M647 Spring 2012 Assignment 5, due Friday Feb. 24

1. [10 pts] Recall from class that for the data in *heights.m* (available on the course web site) we fit son height as a function of midheight with a line

$$y = p_1 + p_2 x.$$

We found $p_1 = 4.0256$ and $p_2 = .9429$. Carry out the following analyses for this fit.

1a. Find 95% confidence intervals for the μ_x estimate $\sum_{j=1}^{m} p_j f_j(x)$ for midheights $x = 59.5$, $x = 69.5$ and $x = 79.5$. Explain why it's reasonable for your errors for $x = 59.5$ and $x = 79.5$ to be significantly larger than your error for $x = 69.5$.

1b. Find 95% confidence intervals for our parameter estimates p_1 and p_2.

2. [10 pts] In Problem 2 from Assignment 3, we modeled the heights data from *heights.m* with a model of the form

$$y = p_1 + p_2 x_1 + p_3 x_2.$$

Carry out the following analyses for this fit.

2a. Find 95% confidence intervals for the μ_x estimate $\sum_{j=1}^{m} p_j f_j(\vec{x})$ for parent heights (ordered (mother,father)) $(60, 60)$, $(60, 70)$ and $(70, 70)$.

2b. Find 95% confidence intervals for our parameter estimates p_1, p_2, and p_3.

3. [10 pts] We saw in Assignment 2 that the four data sets in the M-file *anscombe.m* (available on the course web site) are fit by almost precisely the same models, and give almost precisely the same predictions and prediction errors. We might ask, then, at what stage in our analysis will we recognize that only the first of these data sets can be reasonably described by the regression line? In order to answer this, compute the standardized residuals $\{r_k\}_{k=1}^{11}$ for each of these sets of data, and use these to analyze our assumption of Gaussian errors. Include a residuals plot for each of the data sets.

4. [10 pts] In this problem we will compare fits obtained by transforming an equation to linear form and using linear regression versus fits obtained directly from nonlinear regression. The Malthusian model for population growth is

$$\frac{dy}{dt} = ry; \quad y(0) = y_0,$$

with exact solution $y(t) = y_0 e^{rt}$. (As in our analysis in class of the logistic equation we will regard y_0 as a parameter.)

4a. Using the U.S. population data in *uspop.m* (available on the course web site), find regression values for y_0 and r using the linear relationship

$$\ln y = \ln y_0 + rt.$$

Plot your transformed data along with your regression line, and also plot the curve $y(t) = y_0 e^{rt}$ along with the original data. Compute the error

$$E(r, y_0) = \sum_{k=1}^{23} (y_k - y_0 e^{rt_k})^2.$$

4b. Use *lsqcurvefit.m* to fit the data in *uspop.m* directly to the nonlinear expression $y(t) = y_0 e^{rt}$. Plot your fit along with the data in this case, and compare your result with your result from (a). Also, compare your values of $E(r, y_0)$.

5. [10 pts] The *Gompertz* model for population growth is described through the ODE

$$\frac{dy}{dt} = -ry \ln(\frac{y}{K}); \quad y(0) = y_0,$$

with exact solution

$$y(t; r, K, y_0) = K(\frac{y_0}{K})^{e^{-rt}}.$$

5a. Write the Gompertz ODE in a linear form and use this form and the U.S. population data in *uspop.m* (available on the course web site) to obtain rough estimates for values of the parameters r, K, and y_0. (Notice that as with our analysis of the logistic model in class, we will treat y_0 as a parameter. You can use either forward differences or central differences to approximate the derivative.)

5b. Use *lsqcurvefit* to obtain nonlinear regression values for r, K, and y_0. Plot your fit along with the data. Which model better describes U.S. population growth, logistic or Gompertz?

5c. Linearize $y(t; r, K, y_0)$ about your parameter values from (b) to obtain 95% confidence intervals on your values. You can compute the derivatives symbolically in MATLAB or numerically, but notice that it's not particularly difficult to compute them by hand for this equation.