Regression Summary

February 2, 2005

1 Symbols we'll use again and again

- $\bar{x} = \sum_i^n \frac{x_i}{n}$
- $s_y^2 = \frac{1}{n-1} \sum_i^n (y_i - \bar{y})^2$
- $\text{SXX} = \sum_i^n (x_i - \bar{x})^2 = \sum (x_i - \bar{x}) x_i$
- SYY is defined similarly, and is called the total sum of squares.
- $\text{SXY} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i - \bar{x}) y_i$
- $r = \frac{1}{n-1} \frac{\text{SXY}}{s_x s_y}$
- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, the predicted value at x_i.
- $e_i = y_i - \hat{y}_i$, the residual for observation i
- RSS $\sum e_i^2$, the residual sum of squares

2 Regression as Summary

The regression line is the line $\beta_0 + \beta_1 x$ that minimizes the squared error loss:

$$\sum_i^n (y_i - \hat{y}_i)^{=} \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

The values of β_0 and β_1 that minimize this are

$$\hat{\beta}_1 = \frac{\text{SXY}}{\text{SYY}} = \frac{r s_y}{s_x} = \sum (\frac{x_i - \bar{x}}{\text{SXX}}) y_i$$

.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

The first form for the estimate of the slope is the one that is probably easiest to solve for algebraically speaking, when you are deriving the estimates by taking the partial derivatives of the squared loss function. It is also probably the easiest to use if you ever find yourself forced to use a simple calculator. The second is the form that has the most intuition. It tells us that for groups that differ by one standard deviation in the x-value, the mean y differs by r times a standard deviation. The final form shows that the estimators can be written as linear combination of the y_i's.

A measure of goodness of fit – how tightly clustered the points are about the line, is r^2, literally the correlation coefficient squared. One can show that

$$r^2 = \frac{\text{SYY} - \text{RSS}}{\text{SYY}}$$

which shows that the r-squared, also called the coefficient of determination, measures the percent of total variation (SYY) that is explained by the regression line.

Also, we can relate the residual sum of squares — the variation about the regression line – to the variation of y and the correlation:

$$\text{RSS} = \text{SYY}(1 - r^2)$$

and this also shows that the residual sum of squares will always be less than the total sum of squares as long as r^2¿0 (as long as there is some linear relationship between the x and y).

A plot of the residuals against the predicted values (or the x value) will help you decide if the mean function is really related to x in a linear fashion. It will also help identify outliers and points that are potentially influential.

Cook's distance can help identify influential points. The higher the Cook's distance, the greater the influence a particular point has.

3 Inference

The model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\epsilon_i \sim \text{indpt} N(0, \sigma)$.

We now think of the βs as being population parameters. If we could see the entire population and fit a regression line, these are the values we would get. However, now we think of our points as a randomly selected subset from this population, selected so that the x values are known exactly, and so that subsquent measurements are independent, and the "errors" are normally distributed.

3.1 Variances and their estimates

We've already talked about the estimate of the intercept and slope. Now we also need to estimate the variance of the errors, σ^2:

$$\sigma^2 = \frac{\text{RSS}}{n-2}$$

$$\text{Var}(\hat{\beta}_0) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\text{SXX}})$$

$$\text{Var}(\hat{\beta}_1) = \sigma^2(\frac{1}{\text{SXX}})$$

The standard errors of the estimates are just the square-roots of these variances. (Recall that "standard error" is the term for the standard deviation of an estimator.) In practice, σ is not known, and so we replace it with the square-root of its estimate, given in the first of the three above equations.

3.2 Confidence Intervals

Confidence intervals are of the form estimate $\pm t_{n-2}(\alpha/2)$ times standard error(estimate).

The constant is chosen to get the correct confidence level. It is taken from a t-distribution with $n-2$ degrees of freedom. It also depends on the confidence level you wish. If the level you want is $1 - alpha$, then t is chosen so that $P(T > t) = \alpha/2$, where T is a random variable with a t distribution and n-2 degrees of freedom. In other words, the upper $\alpha/2$ quartile.

For example, for a 95% confidence interval and n = 10, then $alpha = 0.05$ and t is found from R using the command qt(1-.05/2, 8) which is equal to 2.036. The reason for "1−.05/2" rather than just .05/2 is that qt returns the number that has p% below it. We want instead the number that has .025% above it, and this is the same as the number that has 1-.025% below it.

3.3 Estimation and Prediction Intervals

To find a $1-\alpha\%$ confidence interval for the mean y value given a particular x value (lets call it x_0:

$$\hat{y}|x_0 \pm t_{n-2}(\alpha/2) * \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{(n-1)s_x^2}}$$

To find a CI to predict an individual response y given a value x_0,

$$\hat{y}|x_0 \pm t(n-2, \alpha/2) * \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{(n-1)s_x^2}}$$

Note that the margin of errors in both interval is widest when x_0 is far from the average. So it is easiest to estimate for points near the center.

3.4 Hypothesis Tests

To test the null hypothesis that the true value of the parameter (either the slope or the intercept) is 0 against the two-sided alternative that it is not 0, use

$$t = \frac{\text{estimate}}{\text{SE(estimate)}}$$

"SE" means "standard error. This statistic will follow a t distribution with n-2 degrees of freedom if the null hypothesis is true.

3.5 Diagnostics

A qq plot (qqnorm) of the residuals will assess whether the assumption of linearity is sound. If possible, a plot of residuals against the order in which the observations were collected will help detect at least one kind of dependence (which would violate the assumption that observations are independent.) A plot of the residuals against either the predicted values or the x values will detect whether σ is constant for all values of x (and is also good for checking that the model is truly linear).