PLSC 502: “Statistical Methods for Political Research”

Data: Structure and Measurement
September 1, 2016

Data: Introduction

The subject du jour is *data*: what they are, how to think about them, how to collect and organize them, etc.

The vast majority of the time, data are *rectangular*, consisting of a matrix of columns (which correspond to *variables*) and rows (which correspond to *observations*). The intersection of each row and column is a *cell* of the data matrix, and contains the value that that variable takes on for that variable. A single variable is a *column vector* of the data matrix, and contains all observations' data for that variable; a single observation is a *row vector*, and contains all of a particular observation's values for each different variable.

Indices

It is conventional, when discussing data, to use subscripts as *indices*, to indicate particular observations or particular variables. The overwhelming convention in the social sciences is to use the letter i to denote individual observations (rows). Moreover, there are almost inevitably N such observations, such that

$$i \in \{1, 2, 3, ...N\}$$

Similarly, it is common (although somewhat less so) to use the letter k to denote variables, and to let K indicate the total number of such variables (columns) in the data:

$$k \in \{1, 2, 3, ...K\}$$

A typical data structure – with X denoting a variable – looks like this:

i	X_1	X_2	...	X_K
1	X_{11}	X_{21}	...	X_{K1}
2	X_{12}	X_{22}	...	X_{K2}
3	X_{13}	X_{23}	...	X_{K3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	X_{1N}	X_{2N}	...	X_{KN}

Of course we don't actually include the subscripts in the actual data values. We'll spend the day discussing each of these components – variables and observations – at length. The former are a useful entré into a discussion of levels of measurement, while the latter motivate a general discussion on data structures and organization.

Variables and Measurement

Variables are measurable quantities that – as the name suggests – vary (take on different values) across different observations. A variable that does not vary is not a variable at all – it is known as a *constant.* While constants are important, they are, as a rule, far less interesting than variables, so we won't spend any more time on them right now.

Variables are measured at different *levels of measurement*, and in fact the level at which a variable is measured is probably it's single most important trait. In addition, variables can be classified as either *continuous* or *discrete.*

Levels of Measurement

Variables can be classified according to their level of measurement, which reflects the sort of information they contain.

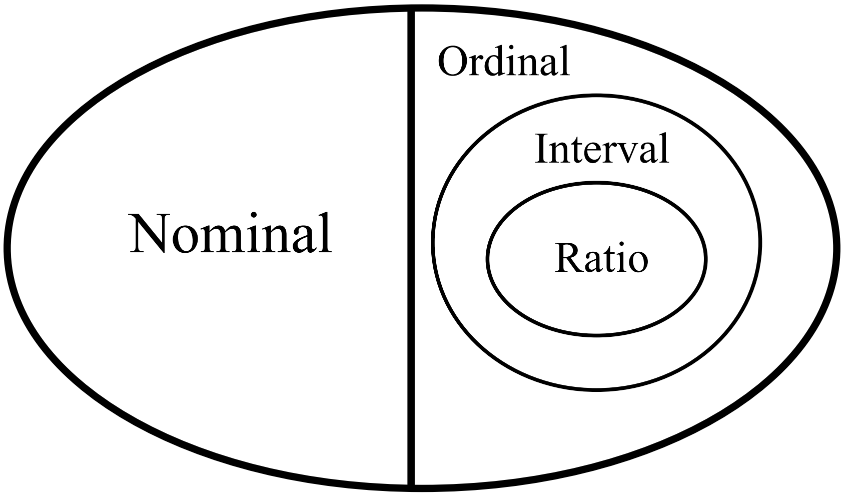
- **Nominal** level variables are variables that *classify* data into two or more distinct, mutually exclusive, exhaustive, and unordered categories. Examples are things like brands of tunafish (Starkist, Chicken of the Sea, or Giant's in-house brand), state political cultures, a la Elazar (moralistic, individualistic, and traditionalistic), the status of the international system (unipolar, bipolar, or multipolar), and so forth. Statistically speaking, nominal variables allow us to examine equality and inequality (is this the same type as that, or not?) but not to say anything about the ordering of the outcomes, or their relative "distance" from each other. In that sense, they represent the lowest level of information a variable can provide. This fact means that statistics for nominal data often look very different from those used on variables measured at higher levels.

  While we often use numerals to reflect the categories of a nominal variable, variables measured at the nominal level do not contain information on the ordering or ranking of the categories into which they classify, nor on the relative "distance" between those categories.

- **Ordinal** level variables are categorical variables that also reflect an ordering of outcomes, but do not impose a metric on that ordering. The easiest way to think of ordinal data is as a *ranking* of the outcomes, where each is higher and/or lower than each other, but where the "distance" between those outcomes is not recorded or reflected.

