BIVARIATE ANALYSIS: RELATIONSHIPS BETWEEN VARIABLES AND MEASURES OF ASSOCIATION

Think back to Problem Set #3A on Identifying Variables. You will recall that in each sentence you were asked to identify *two* variables pertaining to the same unit of analysis. This is because each sentence claims or implies that there is a *relationship* or *association* between these two variables — for example, a case that has a "high" value on one variable is likely to have a "high" value on the other variable also and likewise that a case that has a "low" value on one variable is likely to have a "low" value on the other variable also. Many of the sentences also claim or imply that this association exists because there is a *cause and effect relationship* between the two variables, e.g., that having a "high" (or "low") value on one variable *causes* a case to have a "high" (or "low") value on the other variable. Thus, Problem Set #3A anticipated that we would move beyond *univariate analysis* to *bivariate analysis*.

Direction of Association Between Variables

The generic example of association given in the previous paragraph is more specifically an example of a *positive* association between two variables — that is, "high" goes with "high" and "low" goes with "low." For example, sentences #1 and #7 in PS #3A claim that there is a positive relationship between two variables, which we can diagram in this way:

#1 LEVEL OF SENIORITY (Low to High) ___+___ ======== LEVEL OF PRAGMATISM (Low to High) [*members of Congress*]

#7 DEGREE OF CLOSENESS (Low to High) ___+___ ========= LEVEL OF TURNOUT (Low to High) [*elections*]

However, the *direction of association* can be negative as well as positive — that is, "high" may go with "low"and "low" may go with "high." For example, sentence #2 in PS #3 claims that there is a *negative* relationship between two variables, which we can diagram in this way:

#2 LEVEL OF EDUCATION (Low to High) ___–___ ======== DEGREE OF RELIGIOSITY (Low to High) [*individuals*]

A negative relationship is sometimes called an *inverse* relationship, as in the *<u>Inverse</u> Square Root Law* of random sampling (Handout #2):

SIZE (Low to High) ___–___ ========== MARGIN OF ERROR (Low to High) [*samples*]

Note that all the variables in the above associations have values that run from "Low" to "High" and we made use of this fact in defining positive vs. negative association. All quantitative (interval or ratio) variable have (numerical) values that can run from "Low" (e.g., 0% TURNOUT) to "High" (e.g., 100% TURNOUT) and, as we have seen, many ordinal variables also have values

that run from "Low" to "High." However, some ordinal variables have values that, while they run in a natural order, do not run from "Low" to "High," but rather from "Liberal" to "Conservative" (IDEOLOGY variables), from "Strongly Agree" to "Strongly Disagree" (many ISSUE OPINION variables), from "Never" to "Always" (ABORTION OPINION), and so forth. Moreover, nominal variables by definition have values that do not "run" in any natural ordering at all.

In general, when both of two variables have "matching values" with a *common natural ordering*, an association between them can be characterized as positive or negative. Here are some additional examples involving variables that have matching values but are not of the "Low to High" type (and in which we probably expect positive associations):

PARENTS' PARTY ID (Dem. to Rep.) ==========+========== CHILD'S PARTY ID (Dem. to Rep.) [*parent-child pairs*]

PRESIDENT'S ECONOMIC POLICY (Strongly Approve to Strongly Disapprove) ====+==== PRESIDENT'S FOREIGN POLICY (Strongly Approve to Strongly Disapprove) [*inds*]

On the other hand, pairs of ordinal and nominal variables may have values that, from a logical (as opposed to empirical) point of view, do not have "matching values." Here are some examples:

#14 DIRECTION OF IDEOLOGY (Liberal to Conservative) ======== DIRECTION OF VOTE (Dem. vs. Rep.) [*individuals*]

RELIGIOUS AFFILIATION (Protestant vs. Catholic) ======== PRESIDENTIAL VOTE (Dem. vs. Rep.) [*individuals*]

We certainly would expect IDEOLOGY and VOTE to be quite closely associated but we will have to spell out the direction verbally (e.g., as in sentence #14: "liberals generally vote Democratic and conservatives generally vote Republican"), rather than relying on the semi-mathematical shorthand of "positive vs. negative." Likewise, in the first half of the twentieth century (and outside of the South), there was a clear association between RELIGION and VOTE in the U.S., but again the direction of association has to spelled out in words, namely that Protestants tended to vote Republican and Catholics tended to vote Democratic. In neither of these examples can we meaningfully specify the direction of the relationship in terms of the positive vs. negative shortcut, because the variables have non-matching values; rather we need to spell out its direction verbally.

