MSDS 7333 Quantifying the World

Locally Weighted Regression Smoothers

Curve Fitting

- Two Applications
  - Estimating a PDF
  - Regression
- No previous knowledge of functional form
  - Assume f(x) is continuous
  - Assume observations are an i.i.d. random sample
- Now …. regression

Simple Linear Regression

- Fit the model $Y = \beta_0 + \beta_1 X + \varepsilon$ to the data
  - β_0 is the intercept
  - β_1 is the slope
  - ε is the residual, and $\varepsilon \sim N(0, \sigma^2)$
- Uses
  - Assess the significance and strength of the relationship between Y and X
  - Predict future values of the mean response Y a given hypothetical value of X

OLR Review

- Musclemass Data
  - Age and Musclemass for 60 women ages 40 – 85

```
• site <- "http://www.users.muohio.edu/hughesmr/sta333/musclemass.txt"
• musclemass <-read.table(site,header=TRUE)
• attach(musclemass)
• plot(mass~age,data=musclemass)
• fit <- lm(mass~age, data=musclemass)
• summary(fit)
• predict(fit, newdata=data.frame(age=65), int="conf")
• abline(fit)
• Par(mfrow=c(1,2))
• plot(age,residuals(fit))
• qqnorm(residuals(fit))
```

Assumptions

- Error Assumptions $\varepsilon_i \sim$ iid $N(0, \sigma^2)$
- Normally distributed
- Constant variance
- Independent
- Linearity assumption
- A linear model correctly describes the relationship between x and y
- Unusual, isolated observations have the potential to dramatically alter the fit, or even the choice of model used.
- There is no error associated with X

Nonparametric Curve Smoothing

- We wish to investigate the relationship between variables X and Y
- We have n pairs of data (X_i, Y_i), $i = 1, 2, \ldots, n$
- Assume relationship has the form $Y = \phi(x) + \varepsilon$, where $E(\varepsilon) = 0$
- Various methods to estimate $\phi(x)$
  - Splines: piece together lower-order polynomials
  - LOESS (local regression smoother): weighted linear regression applied to the pairs for a local window of x's
  - Kernel method: analogous to density estimation
- How do we know we have a "good" $\phi(x)$?

Locally Weighted Regression Smoother

- Goal: Estimate $y = \phi(x)$ at $x = x_0$
- $\phi(x)$ can be approximated by a linear function
  $l(x) = \beta_0 + \beta_1(x - x_0)$ when x is near x_0.
- To fit $l(x)$ locally
  - Determine the k values of the X_i's that are nearest to x_0
  - k/n is a specified fraction (the span) of the total number of points
  - Let $N_k(x_0)$ denote this set of k points
  - Let $W(u)$, $0 \leq u \leq 1$ be a nonnegative weighting function with mode at $u = 0$
- Loess approximation finds the $l(x)$ that minimizes

$$\sum_{X_i \in N_k(x_0)} \left[Y_i - l(X_i)\right]^2 W\left(\frac{|x_0 - X_i|}{\Delta_{x_0}}\right)$$

$$\Delta_{x_0} = \max_{X_i \in N_k(x)} |X_i - x|$$

Loess Example

```
plot(mass~age, data=musclemass, main="Muscle Mass vs Age")
out <- loess(mass~age, data=musclemass)
curve(predict(out, newdata=data.frame(age = x)), add=TRUE)
```

- User can control
  - Degree of local polynomial
  - Smoothing parameter (span): defines the size of the neighborhood of a particular X value
  - Weight function
    - Determines how large a role each observation plays in fitting the LOESS curve at a particular point.
    - Gives most weight to points closest to the point of estimation

Pros and Cons

- Pros
  - Does not require the specification of a function describing the relationship.
  - Very flexible, making it ideal for modeling complex processes for which no theoretical models exist.
  - Can verify if a simpler model is reasonable

- Cons
  - Requires fairly large, densely sampled data sets in order to produce good models.
  - Does not produce a regression function that is represented by a mathematical formula.
  - Somewhat prone to the effects of outliers in the data set, like other methods.

Exercise: Respiratory Rates in Children

- A high respiratory rate is a potential diagnostic indicator of respiratory infection in children. To judge whether a respiratory rate is truly “high,” however, a physician must have a clear picture of the distribution of normal respiratory rates.
- To this end, Italian researchers measured the respiratory rates of n = 618 children between the ages of 15 days and 3 years (given in months). The data appear in the R workspace respiratory.RData.

```
load(url("http://www.users.muohio.edu/hughesmr/sta333/respiratory.RData"))
```

- Draw a LOESS curve using different values of the smoothing parameter on the fitted curve. Plot each curve on the same plot.

DataScience@SMU