PLSC 308: Introduction to Political Research

Christopher Zorn

March 22, 2016

Data!

Rectangular Data

- Rows (are *observations*)
- Columns (are *variables*)
- Typically say there are N observations $i \in \{1, 2, 3, ...N\}$
- ...and K variables $k \in \{1, 2, 3, ...K\}$
- Sometimes denoted (e.g.) $\underset{N \times K}{\mathbf{X}}$

Typical Data Structure

	Variables				
	i	X_1	X_2	...	X_K
	1	X_{11}	X_{21}	...	X_{K1}
Observations	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}

Review: Variables

Variable Types

- Discrete
- Continuous

Levels of Measurement

- Nominal
- Ordinal
- Interval
- Ratio

Variables: Examples

Examples of Variables, by Type and Level of Measurement

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

Design and Data Structure

Cross-Sectional Data: 1997 Baseball Survey

```
. list respon age female followbaseball DH_appr

      +--------------------------------------------+
      | respon   age   female   follow~l   DH_appr |
      |--------------------------------------------|
   1. |      1    65        1          0         . |
   2. |      2    63        0          1         1 |
   3. |      3    56        1          1         . |
   4. |      4    24        1          0         . |
   5. |      5    47        0          0         . |
   6. |      6    81        1          1         . |
   7. |      7    28        0          1         1 |
   8. |      8    76        0          1         0 |
   9. |      9    22        1          0         . |
  10. |     10    39        1          0         . |
   .
   .
   .
```

Design and Data Structure

Time-Series Data: Supreme Court Clerks

```
. list Term female white top5law lcclerk

     +-------------------------------------------------+
     | Term    female      white    top5law    lcclerk |
     |-------------------------------------------------|
  1. | 1953         0        100   44.44445       12.5 |
  2. | 1954         0        100   64.70589   44.44445 |
  3. | 1955         0        100   76.47059   41.66666 |
  4. | 1956         0        100   55.55556         20 |
  5. | 1957         0        100   58.82353         30 |
  6. | 1958         0        100   57.89474   27.27273 |
  7. | 1959         0        100   61.11111   44.44445 |
  8. | 1960         0        100   66.66667   7.142858 |
  9. | 1961         0        100   55.55556   21.42857 |
 10. | 1962         0        100   71.42857   21.42857 |
 11. | 1963         0        100   78.94737         25 |
 12. | 1964         0        100       62.5   8.333334 |
 13. | 1965         0        100         70      43.75 |
 14. | 1966   5.88235        100   52.94118   33.33334 |
 15. | 1967         0    95.2381   66.66667   44.44445 |
   .
   .
   .
```

Design and Data Structure

Time-Series Cross-Sectional Data: Countries since 1945

```
. list country ccode year gdppc polity region coldwar

       +-------------------------------------------------------------------+
       |     country   ccode   year      gdppc   polity   region   coldwar |
       |-------------------------------------------------------------------|
 2820. | AFGHANISTAN     700   1946          .      -10        6         1 |
 2821. | AFGHANISTAN     700   1947          .      -10        6         1 |
 2822. | AFGHANISTAN     700   1948          .      -10        6         1 |
 2823. | AFGHANISTAN     700   1949          .      -10        6         1 |
   .
   .
   .
 2871. | AFGHANISTAN     700   1997        901       -7        6         0 |
 2872. | AFGHANISTAN     700   1998        937       -7        6         0 |
 2873. | AFGHANISTAN     700   1999          .       -7        6         0 |
 2874. |     ALBANIA     339   1946          .       -9        3         1 |
 2875. |     ALBANIA     339   1947          .       -9        3         1 |
 2876. |     ALBANIA     339   1948          .       -9        3         1 |
   .
   .
   .
10133. |    ZIMBABWE     552   1997       3153       -6        4         0 |
10134. |    ZIMBABWE     552   1998       3089       -6        4         0 |
10135. |    ZIMBABWE     552   1999          .       -6        4         0 |
       +-------------------------------------------------------------------+
```

Design and Data Structure

```
. list ccode1 ccode2 dyadid dem1 dem2 allies distance

      +-------------------------------------------------------------+
      | ccode1   ccode2   dyadid   dem1   dem2   allies   distance |
      |-------------------------------------------------------------|
   1. |      2       20     2020     10     10        1          0 |
   2. |      2       40     2040     10     -7        0       1135 |
   3. |      2       41     2041     10     -9        1       1437 |
   4. |      2       42     2042     10     -3        1       1477 |
   .
   .
   .
 128. |      2      900     2900     10     10        1       9916 |
 129. |      2      920     2920     10     10        1       8759 |
 130. |     20       40    20040     10     -7        0       1586 |
 131. |     20       41    20041     10     -9        0       1869 |
 132. |     20       42    20042     10     -3        0       1893 |
   .
   .
   .
 259. |     20      900    20900     10     10        0      10019 |
 260. |     20      920    20920     10     10        0       9009 |
 261. |     40       41    40041     -7     -9        0        722 |
 262. |     40       42    40042     -7     -3        0        868 |
 263. |     40       51    40051     -7     10        0        506 |
   .
   .
   .
8754. |    850      900   850900     -7     10        0       3361 |
8755. |    850      920   850920     -7     10        0       4804 |
8756. |    900      920   900920     10     10        1       1444 |
```

Missing Data: Why?

- The observation itself does not exist (what was the per capita GDP of the United States in 1217 A.D.?),
- Data simply don't exist for that observation (e.g., what type of star is my neighbor's sheepdog?),
- Data exist, but are *impossible* to measure, or
- Data exist, but were not measured. Yields:
    - Missing completely at random ("MCAR"),
    - Missing at random ("MAR"), and
    - Informatively (or "non-ignorably") missing.

Missing Data: What To Do?

Listwise Deletion

- Keep only “complete observations”...
- Simple...
- Default option in many cases
- Justifiable if data are MCAR

Missing Data Imputation

- “Fill in” missing values with “likely” values; repeat multiple times and average over the results
- Pros: More efficient, less bias
- Cons: Difficult / complex, not always accepted

Practical Data Tips

- **Use descriptive variable names**.

- **Be consistent in naming variables**.

- **Label everything**.

- **Log everything**.

- **Never overwrite anything**.