INTRODUCTION TO SMOOTHING

One aspect of regression:

How does the "center" of the conditional distributions vary as a function of the explanatory variables?

e.g., How does E(Y|X = x) depend on x?

A *smooth*: A curve constructed (computationally) to go through or close to all points (x, f(x)) for a certain function. e.g.,

- A "mean smooth" goes through or close to all points (x, E(Y|X = x)
- A "median smooth" goes through or close to all points (x , med(Y|X = x).

Example: For fish data, we've seen:

- median smooth (transparency)
- lowess mean smooth (constructed by arc).

Note: The median smooth was easy to construct for the fish data, since there were just a few values of the explanatory variable.

Example: To construct a median smooth for haystack data, number of "slices" is a *smoothing parameter*.

Note:

1. What does the haystack smooth help us see in the data?

2. Arc also has a "slice smooth" function illustrating how a parameter is involved in creating a smooth.

Lowess smooth:

- <u>lo</u>cally <u>we</u>ighted <u>sc</u>atterplot <u>s</u>moother
- found on most statistical software .

Outline of how the lowess curve is calculated

- Start with data points (x_1, y_1), … (x_n, y_n).

- Select a *smoothing parameter* f between 0 and 1. (We'll use f = 0.5 for illustration.)

- For each i,

 a. Look at the half (if f = ½; 1/4 if f = 1/4, etc.) of the data with x values closest to x_i.

 b. Fit a line (using weighted least squares -- we may talk about this later) to these points in a way that give more weight to points with x closest to x_i.

 c. Replace y_i with y_i' = the y-value of the point on this line corresponding to x_i. (So y_i' "adjusts" y_i to be influenced by nearby data points.)

- After doing this separately for each i, repeat the procedure using points (x, y_i') (so the effect of points away from the trend will probably be less.)

- After a few iterations of this process, connect all the current "adjusted" points.