Note 1. Given a dichotomous variable with "yes" and "no" values, "no" is conventionally considered to be "low" and "yes" to be "high." Accepting this conventiona, we can meaningfully say (for example) that LEVEL OF POLITICAL INTEREST is *positively* associated with WHETHER OR NOT VOTED IN ELECTION.

Note 2. One may be able to rename and recode variables so that they have values that run from low to high, with the result that a relationship that previously could not be summarized in positive vs. negative terms now can be so characterized.

DIRECTION OF IDEOLOGY (Liberal to Conservative) ========= LEVEL OF SUPPORT FOR U.N. [*individuals*] (Low to High)

Given that these two variables don't have matching values, we can't characterize the direction of the association between them as positive or negative — we will have to spell it out verbally and say (for example) that liberals tend to more supportive of the UN and conservatives less supportive. However, we can rename and recode the first variable in a way that allows us to summarize the association in positive vs. negative terms.

DEGREE OF LIBERALISM (Low to High) + ========== LEVEL OF SUPPORT FOR U.N. [*individuals*] (Low to High)

Finally note that reversing the "polarity" of either *one* of the two variable reverses the (positive or negative) sign of the association, while reversing *both* preserves it..

DEGREE OF LIBERALISM (Low to High) – ========== LEVEL OF OPPOSITION TO U.N. [*individuals*] (Low to High)

DEGREE OF CONSERVATISM (Low to High) + ======= LEVEL OF OPPOSITION TO U.N. [*individuals*] (Low to High)

DEGREE OF CONSERVATISM (Low to High) – ======= LEVEL OF SUPPORT FOR U.N. [*individuals*] (Low to High)

You should be able to see that all four of the diagrams above make the same substantive empirical claim.

Strength of Association Between Variables

Beyond the question of the direction of an association (if any) between two variables, there is the question of the *strength* of the association between them. If almost all liberals vote Democratic and almost all conservatives vote Republican, there is a *strong association* between IDEOLOGY and VOTE. But if liberals vote Democratic only slightly more than conservatives do, and conservatives vote Republican only slightly more than liberals do, there is only a *weak association* between IDEOLOGY and VOTE.

A great number of different of *bivariate summary statistics* called *measures of association* are used in quantitative research, each with somewhat different properties. Different measures of association are appropriate depending on whether the variables are nominal, ordinal, or interval. Many are defined and discussed in Weisberg, Chapter 12; however, you are asked only to skim this chapter and you are *not* responsible for knowing the different types of measures of association, let alone how to calculate them. We will later study one measure of association that you will be responsible for — the *correlation coefficient* that is used to measure association between two *interval* variables. But in the meantime, you should understand two general properties of all measures of association. Let's use the symbol *a* to designate a generic measure of association.

(a) Every measures of association *a* is *standardized* — that is, every such measure takes on values between 0 and 1. If bivariate analysis shows that there is *no relationship* or *association* between two variables (e.g., if liberals vote Democratic and Republican in the same proportions as conservatives do), then $a = 0$. If it shows that there is *perfect relationship* or *association* between two variables (e.g., if every liberal votes Democratic and every conservative votes Republican), then $a = 1$. As the strength of association ranges between these (empirically unlikely) extremes, the value of *a* ranges between 0 and 1.

(b) If the variables have matching values such that an association can be characterized as *positive* or *negative*, a measure of association carries the appropriate (+ or −) sign. In this event, a measure of associate takes on values that extend from −1 through 0 to +1. If every Democratic parent has Democratic children and every Republican parent has Republican children (the "pure socialization hypothesis"), then $a = +1$; if Democratic and Republican parents are equally likely to have children of either partisanship (the "no impact hypothesis"), then $a = 0$; and if every Democratic parent produces Republican children and every Republican parent produces Democratic children (the "pure rebellion hypothesis"), then $a = -1$.

