Homework – Least Squares Due Thurs. 12/15/16

Be sure to include all MatLab programs used to obtain answers.

1. In the lecture notes, we created the **normal equations**, which were given by the matrix equation:

$$\begin{pmatrix} (n+1) & \sum_{i=0}^{n} x_i \\ \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{n} f_i \\ \sum_{i=0}^{n} x_i f_i \end{pmatrix}.$$

The solution of this matrix equation gives the linear least squares best fit line

$$f(x) = a_0 + a_1 x.$$

We also noted that Statistics texts usually give the following formula for finding the same linear least squares best fit line. Define the averages

$$\bar{x} = \frac{1}{n+1} \sum_{i=0}^{n} x_i \qquad \text{and} \qquad \bar{f} = \frac{1}{n+1} \sum_{i=0}^{n} f_i.$$

The best fitting slope and intercept are

$$a_1 = \frac{\sum_{i=0}^{n}(x_i - \bar{x}) f_i}{\sum_{i=0}^{n}(x_i - \bar{x})^2} \qquad \text{and} \qquad a_0 = \bar{f} - a_1 \bar{x}.$$

Show these two methods are equivalent.

2. Work **Problem 5.7** from the text. It states to generate 11 data points, $t_k = (k-1)/10$, $y_k = \text{erf}(t_k)$, $k = 1, ..., 11$.

a. Fit the data in a least squares sense with polynomials of degree 1 through 10. Compare the fitted polynomial with $\text{erf}(t)$ for values of t between the data points. How does the maximum error depend on the polynomial degree?

b. Because $\text{erf}(t)$ is an odd function of t, that is $\text{erf}(x) = -\text{erf}(-x)$, it is reasonable to fit the data by a linear combination of odd powers of t:

$$\text{erf}(t) \approx c_1 t + c_2 t^3 + \cdots + c_n t^{2n-1}.$$

Again, see how the error between data points depends on n.

c. Polynomials are not particularly good approximants for $\text{erf}(t)$ because they are unbounded for large t, whereas $\text{erf}(t)$ approaches 1 for large t. So using the same data points, fit a model of the form

$$\text{erf}(t) \approx c_1 + e^{-t^2}(c_2 + c_3 z + c_4 z^2 + c_5 z^3),$$

where $z = 1/(1+t)$. How does the error between the data points compare with the polynomial models?

3. Work **Problem 5.8** from the text. It states that here are 25 observations, y_k, taken equally spaced values of t.

```
t = 1:25
y = [ 5.0291 6.5099 5.3666 4.1272 4.2948
      6.1261 12.5140 10.0502 9.1614 7.5677
      7.2920 10.0357 11.0708 13.4045 12.8415
     11.9666 11.0765 11.7774 14.5701 17.0440
     17.0398 15.9069 15.4850 15.5112 17.6572]
y = y';
y = y(:);
```

a. Fit the data with a straight line, $y(t) = \beta_1 + \beta_2 t$, and plot the residuals, $y(t_k) - y_k$. You should observe that one of the data points has a much larger residual than the others. This is probably an *outlier.*

b. Discard the outlier, and fit the data again by a straight line. Plot the residuals again. Do you see any pattern in the residuals?

c. Fit the data, with the outlier excluded, by a model of the form

$$y(t) = \beta_1 + \beta_2 t + \beta_3 \sin(t).$$

d. Evaluate the third fit on a finer grid over the interval [0, 26]. Plot the fitted curve, using line style '-', together with the data, using line style 'o'. Include the outlier, using a different marker, '*'.