PLSC 502: "Statistical Methods for Political Research"

Measures of Association: Interval/Ratio-Level Variables
November 8, 2016

Relationships between Interval/Ratio-Level Variates

We'll spend the day discussing relationships between interval- and/or ratio-level variates.

Linear Relationships

The simplest form of a monotonic relationship between Y and X is a *linear* relationship. We can think of the relationship as one akin to

$$Y = mX + b \tag{1}$$

where (just like in geometry class) m is the "slope" of the line and b is the "intercept." The reason we characterize this as the "simplest" form of relationship is because, for a linear relationship,

$$\frac{\partial Y}{\partial X} = m;$$

that is, the change in Y associated with a one-unit change in X (that is, the slope of the function) is just m, a constant. This is true irrespective of the "location" at which the change in X takes place.

Nonlinearity

Of course, linearity is only one form of a relationship that interval/ratio-level variates might have. Two others are in Figures 1 (logarithmic) and 2 (exponential) in the slides. Note that:

- In a *logarithmic* relationship, we observe *diminishing* returns to Y in X. That is, the change in Y associated with a one-unit change in X is decreasing in X. Formally, this implies that, irrespective of $\frac{\partial Y}{\partial X}$,

$$\frac{\partial^2 Y}{\partial X \partial X} < 0.$$

- In an *exponential* relationship, we observe *increasing* returns to Y in X. That is, the change in Y associated with a one-unit change in X is increasing in X. Formally, this implies that

$$\frac{\partial^2 Y}{\partial X \partial X} > 0.$$

One can also imagine:

- Curvilinear relationships with more "bends" (a la polynomials),
- "Step-functions,"

- "Threshold" effects, and/or
- combinations of these.

All of which counsels that it's always good to "look" at your data.

Pearson's r

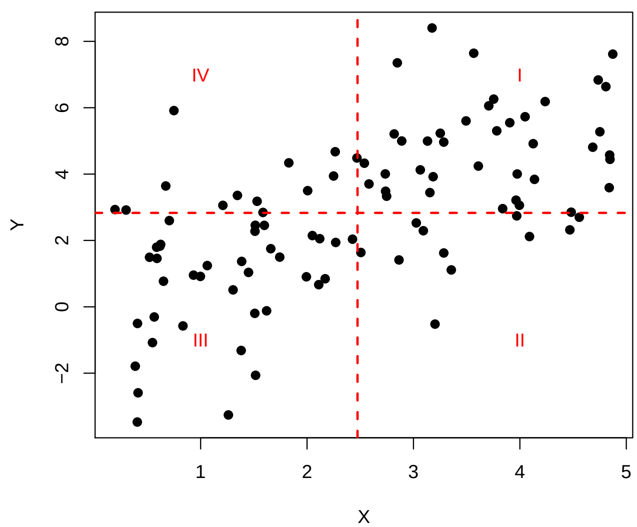
Pearson's product-moment correlation (much better known as *Pearson's r*) is the workhorse of bivariate association measures between two continuous variates. It is a summary measure of the direction and strength of the linear association between two variables. Formally,

$$r = \frac{\sum_{i=1}^{N} \left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)}{N-1} \tag{2}$$

where (as before) s_X and s_Y are the sample standard deviations of X and Y, respectively.

The intuition of Pearson's r is illustrated in Figure 1. The red dashed lines indicate the means of Y and X. Observations in quadrant I have both $X_i - \bar{X}$ and $Y_i - \bar{Y} > 0$; observations in quadrant II have $X_i - \bar{X} > 0$ and $Y_i - \bar{Y} < 0$, and so forth. The product of these two terms "signs" each observation's product of deviations-from-means; summing across observations means that relatively large numbers of observations in quadrants I and III will yield positive values of r, while larger numbers in quadrants II and IV will yield negative values.

Figure 1: Intuition: Pearson's r



Characteristics of r are:

- $r \in [-1, 1]$

- $r = 0 \leftrightarrow$ no association between Y and X.
- The sign of r indicates direction of the (*linear*) relationship; while
- The magnitude of $|r|$ indicates the strength of the (again, *linear*) relationship.
- These are illustrated graphically in the slides.

Note that the fact that r measures linear association means that:

- It cannot tell you anything about nonlinear relationships in the data (as in the figure in the slides).
- It can also be unduly influenced by *outliers* (which one might think of as a form of nonlinearity, depending on the circumstances) (Figure 5).
- Finally, note that the "slope" of the linear relationship has little or no bearing on r; two "perfect" positive, linear relationships between Y_1 and X_1 and Y_2 and X_2, each with different values of m in (1), will nonetheless both have $r = 1.0$. In other words, Pearson's r measures the degree of clustering around a line characterizing the relationship between Y and X, but not the *slope* of that line.