  As a practical matter, we use numerals to reflect the categories of an ordinal-level variable, but those numbers do not have the cardinal meaning they do in the interval and ratio cases. That is, we might create an ordinal variable where 1 = "Strongly disagree," 2 = "Disagree," 3 = "Neutral," 4 = "Agree," and 5 = "Strongly agree," but of course we wouldn't say that "Disagree" (= 2) is "half as much agreement" as "Agree" (= 4).

Figure 1: Obligatory Venn Diagram: Levels of Measurement



- **Interval** level variables are variables that both order and impose a metric on outcomes. That is, each value on the measure reflects an equal “distance” between outcomes.[1] This means that we can add and subtract values, average them, and so forth, and get meaningful quantities. (Note, however, that we can’t divide them...).

  A common example of an interval-level measure is the calendar **year**.

- **Ratio** level variables are identical to interval-level variables, except that they also include a “true zero point.” That is, a variable is measured at the ratio level if a value of “0” on that variable represents the absence of the characteristic being measured (or otherwise is a “true” zero value). Thus, while (say) temperature in degrees Celsius is an interval-level measure (because 0 degrees does not indicate the absence of temperature), temperature in degrees Kelvin is a ratio-level scale (since 0K is the absence of molecular movement – that is, energy/heat).

  As a practical matter, there is relatively little difference between interval- and ratio-level variables. The biggest one is that (as the name suggests) ratio-level variables can be used to construct ratios, while interval-level ones (generally) should not.

[1] Geometrically, interval-level measures are *affine*.

As indicated in Figure 1, levels of measurement "nest" to some extent: All ratio-level variables are also interval-level, and all interval-level variables are also ordinal (which, of course, means that ratio-level variables are ordinal as well). Note as well that all variables are (nominally) nominal, in that they classify data into mutually exclusive, exhaustive categories, but they do not "nest" into the nominal category in the same functional way.

The level at which a variable is measured is arguably the most critical aspect of a variable. It tells us something about how much information the variable contains, and – most important – it lets us know how we can combine information across observations to summarize, describe, and analyze data.

Discrete vs. Continuous Variables

- A **discrete** variable is one that can take on only a finite, or countably infinite, number of values in its range. Note that:
    - All nominal- and ordinal-level variables are discrete.
    - Similarly, some ratio-level variables are also discrete, most notably **counts** (which can take on only whole-numbered values, i.e., values in the non-negative integers).
- By contrast, a **continuous** variable can (in theory) take on any value in its range. An alternative (related) way of thinking of a continuous variable is as one that is differentiable over it's entire range – that is, it has no "gaps," "holes," "breaks," etc.[2]
    - As a practical matter, the "continuousness" of a variable is limited by the precision of the measuring instrument: a bathroom scale may only record pounds (or tenths of a pound), even though weight is infinitely divisible.

Some examples of different classes of variables, distinguished by level of measurement and type, are presented in the table below.

[2] Note that this is not to say that every value *must be* represented in the data, only that every possible value *could be* present. Data on daily high temperatures for some location, for example, are continuous even if there was no day on which the high was exactly 39.41 degrees.

Examples of Variables, by Type and Level of Measurement

Level of Measurement	Discrete	Continuous
Nominal	$\{\text{Blonde}, \text{Brunette}, \text{Redhead}\}$	n/a
Ordinal	Social Class (Upper, middle, lower)	n/a
Interval	Year	Temperature, degrees F
Ratio	Counts of things	Height, weight, distance, etc.

Dichotomous Variables

A dichotmous variable (often also referred to as a "dummy" or "binary" variable) is a variable that can take on only two values. Typically, these values are coded as zero and one, where the "natural coding" has zero indicating the absence of the trait and one the presence of the same trait.

Note a few things about dichotomous variables:

- Dummy variables occupy an ambiguous place in the levels of measurement. They are clearly nominal (as with, say, sex), but are often also ordinal (where zero indicates, say, "has never watched *Desperate Housewives*" and one indicates "has watched at least one episode of *Desperate Housewives*") and occasionally even interval or ratio (e.g., "number of first marriages").
- Any discrete variable with ℓ categories can be identically replaced by $\ell - 1$ dichotomous variables, one for each category (minus one), with no loss of information. This is often known as coding a nominal-level "as a factor."
- Occasionally, particularly in psychology and related fields, one will see a dichotomous variable coded as $Y \in \{-1, 1\}$. This is called *effect coding.*
- We'll discuss approaches for describing and analyzing dichotomous variables a bit later in the course.

