Stat 400, section 12.5 Correlation

notes by Tim Pilachowski

First of all, a look back at Lecture 5.2.

The variance of a single random variable *X* gives an indication of how the values vary in relationship to the mean. Given two random variables *X* and *Y*, we'll be interested in how the two vary in relationship to each other. The *covariance* between two random variables is

$$\text{discrete} \qquad \operatorname{Cov}(X, Y) = \sum_{\text{all } x} \sum_{\text{all } y} (x - \mu_X) * (y - \mu_Y) * p(x, y)$$

$$\text{continuous} \qquad \operatorname{Cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X) * (y - \mu_Y) * f(x, y)\, dx\, dy$$

However, as with single random variables, these calculations can become quite onerous. If we were to multiply $(x - \mu_X) * (y - \mu_X)$, then find expected value for each term separately before recombining, we'd get a shortcut:

$$\operatorname{Cov}(X, Y) = E(XY) - \mu_X * \mu_Y .$$

Covariance of two random variables has a basic problem that it shares with variance of one random variable: the value alone doesn't tell us much. Given a random variable *X* with variance $\sigma_X^2 = 10$, and a linear transformation $Y = 5X$, then $\sigma_Y^2 = 5^2 * \sigma_X^2 = 250$. In other words, the size of the values of *X* has a direct effect on the values of *E*(*X*) and *V*(*X*).

For one random variable, we found standard deviation, then used $Z = \dfrac{X - \mu_X}{\sigma_X}$ to standardize. We'll do something similar for two random variables considered jointly. The *correlation coefficient* of two random variables *X* and *Y* is defined as $\operatorname{Corr}(X, Y) = \rho_{X,Y} = \rho = \dfrac{\operatorname{Cov}(X, Y)}{\sigma_X * \sigma_Y}$.

It is extremely important that we note here that the formulas given above were applied to *populations*.

Now that we are considering *samples*, we consider again the question first posed in chapter 6. Since we don't usually know the actual value of a population parameter (in this case population correlation ρ), can we find a way to take sample data and calculate a point estimate?

The answer is "Yes". Our point estimate of the population correlation coefficient ρ is the sample correlation coefficient *r*. The sample correlation coefficient *r* is a measure of the strength of the relationship between the x_i and y_i values in a sample.

Given *n* numerical pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, it is natural to speak of *x* and *y* as having a positive relationship if large *x*'s are paired with large *y*'s and small *x*'s with small *y*'s. Similarly, if large *x*'s are paired with small *y*'s and small *x*'s with large *y*'s, then a negative relationship between the variables is implied.

Consider the quantity $S_{xy} = \sum (x - \bar{x})(y = \bar{y}) = \sum xy - \dfrac{\sum x * \sum y}{n}$. If the relationship is strongly positive, an x_i above the mean $\bar{x}$ will tend to be paired with a y_i above the mean $\bar{y}$. As a consequence, we would expect $(x - \bar{x})(y - \bar{y}) = (+)(+) > 0$. This product will also be positive whenever both x_i and y_i are below their respective means: $(x - \bar{x})(y - \bar{y}) = (-)(-) > 0$. In other words, a positive relationship between x_i and y_i values in a sample implies that S_{xy} will be positive.

An analogous argument shows that when the relationship is negative, S_{xy} will be negative, since most of the products $(x - \bar{x})(y - \bar{y})$ will be negative.

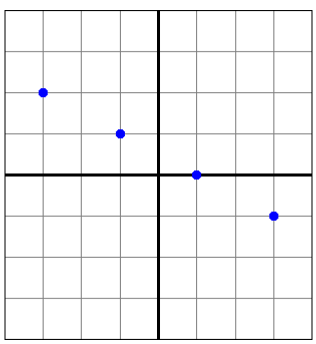
Unfortunately, S_{xy} has the same defect as covariance did in chapter 5: the value alone doesn't tell us much. By changing the unit of measurement for either *x* or *y*, S_{xy} can be made either arbitrarily large in magnitude or arbitrarily close to zero. So, just as we did to find population correlation coefficient in chapter 5, we use a denominator to "standardize" so that the calculated value will not depend on the particular units used to measure *x* and *y*.

Definition: The **sample correlation coefficient** for the *n* pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ is given by

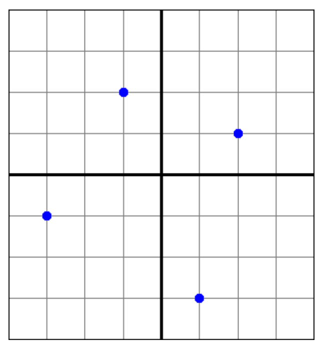
$$\hat{\rho} = r = \frac{\sum (x-\bar{x})(y=\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}} = \frac{\sum xy - \frac{\sum x * \sum y}{n}}{\sqrt{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}\sqrt{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}.$$

Non-technical explanation: The numerator expresses the relationship between *x* and *y* when they are considered as a system. The denominator expresses the relationship between *x* and *y* when they are considered separately. If the "system" relationship is a large portion/fraction of the "separate" relationship, we'll conclude that the "system" relationship is very strong. If the "system" relationship is a small portion/fraction of the "separate" relationship, we'll conclude that the "system" relationship is very weak.

