ESTIMATING CONDITIONAL MEANS

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

Sampling Distribution of Estimate of Conditional Mean:

- $\hat{E}(Y|x) = \hat{\eta}_0 + \hat{\eta}_1 x$ is our estimate of E(Y|x). Note that this is a random variable (varying according to our choice of y_i's), so has a sampling distribution.

- Since $\hat{\eta}_0$ and $\hat{\eta}_1$ are linear combinations of the y_i's, so is $\hat{E}(Y|x)$. Hence $\hat{E}(Y|x)$ has a normal distribution. (Why doesn't this follow just from normality of $\hat{\eta}_0$ and $\hat{\eta}_1$?)

- $E(\hat{E}(Y|x)| x_1, \ldots, x_n) = E(\hat{\eta}_0 + \hat{\eta}_1 x| x_1, \ldots, x_n)$
  $= E(\hat{\eta}_0| x_1, \ldots, x_n) + E(\hat{\eta}_1| x_1, \ldots, x_n)x$
  $= \eta_0 + \eta_1 x = E(Y|x)$

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

- Calculations (left to the interested reader; you need to consider covariances) will show that

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

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 (under our model assumptions)

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

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

Confidence Bands

If we plot the least squares regression line, and then for each point (x,y) on the line plot the points (x,y ± s.e ($\hat{E}$(Y|x)), we will get two curves, with equations

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

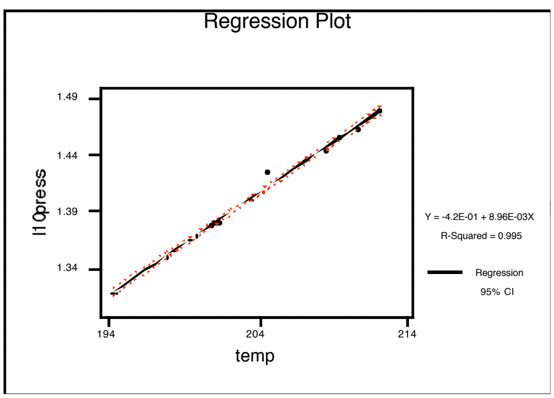
What kinds of curves are these? We will answer this a little more generally, looking at curves of the form

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

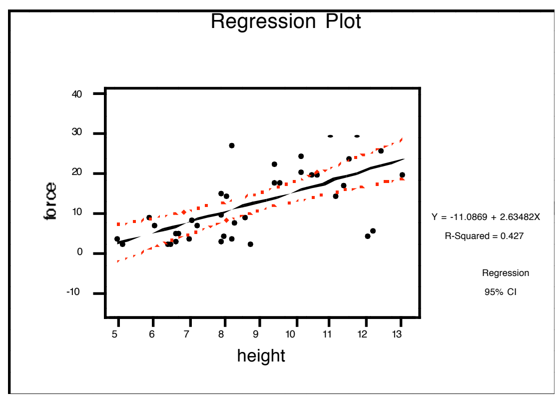
for some constant c. These are called *confidence bands*. For example, if we choose c = $t\hat{\sigma}$, where t is the 95th percentile for the t(n-2) distribution, then the curves will show the 90% confidence intervals for $\hat{E}$(Y|x) as x varies.

Example of confidence bands from Minitab:

Forbes data



Another example



We need the following 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). (For more information, see the Mathworld website at http://mathworld.wolfram.com/QuadraticCurveDiscriminant.html)

A little algebraic manipulation puts our equation (*) in the form

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

More algebra gives

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

So A = $\hat{\eta}_1{}^2 - \frac{c^2}{SXX}$, B = $- 2\hat{\eta}_1$, and C = 1, giving

$$B^2 - 4AC = 4\hat{\eta}_1{}^2 - 4[\hat{\eta}_1{}^2 - \frac{c^2}{SXX}] = \frac{c^2}{SXX} > 0,$$

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.]