Observations and Data Structure

Data are structured by their *units of analysis* (or *unit of observation*) – the "things" that we are actually observing and measuring when we collect our data. While data can take on a wide range of "shapes," the four discussed below encompass probably 99 percent of all data encountered and used in the social sciences.

Cross-Sectional Data

Data where each unit of analysis appears in the data once (i.e., as one row of data) are what I broadly term *cross-sectional* data. Common examples include single-shot opinion surveys, one-observation-per-country international data, and so forth.

An example – from a 1997 CBS/NYT survey – is given in the slides. In these data, each row represents a single survey respondent, who is measured a single time on each of the variables (here, survey questions). As indicated above, cross-sectional data are typically indexed by the letter i:

$$X_i \in X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix}$$

Time-Series Data

Time-series data are data consisting of repeated measurements on a single unit of analysis over time. Common examples are macro-level economic and political data (on things like inflation rates, GDP growth, and the like), as well as system-wide measures in international relations (things like unipolarity vs. bipolarity).

Most time-series data are measured at regular intervals – days, years, etc. – and are indexed by t, with T denoting the number of time periods observed:

$$t \in \{1, 2, 3, ...T\}$$

An example of time-series data on U.S. Supreme Court clerks is given in the slides. Time-series data can be thought of as the "opposite" of cross-sectional data, in one respect: whereas cross-sectional data have *one* observation for *each* unit of analysis, time-series data have T observations of data on *one* unit of analysis/observation.

That said, time-series data are often (usually) treated no differently than cross-sectional data. We'll get into some differences a bit later, but for now, they can be viewed (and analyzed) equivalently.

Time-Series Cross-Sectional (or "Panel") Data

Increasingly, social scientists have available data that combine cross-sectional and time-series elements; these are known as *time-series cross-sectional* ("TSCS," or – sometimes – "panel") data. Such data consist of repeated observations on more than one unit of analysis; you can think of this either as a repeated cross-section or as a "pooling" (combining) of multiple, separate time-series. Such data are typically doubly indexed, with i denoting the unit and t denoting the time point at which that unit was observed:

$$X_{it} \in X = \begin{pmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1T} \\ X_{21} \\ X_{22} \\ \vdots \\ X_{NT-1} \\ X_{NT} \end{pmatrix}$$

TSCS data thus consist of T observations for each of N units. In some cases, TSCS data are *unbalanced*; that is, some units have longer time-series than others. In such cases, we often denote the number of (temporal) observations on unit i as T_i. For example, if we were examining all countries in the international system since the end of World War II, we would have data from 1946-2009 ($T_i = 63$) on the U.S., China, etc., but only data from 1989 or 1990 to the present for countries of the former Soviet Union (so, $T_i \approx 19$ for, say, Romania).

TSCS data are typically sorted first by observation i, and then, within observation, by time t (as illustrated above). The slides have an example of TSCS data on countries in the international system between 1946 and 1999.

Relational Data

Relational data is a phrase used to describe data where each observation reflects a "pairing" (or relationship) between two units of interest. For example, studies of international trade often analyze trade flows from one country to another. Because each country has at least the potential to trade with every other, the relevant unit of analysis is the "country pair" (sometimes referred to as the "dyad").

In a sample of N units of analysis, there are $\frac{N(N-1)}{2}$ relational observations – one for each unique "pair" of observations in the data. These are typically denoted by an i for the "first" unit in the pair, and a j for the "second" unit:

$$X_{ij} \in X = \begin{pmatrix} X_{12} \\ X_{13} \\ \vdots \\ X_{1N} \\ X_{23} \\ X_{24} \\ \vdots \\ X_{2N} \\ X_{34} \\ X_{35} \\ \vdots \\ X_{N-2,N-1} \\ X_{N-2,N} \\ X_{N-1,N} \end{pmatrix}$$

Note that the nature of the pairing is "non-directional": X_{12} is the same "pair" as X_{21}, so the latter does not appear in the data. Occasionally, we have reason to be interested in "directional" pairings: psychologists who study aggression in children, for example, might look at pairs of classmates, where it is necessary to distinguish between A bullying B and vice-versa. These are (unsurprisingly) referred to as *directional relations*, or *directional dyads*; for a sample of N units of analysis, there are $N(N-1)$ directed pairings of this sort.

Finally, note that we can combine the various data types above. For example, a commonly-used approach in quantitative international relations is to study "dyad-years," that is, pairs of countries measured in each year. These are essentially dyadic-level data with repeated measures on each dyad over time.

