Statistical Analysis of Genetic Data
Math 155, Spring 2008
Jo Hardin
HW #4

Due Thursday, Feb 28

1. Find the slope and intercept that minimize the sum of squares in a weighted simple linear regression model.

2. Generate three different data sets: cubic, sin, uniform scatter. For each one, determine the "best" degree and span for fitting a loess smoother to the data. Include a plot of the best fits as well as a few sentences explaining why you think the fit is best.

   (a) cubic:
   ```
   x <- runif(50,-10,10)
   y <- x^3 + rnorm(50,0,4)
   ```

   (b) sin:
   ```
   x <- runif(50,0,10)
   y <- sin(x) + rnorm(50,0,.5)
   ```

   (c) uniform:
   ```
   x <- runif(50,0,1)
   y <- runif(50,0,1)
   ```

3. Five fourth grade children were selected at random from the entire class and timed in a short race. The times in seconds were 6.3, 4.2, 4.7, 6.0, and 5.7. Using Kolmogorov-Smirnov, test the two-sided hypothesis that the distribution of times is uniform on the interval from 4 to 8 seconds. Note that such a distribution is given by:

$$\begin{aligned} F(x) &= 0 \quad \text{for} \quad x < 4 \\ &= (x-4)/4 \quad \text{for} \quad 4 \leq x < 8 \\ &= 1 \quad \text{for} \quad 8 \leq x \end{aligned}$$

4. A friend says to you, "I made a qq-plot of my 10 data points versus normal quantiles. Then I calculated the correlation of the two variables, and I got 0.90. Since only 2.5% of the null test statistic values are above my test statistic, I can reject the null hypothesis that the data are normal." Explain what is wrong with the previous statement. In your explanation, make sure to demonstrate that you understand the critical values from the table.