12: SIMPLE LINEAR REGRESSION IV

The Coefficient of Determination

Once we have decided that β_1 is not zero, so that a linear relationship seems to exist between x and y, it is useful to measure the *strength* of this linear relationship. Such a measure is provided by the **coefficient of determination,** R^2.

To understand R^2, note that one of the aims of regression analysis is to study the relationship between x and y, i.e., to try to use the value of x to "explain" y.

- Keep in mind, though, that this "explanation" may

not be one of cause and effect.

Now, consider an example where x and y are the heights and weights of 500 people. If we look at the y's (the weights) as a data set, we note that in general they are not all the same; the y's exhibit variability. A rough measure of this variability is

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \ .$$

Note that SST is $(n-1)$ times the sample variance of the y's.

If there is a linear relationship between x and y, then the variability of the y's is not due entirely to chance fluctuations.

Instead, the fact that the weights are different can be partially "explained" by the fact that the heights (x) are different. Of course, weight is not completely explained by height, so part of the variability in the weights remains unexplained.

- Interestingly, the variability in people's weights can be broken into two parts, the first attributed to differences in height, and the second attributed to other factors not yet accounted for.

• We have the following important formula:

$$\sum(y_i-\bar{y})^2=\sum(\hat{y}_i-\bar{y})^2+\sum(y_i-\hat{y}_i)^2\ ,$$

or $$SST=SSR+SSE\ .$$

Interpretation:

The variability of the y's (SST) can be broken into two parts, $SSR+SSE$.

• The first part is the regression sum of squares, $SSR=\sum(\hat{y}_i-\bar{y})^2$. This is simply $(n-1)$ times the sample variance of the fitted values $\{\hat{y}_i\}$.

Since $\hat{y}_i=b_0+b_1x_i$, the fluctuation in the $\{\hat{y}_i\}$ values is completely "explained" by the fluctuation in the $\{x_i\}$ values.

Thus, SSR is the part of the variability of y which can be "explained" by x (or, more precisely, by the *regression* of y on x).

- The second part is the residual sum of squares, $SSE = \sum (y_i - \hat{y}_i)^2$.

The residuals $e_i = y_i - \hat{y}_i$ represent the part of y which remains after we try to "explain" y based on x.

(This is clear, since $y_i = \hat{y}_i + e_i$.)

Thus, *SSE* represents the part of the variability of y which is "unexplained" by x.

This unexplained variability is attributed to chance, or to other factors not yet considered.

Now, we define the

coefficient of determination by

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{SSR}{SST} \ .$$

- We see that R^2 measures the proportion of the variability of y that is "explained" by x.

An equivalent definition is

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \ .$$

• It can be shown that $0 \le R^2 \le 1$.

We will get $R^2=1$ if, and only if, all points lie exactly on a straight (non-horizontal) line.

The closer R^2 is to 1, the stronger the linear relationship between x and y.

If R^2 is near zero, then almost none of the variability of y is explained by x, so the linear relationship is weak.

We will get $R^2=0$ if, and only if, $b_1 = 0$. This can happen in a variety of ways, including:

(1) All y's lie on a horizontal line;

(2) The data points lie on a parabola $y = a + bx^2$, which peaks in the middle of the range of the equally-spaced x's.

• Note that in (2), there is a clear nonlinear relationship but no linear relationship whatsoever! So keep in mind that R^2 only measures the strength of the *linear* relationship.

If R^2 is large, we say that x and y are "highly correlated". In this case, there is a strong linear relationship between x and y.

If R^2 is near zero, we say that x and y are nearly "uncorrelated". In this case,

the linear relationship is weak.

Note that R^2 is the square of the **Pearson correlation coefficient,** $r = s_{xy}/(s_x s_y)$.

Note that a high correlation should not be taken as evidence of a causal relationship. Consider the Nielsen and People Meter ratings. Another example is #TV sets (x) and #Cars (y) in a household. If you want more cars, can you get them by buying more TV sets?

Eg: For the Stock Market example, the Minitab output shows that R^2=.015. Only 1.5% of the variability in Today's returns is "explained" by

Yesterday's returns. The linear relationship is so weak, that it is essentially nonexistent.

For the TV example, R^2=.9490. This indicates a strong linear relationship, with 95% of the variability in the People Meter ratings being "explained" by the Nielsen ratings. But we still need the quotes around the word "explained". In fact, the ultimate explanation of a high People Meter rating is presumably that the show was popular. This is presumably what causes the Nielsen ratings to also be high. So it's not that high Nielsen ratings are the *cause* (or the explanation, in any meaningful sense) of high People Meter ratings.