Simple Regression and Correlation

Today, we are going to discuss a powerful statistical technique for examining whether or not two variables are related. Specifically, we are going to talk about the ideas of *simple regression* and *correlation*.

One reason why regression is powerful is that we can use it to demonstrate causality; that is, we can show that an independent variable causes a change in a dependent variable.

Scattergrams

The simplest thing we can do with two variables that we believe are related is to draw a *scattergram*. A scattergram is a simple graph that plots values of our dependent variable Y and independent variable X.

Normally we plot our dependent variable on the vertical axis and the independent variable on the horizontal axis.

Shilling for the incontinence industry

For example, let's make a scattergram of the following data set:

Individual	No. of Sodas Consumed	No. of Bathroom Trips
Rick	1	2
Janice	2	1
Paul	3	3
Susan	3	4
Cindy	4	6
John	5	5
Donald	6	5

Eyeballing a regression line

Just from our scattergram, we can sometimes get a fairly good idea of the relationship between our variables. In our current scattergram, it looks like a line that slopes up to the right would "fit" the data pretty well.

Essentially, that's all of the math we have to do: figure out the best fit line, i.e. the line that represents an "average" of our data points.

Note that sometimes our data won't be linearly related at all; sometimes, there may be a "curvilinear" or other nonlinear relationship. If it looks like the data are related, but the regression doesn't "fit" well, chances are this is the case.

Simple linear regression

While our scattergram gives us a fairly good idea of the relationship between the variables, and even some idea of how the regression line should look, we need to do the math to figure out exactly where it goes.

To figure it out, first we need an idea of the general equation for a line. From algebra, any straight line can be described as:

$Y = a + bX$, where a is the intercept and b is the slope

Figuring out a and b

The problem of regression in a nutshell is to figure out what values of a and b to use. To do that, we use the following two formulas:

$$b = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

Again, this looks ugly, but it's all the same simple math we already know and love: just use PEMA, and we will get the right answer.

Solving our example

So, let's revisit our example data and figure out the slope and intercept for the regression line.

Individual	No. of Sodas Consumed	No. of Bathroom Trips
Rick	1	2
Janice	2	1
Paul	3	3
Susan	3	4
Cindy	4	6
John	5	5
Donald	6	5

Solving our example (cont'd)

$$b = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{104 - \frac{(24)(26)}{7}}{\sum 100 - \frac{(24)^2}{7}}$$
$$= \frac{104 - (624/7)}{100 - (576/7)} = \frac{104 - 89.142}{100 - 82.285} = \frac{14.86}{17.71} = 0.8387$$

Now that we have calculated b, calculating a is pretty simple; we just solve $a = (26/7) - 0.8387(24/7) = 3.7142 - (0.8387)(3.4285) = 3.7142 - 2.8754 = 0.8388$.

Pearson's r

Now that we've found a and b, we know the intercept and slope of the regression line, and it appears that X and Y are related in some way. But how strong is that relationship?

That's where Pearson's r comes in. Pearson's r is a measure of correlation; sometimes, we just call it the correlation coefficient. r tells us how strong the relationship between X and Y is.

Calculating Pearson's r

The formula for Pearson's r is somewhat similar to the formula for the slope (b). It is as follows:

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\sum X^2 - (\sum X)^2/n}\sqrt{\sum Y^2 - (\sum Y)^2/n}}$$

We already know the numerator from calculating the slope earlier, so the only hard part is the denominator, where we have to calculate each square root separately, then multiply them together.

Solving for our example

So, from our example:

$$
\begin{aligned}
r &= \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\sum X^2 - (\sum X)^2/n}\sqrt{\sum Y^2 - (\sum Y)^2/n}} \\
&= \frac{104 - \frac{(24)(26)}{7}}{\sqrt{100 - (24)^2/7}\sqrt{104 - (26)^2/7}} \\
&= \frac{14.8572}{\sqrt{17.7143}19.4286} = \frac{14.8572}{(4.2088)(4.4077)} = \frac{14.8572}{18.551} = 0.8008
\end{aligned}
$$

Correlations and determinations

A correlation coefficient of around .8 indicates that the two variables are fairly highly associated. If we square r, we get the *coefficient of determination* r^2, which tells us how much of the variance in Y is explained by X. In this case, $r^2 = .6412$, which means that we estimate that 64% of the variance is explained by X, while the remainder is due to error.

The only other thing we might want to do is find out if the correlation is statistically significant. Or, to put it in terms of a null hypothesis, we want to test whether $H_0 : r = 0$ is true.

Significance testing for t

To test whether or not r is significantly different from zero, we use the t test for Pearson's r:

$$t_{\text{ob}} = r\sqrt{\frac{n-2}{1-r^2}}$$

Since this is like any other hypothesis test, we want to compare t_{ob} with t_{crit}. For this test, we use our given alpha level (conventionally, .05 or .01) and df $= n - 2$. In this case, we subtract 2 from our sample size because we have two variables.

So, with $\alpha = .05$, is the correlation significant?

Example significance test

$$t_{\text{ob}} = r\sqrt{\frac{n-2}{1-r^2}} = .8008\sqrt{\frac{7-2}{1-.6412}} = .8008\sqrt{\frac{5}{.3588}}$$
$$= .8008\sqrt{13.9353} = (.8008)(3.733) = 2.9893$$

Now, like in other significance tests, we find our critical value of t from the table $(\alpha = .05, \text{df} = 5 : 2.571)$ and compare it to the obtained value. Since $2.571 \leq 2.9893$, we reject the null hypothesis and conclude that the correlation is statistically significant.

Homework

Most of the time when you see regressions, they'll be more complex than this example. Rather than testing the significance of r, with multiple explanatory variables we test the significance of each independent variable, but the principle is exactly the same. We'll talk more about this on next Monday.

There is no class on Wednesday. You can spend the time working on your papers, studying for the final, or frolicking in the sunlight.

If you want to practice this material, read through the second example and do question 2 on page 180.