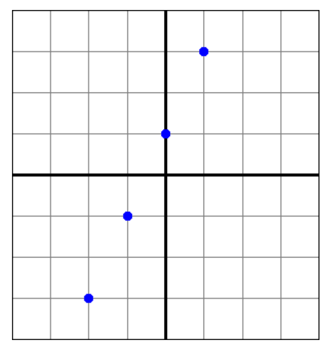
Specifically, if *x* and *y* have a very strong relationship, the value of *r* will be close to 1 or –1. If *x* and *y* have a very weak relationship, the value of *r* will be close to 0.



$r = -1$



$r = 0$



$r = 1$

The points in the first and third scatterplots all fall exactly on the estimated regression line. In this case, all of the relationship/connection/correlation between *x* and *y* can be attributed to their existing in a system. In the middle scatterplot, none of the relationship/connection/correlation between *x* and *y* can be attributed to their existing in a system. [An analogy for "zero correlation" from chapter 5 is the concept of "independence".]

Of course, in actuality, we'd almost never encounter any of these extremes using sample data.

Another important property of the sample correlation coefficient *r*: The value of *r* does not depend on which of the two variables under study is labeled *x* and which is labeled *y*. This is very different from regression analysis, where virtually all quantities of interest ($\hat{\beta}_1$, $\hat{\beta}_0$, s^2 ,etc.) depend on which of the two variables is considered the independent variable *X* and which is treated as the dependent variable *Y*.

One more note: The square of the sample correlation coefficient *r* equals the value of the coefficient of determination that would result from fitting the simple linear regression model. Symbolically, $(r)^2 = r^2$.

The sign of *r* indicates direction of the correlation; the magnitude of *r* indicates strength of the correlation. This is the text's rule of thumb: $|r| < 0.5$ weak, $0.5 < |r| < 0.8$ moderate, $|r| > 0.8$ strong correlation.

IMPORTANT: Correlation is not the same thing as causation. No matter how strong the correlation may be, we still cannot say that "*X* causes *Y*" or "*Y* causes *X*". While it may be so, it might also be true that "both *X* and *Y* are the result of some other unknown factor(s)". Consider unemployment vs. inflation, or weight vs. height.

(Go to Example A-1 below.)

Two things to note about the interpretation of sample correlation coefficient r.

1) The strength or weakness of the correlation between sample x and sample y values does not tell us about the value of the linear regression slope estimate $\hat{\beta}_1$. The value of r tells us how closely (or not) the dots in the scatterplot are aligned, not how steep (or shallow) the alignment is.

2) A sample correlation value r might indicate a strong relationship, but we still need to determine whether the correlation is statistically significant. (This will depend in part on both value of r and sample size n).

The small-sample intervals and test procedures presented in Chapters 7–9 were based on an assumption of population normality. To test hypotheses about r, an analogous assumption about the distribution of pairs of (x, y) values in the population is required. We are now assuming that *both* X and Y are random variables, whereas much of our regression work focused on x fixed by the experimenter.

The 8th edition of the text includes an explanation and graphic of an assumed normal bivariate distribution of X and Y. The 9th edition simply refers back to section 5.2. Here's the important idea: $\rho = 0$ implies X and Y are independent.

Proposition: When $H_0 : \rho = 0$ is true, the test statistic $T = \dfrac{R\sqrt{n-2}}{\sqrt{1-R^2}}$ has a t distribution with $n - 2$ degrees of freedom.

Because ρ measures the extent to which there is a linear relationship between the two variables in the population, the null hypothesis $H_0 : \rho = 0$ states that there is no such population relationship.

In Section 12.3, we used the t ratio $T = \dfrac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$ to test for a linear relationship between the two variables in the context of regression analysis. The test procedures of 12.5 and 12.3 are completely equivalent. With a lot of algebraic manipulation, we could show that $\dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$.

When interest lies only in assessing the strength of any linear relationship rather than in fitting a model and using it to estimate or predict, the 12.5 test statistic formula just presented requires fewer computations than does the t-ratio of 12.3.

The test of 12.5 can be useful to a researcher who wants to verify the significance of a correlation as a means of deciding whether development of the linear regression equation is worthwhile.

IMPORTANT: The sample correlation coefficient r and its test of significance will only tell us about a *linear* relationship between random variables X and Y. Other procedures and tests would be needed to investigate quadratic, exponential or logarithmic relationships.

IMPORTANT: Correlation is not the same thing as causation. No matter how strong the correlation may be, we still cannot say that "X causes Y" or "Y causes X". While it may be so, it might also be true that "both X and Y are the result of some other unknown factor(s)". Consider unemployment vs. inflation, or weight vs. height.

(Go to Example A-2 below.)

Example A: A paper in *Measurement and Evaluation in Counseling and Development* (Oct 90, pp. 121–127) discussed a survey instrument called the *Mathematics Anxiety Scale for Children* (MASC). Suppose the MASC was administered to ten fifth graders with the following results:

MASC Score	67	37	70	40	35	65	40	35	30	40
Math grade (%)	75	85	60	90	80	75	70	90	95	80

From Lectures 12.1 and 12.2, we have $S_{xy} \approx -1075$, $S_{xx} \approx 2064.9$, $S_{yy} \approx 1000$.

1. Calculate the sample correlation coefficient *r* and interpret its value in the context of anxiety score vs. Math grade.

2. Test the statistical significance of the sample correlation coefficient *r* calculated in 1) above.