Simple Linear Regression

The focus for now is on describing a linear relationship within a sample. We are not thinking of this as a sample from a larger population – this is purely descriptive.

We assume that there is good reason for thinking that the relationship is linear – either a theoretical reason, or preliminary scatterplots (with lowess curves, maybe) suggest that this is not out of the ballpark. However, even if this is not the case, we can still describe it as if it were, knowing that our description is a crude "first order" summary. In particular, when we say "the relation" we're talking about the model that describes how the mean response at a particular value of x, $E(Y|x)$, depends on x.

Goal is to find the "best" line to fit the data. Best is subjective, but for a few centuries now a popular approach has been to minimize the least-squares distance between the observations y, and the predictions, $\hat{y}$:

$$\sum_{i}^{n}(y_i - \hat{y}_i)^2$$

Later we'll see that this results in estimates with some nice properties.

Since we believe the relationship to be linear, we model our predictions $\hat{y} = \beta_0 + \beta_1 * x$.

The least squares approach finds the parameters that minimize the "residual sum of squares":

$$\text{RSS}(\beta_0, \beta_1) = (\sum_{i} y_i - \beta_0 - \beta_1 x_i)^2$$

Find the beta's that minimize this. To do this, differentiate with respect to both and set equal to 0. Solve simultaneously.

Solution can be "cleaned up" to look like this:

$$\hat{\beta}_1 = (r s_y)/s_x$$

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

Non-linear mean functions might not work out so easily, and might require numerical algorithms – which is where the art of statistical computation comes in.

In the future, it will be nice to think about these formulas slightly differently. So we're going to define some new terms:

SYY : $\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \bar{y}) y_i$ = "total variation (in y)"

SXX: defined similarly to SYY

SXY = $\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i (x_i - \bar{x}) y_i$. sum of cross products

residual $\hat{e}_i = y_i - \hat{y}_i$

RSS $\sum (\hat{e}_i)^2$

Now we can re-write some of your favorite formulas:

$$s_x^2 = \text{SXX}/(n-1)$$

$$r = \frac{\text{SXY}/s_x s_y}{n-1}$$

Can also show that

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

This formulation is interesting because it shows that the estimated slope (and therefore the intercept) are linear functions of the response. Estimators that are linear functions of the response are a special class – mostly special because they are easy to work with.

For example, for the Galton data of father and son heights:

predicted son's height = 33.88660 + .51409*father's height

Interpretation of the line

First, note what happens if we plug $\bar{x}$into the regression equation:

$$\hat{y} = (\bar{y} - \hat{\beta}_1\bar{x}) + \hat{\beta}_1\bar{x} = \bar{y}$$

In this case, we learn that fathers of average height have sons of average height, on average. (Lots of "average".)

Recall that average father = 67.6871, average son = 68.68407

So 33.88660+ .51409*67.6871 = 68.68

More generally, this shows that the regression line passes through the point $(\bar{x}, \bar{y})$.

In a way that we'll make more precise later, we can find the average y response for any x value. So we can find the average son height for any father height simply by plugging into the equation.

Ex: Find the average height of sons whose fathers are six feet tall (72")

33.8866+.51409*72 = 70.901

What the intercept tells us:

Well, not much. Often. It represents the predicted y value when x = 0. Here x=0 is far outside our range, and so we learn nothing. But sometimes it is useful.

What the slope tells us:

Here is where most of the interesting information lies.

First, think what it would mean if the slope were 0. It would mean that no matter what value of x you plugged in, you would always get the same predicted value of y. So x was no help in predicting y, and so there is no (linear) relation. So one of the first things we look at is whether or not the line is 0 or not.

If the value is positive, we have a positive association, and negative a negative association.

More precisely, the slope relates how the mean differs for different x-values. So if x-values differ by 1 unit, the mean response differs by (slope) units.

For the Galton data, this tells us that father's who were 1 inch taller had son's whose average heights were .51 inches taller.

There are many common misinterpretations. For example, "a one inch increase in x leads to a .51 increase in the mean of y". Or, "the sons of fathers who were 1" taller were .51 inches taller."

Also correct: "differences in 1 inch in the fathers' heights were associated with average differences of .51 inches in the the sons' heights."

Pay attention to what is being measured and how data were collected when you interpret. FAther's don't grow. Even if we're talking about something that could grow, like plants, ask whether or not we observed them growing. Is the regression line comparing changes for particular plants? Or is it comparing plants of different heights?

Note that the equation for slope tells us that when the SDs are equal in both groups (as they are in the father/son data) then the slope is the same as the correlation. Otherwise, not. But also note that what it tells us is that if the x variables are one SD apart, the means of the y's will be less than one SD apart. This is what Galton called regression to the mean.

The residuals contain what is "left over". If our model is complete, they should contain nothing but "noise" – random variation. Otherwise, the residuals could still have information in them. For this reason, they are a useful tool for evaluating the success of the model.

One thing to note (and prove)

$$\sum_i e_i = 0$$

. Which means a plot of the residuals against x will show that they are "centered" around the line $y = 0$. There should not be any patterns to such a plot – patterns suggest there are features in the data yet to be modeled.

There is a reslationship between the residuals and the correlation:

$$\text{RSS} = \sum_i \hat{e}_i^2 = \text{SYY}(1 - r^2)$$

This shows that the variation that is left over is always less than the total variation, unless the correlation is 0. Even better, solve for r^2

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

This shows that the squared correlation is a fraction of whatever variation is left over. This is what we mean when we say that r-squared is the percentage of variation explained by the fit. r-squared is called the "coefficient of determination".