_x\vec{f}(\vec{x}).$$

(4c) For this part, observe that the least squares error (from class)

$$E(\vec{p}) = \sum_{k=1}^{N}(y_k - \sum_{j=1}^{m} p_j F_{kj})^2,$$

can be expressed more compactly in the vector form

$$E(\vec{p}) = |\vec{y} - F\vec{p}|^2,$$

where $|\cdot|$ denotes Euclidean norm on $\mathbb{R}^N$. Show that

$$D_pE(\vec{p}) = 0 \Rightarrow \vec{p} = (F^{tr}F)^{-1}F^{tr}\vec{y}.$$

Keep in mind that we've been viewing $\vec{y}$ and $\vec{p}$ as column vectors, so you'll want to transpose to use Part (b).

5a. [5 pts] 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

$$S = p_1 + p_2M + p_3F,$$

where S denotes the son's M denotes mother's height, and F denotes father's height. 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 $(M, F) = (60, 70)$; 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.

5b. [5 pts] For the same data as in Part (a) find parameter values for a multidimensional polynomial fit of the form

$$S = p_1 + p_2M + p_3F + p_4M^2 + p_5FM + p_6F^2.$$

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