Independent vs. Dependent Variables

Association between variables is *symmetric* — if PARENTS' PARTY ID is associated with CHILD'S PARTY ID, then it is equally true that CHILD'S PARTY ID is associated with PARENTS' PARTY ID; and if the association between variables *X* and *Y* is $a = +0.7$, then the association between *Y* and *X* is also $a = +0.7$.

However, if there is an association between variables *X* and *Y*, it *may* (but also may not) be due to the fact that *X influences* (or has *causal impact)* on *Y* or that *Y* influences *X*. Such a *cause and effect* relationship clearly is *not symmetric* — saying that variable *X* influences variable *Y* is different from saying that *Y* influences *X*. Consider sentence #6 in PS #3A, which says that "hard studying makes for good grades." In terms of the points made here, this sentence is saying *three* distinct things.

(i) There is a *relationship* or *association* between DEGREE OF STUDYING and LEVEL OF GRADES, i.e., $a \neq 0$.

(ii) The relationship or association between the two variables is *positive*, i.e., $a > 0$.

(iii) This positive association exists (in part, at least) because DEGREE OF STUDYING has (positive) *causal impact* on ("makes for") or *influences* LEVEL OF GRADES.

The variable that is the (hypothesized) cause is called the *independent variable*, and the variable that is the (hypothesized) effect is called the *dependent variable*. In diagrams such as we have been using, it is conventional to draw an arrow from the independent to the dependent variable in this manner:

DEGREE OF STUDYING + LEVEL OF GRADES [*students*]

==========>

(Low to High) (Low to high)

In this context (but not in some others), the independent variable is conventionally put on the left side and the dependent variable on the right, so the arrow points to the right.

Here are several ways of characterizing independent and dependent variables:

(i) The independent variable is the *cause*, the dependent variable is the *effect*.

(ii) The independent variable *influences* (or *has an impact on*) the dependent variable.

(iii) We want to explain why cases have particular values on the dependent variable; we will use their values on the independent variable as (part of) the *explanation*.

Some of the sentences in PS #3A (and PS #9) contain words and phrases that clearly indicate the direction of (claimed) causality (and whether the causal effect is positive or negative), for example:

LEVEL OF EDUCATION *undermines* DEGREE OF RELIGIOSITY

DEGREE OF STUDYING *makes for* LEVEL OF GRADES

In other sentences, the direction of causality is (at best) implicit only, e.g., #1 and #14.

In some contexts, we may have good reason to study the association between two variables even if we don't regard one as independent and the other as dependent. For example, we would expect the following to be true:

SCORE ON PART I OF TEST (Low to High) + ======== SCORE ON PART II OF TEST (Low to High) [*students*]

(In fact, on POLI 300 tests, the correlation (association) between these two variables is typically very high but not perfect, e.g., +0.9) However (except for the fact that there is a temporal sequence to the two tests), it doesn't make much sense to regard one of these variables as independent and the other dependent.

Finally, it should be emphasized that variables are not *intrinsically* independent or dependent, but rather they assume one or other role in different hypotheses, theories, or research projects. For example, once we have developed a concept of PARTY IDENTIFICATION, have figured out how to measure it, have collected appropriate data, and have completed basic univariate analysis, we may turn to further *bivariate* research questions pertaining to PARTY ID.

One set of such questions concerns the *causes of* or, or *influences on*, or *explanations for* PARTY ID. Why do some people think of themselves as Democrats and other as Republicans? Is PARTY ID influenced by PARENTS' PARTY ID, LEVEL OF EDUCATION, LEVEL OF INCOME, RELIGION AFFILIATION, IDEOLOGY, etc? Here we are treating PARTY ID as the *dependent* variable and are looking for *independent* variables that may affect the direction and strength of PARTY ID.

Another set of questions concerns the *consequences* or *effects* of PARTY ID. For example, does PARTY ID affect how likely people are to turn out and vote? Does PARTY ID effect how people vote? Does PARENTS' PARTY ID affect the party identification and other political attitudes of their children? Here we are treating PARTY ID as the *independent* variable and are looking for *dependent* variables that may be affected by PARTY ID.