Missing Data

Sometimes, some of the cells of our data are empty – that is, the data in them are "missing." Missing data is a potentially severe issue, and a complicated one. For now, we'll just discuss two questions one might initially have about missing data.

Why Are Data Missing?

1. The observation itself does not exist (what was the per capita GDP of the United States in 1217 A.D.?),

2. Data simply don't exist for that observation (e.g., what type of star is my neighbor's sheepdog?),

3. Data exist, but are *impossible* to measure, or

4. Data exist, but were not measured. In this case, the classic categorization of missing data (due mainly to Don Rubin and his coauthors) is into three types:
    (a) Missing completely at random ("MCAR"),
    (b) Missing at random ("MAR"), and
    (c) Informatively (or "non-ignorably") missing.

We'll talk at greater length about these later in this course, and in others...

What to Do About Missing Data?

There are three primary ways of dealing with missing data; each has its advantages and disadvantages.

Listwise Deletion

One alternative if data are missing is simply not to use that observation in the analysis. Listwise deletion:

- is the "default" position of most software packages, and
- is often the simplest thing to do. Also,
- it is justifiable if the reason that data are missing can plausibly be considered unrelated to the thing that is being studied (that is, if the data are "MCAR").

Most of the time, then, listwise deletion is what we do. That said, listwise deletion is not a very good solution to missing data in most instances. In particular, listwise deletion:

- is, at best, not an "efficient" use of the data, and
- at worst, will cause the researcher to come to biased conclusions about what is being studied.

The conditions for this last thing to occur are when the reason that the data are missing is related to the thing being studied. So, for example, if you wanted to explore race and representation in the Deep South in the 1940s and 1950s, you could examine voters' tendencies to vote for anti-segregation candidates; however, since blacks were *both* systematically excluded from voting *and* tended to prefer anti-segregation candidates, the conclusions one would draw from only looking at actual voters (i.e., non-missing data) would be incorrect.

Interpolation / Replacement Values

When the degree of missingness is small, researchers occasionally simply “fill in” missing values with values that seem “reasonable.” The most commonly-used example of this approach is *interpolation*, which is used primarily with time series data. Suppose we have a series that looks like this:

t	Y_t
1	275
2	289
3	305
4	321
5	NA
6	350
7	366
.	.
.	.
.	.

It’s pretty clear that this particular time series is increasing in an almost exactly linear fashion (i.e., by about 15 units per period). Given this, many time series analysts would simply insert a value of (say) 335 at $t = 5$, rather than have the data be missing.

There are a number of different ways to interpolate data, and in fact one can think of imputation (discussed below) as a particular approach to interpolation. In general, however, interpolation is frowned upon, largely because the values ones “inserts” are often the product of relatively arbitrary decisions on the part of the researcher.

Missing Data Imputation

An alternative to listwise deletion is *imputation*, where missing values are “filled in” with likely values for the missing cells based on (usually complicated statistical) procedures. We won’t go into it now, but in a nutshell, imputation-based approaches to missing data:

- are more efficient than listwise deletion,
- reduce or eliminate the possibility of bias due to missingness, and
- are generally better alternatives than listwise deletion.

That said, listwise deletion remains the “usual” way of dealing with missing data in the social sciences. (Sigh).

Wrap-Up: A Few Practical Tips for Dealing with Data

Finally, some tips for working with data. Most of these are distilled from the Nagler (1995) article, which you should read three times, at least.

1. **Use descriptive variable names**. That means:
   - Spell it out: Don't call a variable "`gdppc00sq`" when you can call it "`GDPPerCapita2000Dollars-Squared`."
   - Use "directional" names: Don't call a variable "`sex`" when you can call it "`female`," or "`partyID`" when you can use "`GOP`."
2. **Be consistent in naming variables**. If you use "`lnXXXX`" to mean "the natural log of XXXX," then *always* use the "`ln`" prefix for logs.
3. **Label everything**. That means label datasets, variables, values of specific variables, etc. Similarly, comment / document any code you use to build / create / transform data.
4. **Never overwrite anything**. Electronic storage is essentially free; always start transforming, recoding, or otherwise modifying a variable by creating an exact copy of the original variable, then working on the copy. (That way, you can always go back and redo what you did). Similarly, a good practice is to begin modifying an entire dataset (e.g., merging, etc.) by creating a copy of the original, and then working on the copy instead.
5. **Log everything**. Keep a record of everything you do; every command, transformation, merge, etc. At best, do this using command/script files. But, if you (as I sometimes do) tend to prefer to work interactively, at a bare minimum keep log files of everything you do, so you can "retrace your steps" when necessary.

Next class: Data, *described...*