M442 Fall 2017 Assignment 3, due Friday Sept. 22

1. [10 pts] In this problem, we'll practice various R^2 (coefficient of determination) calculations associated with Galton's heights data (available from the course web site in *heights.m*).

a. Compute R^2 for both the model of son height as a function of midheight and the multivariate fit of son height as a function of both mother height and father height.

b. Compute the adjusted R^2 value for the same fits described in (a). What do you conclude from your results?

c. Compute the adjusted R^2 value for a fit of son height as a function of mother height alone, and also for a fit of son height as a function of father height alone. What do you conclude from your results?

2. [10 pts] Let E denote the usual SSR and let T denote the total sum of squares

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

For line regression, the *regression sum of squares* is

$$S = \sum_{k=1}^{N} (\mu_y - mx_k - b)^2.$$

Show that for line regression

$$T = E + S.$$

3. [10 pts] Suppose we have data $\{(t_k, y_k)\}_{k=1}^N$ for which the independent variables are equally spaced, with

$$h = t_{k+1} - t_k$$

for $k = 1, 2, \ldots, N-1$. Show that

$$y'(t_k) = \frac{y(t_k + h) - y(t_k - h)}{2h} + \mathbf{O}(h^2),$$

for $k = 2, 2, \ldots, N-1$.

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.)

a. 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

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

(Notice that MATLAB uses *log* for natural logarithm *ln*.) 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

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

b. Use *lsqcurvefit.m* to fit the data in *uspop.m* directly to the nonlinear expression $y(t) = y_0 e^{rt}$. (Use your parameter values from Part (a) as an initial approximation.) 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

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

a. 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 and K. (You can use either forward differences or central differences to approximate the derivative.)

b. Use *lsqcurvefit* to obtain nonlinear regression values for 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.) Compute the standard deviation for your fit s and also the coefficient of determination and the adjusted coefficient of determination (even though it's not entirely clear what these mean in the context of nonlinear regression.) Plot your fit along with the data. Which model better describes U.S. population growth, logistic or Gompertz?

c. Use your model to predict the population at $t = 230$ (2020), and use standard deviation as an error estimate.