Sampling Distribution of r

The sampling distribution is a bit complicated, since r is necessarily bounded between -1 and 1. In particular, if the population value of r is very high or very low (that is, if $|r| \approx 1.0$), the sampling distribution is *skewed.*

Fisher (yes, *that* Fisher) showed that, even when the sampling distribution of the estimator $\hat{r}$ is skewed,

$$\hat{w} = \frac{1}{2} \ln \left(\frac{1 + \hat{r}}{1 - \hat{r}} \right) \tag{3}$$

is approximately $\mathcal{N}$ormally distributed with a mean of $\frac{1}{2} \ln \left(\frac{1+\hat{r}}{1-\hat{r}} \right)$ and a standard error of $\frac{1}{\sqrt{N-3}}$. This transformation is illustrated in Figure 2; you can see that it looks like a sideways "S-curve," with vertical asymptotes at -1.0 and 1.0. Thus,

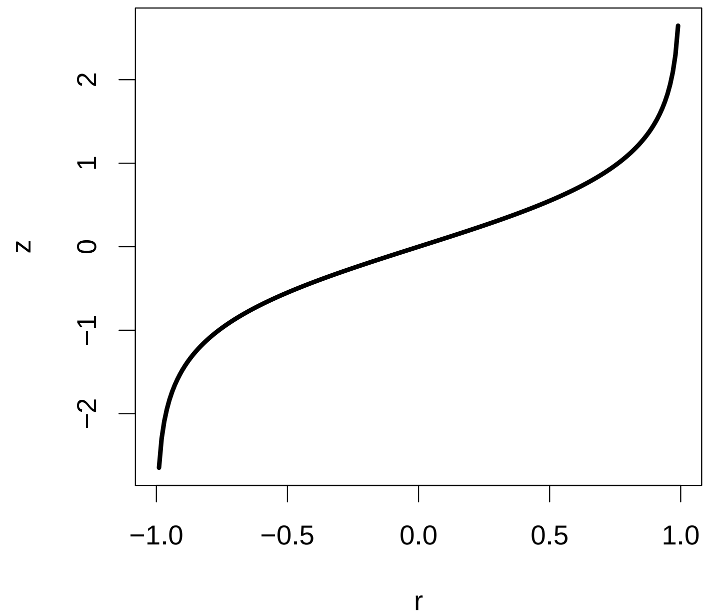
$$z_r = \frac{\frac{1}{2} \ln \left(\frac{1+\hat{r}}{1-\hat{r}} \right) - \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)}{\sqrt{\frac{1}{N-3}}} \sim \mathcal{N}(0, 1)$$

We can also use w to construct confidence intervals around $\hat{r}$, in the usual fashion. An alternative (and more exact) form of the sampling distribution of r – and one that is the default in the `cor.test` routine in R and `pwcorr` in Stata – is:

$$\frac{\hat{r}\sqrt{N-2}}{\sqrt{1-\hat{r}^2}} \sim t_{N-2}. \tag{4}$$

Most software packages will calculate p-values for the usual null hypothesis $r = 0$ automatically.

Figure 2: Fisher's z Transformation for Pearson's r



An Alternative to r: Spearman's ρ

An alternative to Pearson's r is the rank-based test known as *Spearman's* ρ.

Imagine sorting the data on both Y and X, and on the basis of their position on each variable, assigning them a *rank* on each, denoted R_{Y_i} and R_{X_i}, respectively. Thus, the observation in the data with the highest value of Y would be $R_{Y_i} = 1$, the next-highest would be $R_{Y_i} = 2$, and so forth, with a similar procedure for R_{X_i}. Spearman's ρ is then equal to:

$$\rho = 1 - \frac{6 \sum_{i=1}^{N} D_i^2}{N(N^2 - 1)} \tag{5}$$

where D_i is the difference in ranks of observation i between Y and X (that is, $R_{Y_i} - R_{X_i}$).

Characteristics of Spearman's ρ:

- $\rho \in [-1, 1]$
- It has the same interpretation as r.
- ρ is also appropriate for use with ordinal data, since they can also be used to rank observations. However,
- When many "ties" occur, a better alternative is to calculate Pearson's r on the ranks R_{Y_i} and R_{X_i}, and assign "partial" (or "half") ranks to tied individuals.

Summary: Some Measures of Association

		X			
		Nominal	Binary	Ordinal	Interval/Ratio
Y	Nominal	χ^2	χ^2	χ^2	t-test (and η)
	Binary	χ^2	ϕ, Q	γ, τ_c	t-test
	Ordinal	χ^2	γ, τ_c	γ, τ_a, τ_b	Spearman's ρ
	Interval / Ratio	t-test (and η)	t-test	Spearman's ρ	r

Example: Back to Africa

See the slides for some simple examples of how to estimate r (and rho) using R ...