Loading and Manipulating Data

Thomas J. Leeper

Department of Political Science and Government
Aarhus University

November 14, 2013

Loading Data

Basic Data Summaries

Data Manipulation

Loading Data

- Typically use data in a "dataframe"
- Not restricted to one dataframe at a time
- Functions to load data all create a dataframe
- Statistical functions accept vectors or dataframes

Try on your own

Understand dataframe objects:

- "Dataframe Structure"
- "Rearranging Dataframes"

Questions so far?

Loading Data

- There's no "open" button
- A functions for each file format:
  - CSV: `read.csv`
  - TSV: `read.delim`
  - Stata: `read.dta` (from **foreign**)
  - SPSS: `read.spss` (from **foreign**)
- Almost anything can be loaded

Try on your own

Understand dataframe objects:
Do the "Loading Data" Tutorial

Questions so far?

Summary Statistics

- Lots of built-in functions to summarize data
- One important function: `summary`

Try on your own

Understand basic data summaries:

- "Univariate data summaries"
- "Correlations"

Summary Tables

- Tabulation is easy with `table`
- Creates univariate tables and cross-tables
- Tables are objects (of class "table"), so we can work with them like any other object

Try on your own

Understand tabulation and cross-tabulation:

- "Tabulation"

Summary Plots

- Visualization is one of R's greatest strengths
- One important function: `plot`
- Many other functions for specific types of plots
- Basic plots look okay
- Plots can be made beautiful with a little work

Try on your own

Understand basic plots:

- "Summary plots"
- "Plotting colors"

Questions so far?

Recoding vectors

- Recoding is all about indexing
- Several different ways to do it.
- The **car** package has a nice function: `recode`:

```
outvec <- recode(invec, "old1=new1;
old2=new2; else=NA")
```

Try on your own

Understand recoding:

- "Vector recoding"

Questions so far?

Scale Construction

- Vectorization makes scaling easy
- Use the usual operators: `+ - * / ^`
- Convenience functions for sums and means

Try on your own

Understand scale construction:

- "Basic Scale Construction"

Questions so far?

Missing Data

- R has one missing data value: `NA`
- Best to handle missing data during preprocessing

Try on your own

Understand Missing Data:

- "Missing Data"
- "Handling Missing Data"
- "Multiple Imputation"

Questions so far?