ESTIMATING CONDITIONAL MEANS

Model Assumptions: Linear mean, constant variance, independence, and normality.

Sampling Distribution of Estimate of Conditional Mean:

- $\hat{E}(Y|x) = \hat{\eta}_0 + \hat{\eta}_1 x$ -- estimate of E(Y|x).

- $\hat{\eta}_0$ and $\hat{\eta}_1$ are linear combinations of the y_i's

  $\Rightarrow \hat{E}(Y|x)$ linear combinations of the y_i's

  $\Rightarrow \hat{E}(Y|x)$ has a normal distribution.

  (Why doesn't this follow just from normality of $\hat{\eta}_0$ and $\hat{\eta}_1$?)

- $E(\hat{E}(Y|x)|\, x_1, \ldots, x_n) = E(\hat{\eta}_0 + \hat{\eta}_1 x|\, x_1, \ldots, x_n)$

  $= E(\hat{\eta}_0|\, x_1, \ldots, x_n) + E(\hat{\eta}_1|\, x_1, \ldots, x_n)x$

  $= \eta_0 + \eta_1 x$

  $= E(Y|x)$

Thus: $\hat{E}(Y|x)$ is an unbiased estimator of E(Y|x).

- Calculations (left to the interested reader; consider covariances) will show:

$$\mathrm{Var}(\hat{E}(Y|x)|\, x_1, \ldots, x_n) = \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}\right)$$

Comments:

1. What does this say when x = 0?

2. The further x is from $\bar{x}$, the ___________________ the variance of the conditional mean estimate.

3. How does $\mathrm{Var}(\hat{E}(Y|x))$ depend on n and the spread of the x_i's?

Define the standard error of $\hat{E}(Y|x)$:

$$\text{s.e}\,(\hat{E}(Y|x)) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}}$$

As with $\hat{\eta}_0$ and $\hat{\eta}_1$, one can show that (with our model assumptions)

$$\frac{\hat{E}(Y \mid x) - E(Y \mid x)}{s.e.(\hat{E}(Y \mid x))} \sim t(n\text{-}2),$$

so we can use this as a test statistic to do inference on E(Y|x).

Confidence Bands

Plot the least squares regression line.

For each point (x,y) on the line plot the points

$$(x, y \pm s.e\,(\hat{E}(Y|x))).$$

This gives curves with equations

$$y = \hat{\eta}_0 + \hat{\eta}_1 x \pm \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{SXX}}.$$

More generally:

Consider *confidence bands*: Curves of the form

$$y = \hat{\eta}_0 + \hat{\eta}_1 x \pm c\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{SXX}},$$

for some constant c.

Example: If $c = t\hat{\sigma}$, where t is the 95th percentile for the t(n-2) distribution, then the confidence bands will show the 90% confidence intervals for $\hat{E}(Y|x)$ as x varies.

[Pictures]

Criterion for determining what type of curve a quadratic equation in x and y describes:

Given the quadratic equation

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0,$$

if the *discriminant* $B^2 - 4AC$ is positive, then the graph of the equation is a hyperbola (or a pair of intersecting lines in the degenerate case).

Put $y = \hat{\eta}_0 + \hat{\eta}_1 x \pm c\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}}$ in this form:

1) $$(y - \hat{\eta}_0 - \hat{\eta}_1 x)^2 = c^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}\right).$$

2) $$y^2 - 2\hat{\eta}_1 xy + \hat{\eta}_1^2 x^2 - \frac{c^2}{SXX}x^2$$
+ (terms of degree 1 and 2) = 0.

So $A = \hat{\eta}_1^{\,2} - \frac{c^2}{SXX}$

$B = -2\hat{\eta}_1$

$C = 1$

$$B^2 - 4AC = 4\hat{\eta}_1^{\,2} - 4[\hat{\eta}_1^{\,2} - \frac{c^2}{SXX}]$$

$$= \frac{c^2}{SXX} > 0$$

so the confidence bands are a hyperbola.

Note: The least squares regression line is *not* one of the axes of the hyperbola, since the confidence bands are "equidistant " from the line vertically, but not in the perpendicular direction.