Intro to Social Science Data Analysis

Lecture 3: Gathering Data

Christopher Gandrud

September 17, 2012

1. Recap
2. Reminder: First Assignment
3. Data Gathering Principles
4. Response & Explanatory Variables
5. Importing Data Into R
6. Merging & ID Variables
7. Saving Data

Review

Last week we:

- Discussed why we care about data.
- Variables, observations, levels of measurement.
- Compared matrices to data frames.
- Learned basic data frame manipulation techniques.

Review

Last week we:

- Discussed why we care about data.
- Variables, observations, levels of measurement.
- Compared matrices to data frames.
- Learned basic data frame manipulation techniques.

Review

Last week we:

- Discussed why we care about data.
- Variables, observations, levels of measurement.
- Compared matrices to data frames.
- Learned basic data frame manipulation techniques.

Review

Last week we:

- Discussed why we care about data.
- Variables, observations, levels of measurement.
- Compared matrices to data frames.
- Learned basic data frame manipulation techniques.

Quick Quiz 1

What **level of measurement** are these variables at:

1. The number of people living in poverty in a city.
2. Whether or not a country is at war.
3. The names of a continent's major rivers.

Quick Quiz 2

Comment this code:

```r
Subject <- c("A", "B", "C")
Height <- c(3.2, 4.5, 3.1)

Data <- data.frame(Subject, Height)

DataB <- Data[2, ]
```

First Assignment

Due: Monday 24 September

Create a new data frame with country-level data from at least **two** different sources.

Create a folder in your Dropbox and **email me the link**.

The folder needs to include:

1. The new data frame saved as a `.csv` file.
2. A text file **describing the variables and their sources**.
3. A notebook `.html` file detailing how you created the data frame and saved it as a`.csv`.

Marking Criteria

The assigment will be marked based on two criteria:

1. How well you gathered and merged the data so that the new data set can be used for *statistical analysis with minimal data loss*.
2. How well you *describe the data* and *document your data gathering process*. I should be able to understand your data, code, and all steps you took.

Marking Criteria

The assigment will be marked based on two criteria:

1. How well you gathered and merged the data so that the new data set can be used for *statistical analysis with minimal data loss*.
2. How well you *describe the data* and *document your data gathering process*. I should be able to understand your data, code, and all steps you took.

Using Dropbox

This Week

Last week we learned basic data frame handling skills.

But how do we usually **get data into R** and make it useable for data analysis?

Populations

Your **research question** should usually **guide your data gathering**.

Your research question can indicate what the relevant **population** is.

- Does income-level explain voting behaviour in Korea?
- What do Southern European countries have higher deficits than Northern European ones?
- Why do civil wars happen?

Populations

Your **research question** should usually **guide your data gathering**.

Your research question can indicate what the relevant **population** is.

- Does income-level explain voting behaviour in Korea?
- What do Southern European countries have higher deficits than Northern European ones?
- Why do civil wars happen?

Populations

Your **research question** should usually **guide your data gathering**.

Your research question can indicate what the relevant **population** is.

- Does income-level explain voting behaviour in Korea?
- What do Southern European countries have higher deficits than Northern European ones?
- Why do civil wars happen?

Populations

Your **research question** should usually **guide your data gathering**.

Your research question can indicate what the relevant **population** is.

- Does income-level explain voting behaviour in Korea?
- What do Southern European countries have higher deficits than Northern European ones?
- Why do civil wars happen?

Data Gathering Principles 2

Samples

It may **not be possible** to gather data on the entire population of interest.

So we gather data on a subset of the population–a **sample**.

Anecdotal Evidence

If we want to answer questions about populations we should generally **avoid anecdotal evidence**. For example:

- The rich people I know vote for Party A not Party B.

Anecdotal evidence is gathered in a **haphazard way**. It is **not representative** of the population and usually includes **extreme cases**.

Because anecdotal evidence is not representative of the population of interest and **should not** be used to make **generalizable** conclusions.

Random Sample

To avoid selecting extreme cases & biasing our sample in other ways we should try to use **random sampling**.

A sample is random if every member of a population has an **equal probability** of being selected.

Sampling in this Course

In this course we will mostly be using **observational data**.

We will also usually use **convenience samples** where we try to gather data on as much of the population as possible (attempt **exhaustability**).

For example, if we want to research how democracy may affect economic development, we would try to gather data on

- level of democracy
- level of development

for **as many** countries and years as we can.

Sampling in this Course

Especially when we use convenience samples we **must always look out for factors that might bias our results**.

We must be he honest. **Clearly state any potential biases in our presentation of results.**

For example:

We wanted to study all countries, but usually found data only for wealthy countries.
We need to clearly state that our findings might only be generalizable to wealthy countries, not all countries.

Observational Data 2

When we use observational data it is especially important to gather data not only on **what we want to explain** and **what we think explains it**, but also **other factors** that may affect any relationship between these things that we might observe.

Response Variables

The phenomenon we are interested in explaining is operationalized by the **response variable**.

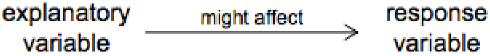
The response variable is also sometimes called the **dependent variable** and denoted by Y.

Explanatory Variables

The factor we believe may explain our phenomenon of interest is operationalized by the **explanatory variable**.

The explanatory variable is also sometimes called the **independent variable** and denoted by X.

Response & Explanatory Variables



explanatory variable

might affect

response variable

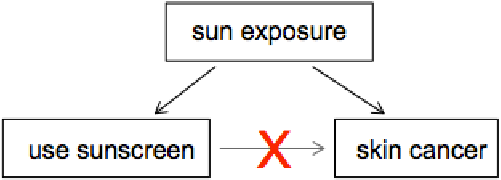
Source: Diaz et al. (2011, 29)

Response & Explanatory Variables

However, **correlation does not equal causation**.

In particular, we need to look out for **confounding factors**.

Confounding factors are **associated with both** the response and explanatory variables.



Source: Diaz et al. (2011, 31)

Control Variables & Data Gathering 1

In parts II & III of the course we look at ways to "**control**" for confounding factors.

At this point, when we gather data we need to consider not just the main response and explanatory variables, but also confounding factors.

Control Variables & Data Gathering 2

Note: We want to control for as many confounders as possible, but there is a data gathering **trade off**:

The more variables we gather, the more likely it is that we will have **missing data**.

If data is **not missing at random**, we will add sampling bias.

Ok, let's get some data into R.

Practical Tips

- Save your data as simple columns and rows (no fancy colours, etc.).
- Save it in a text format.
- Document:
  - Where the data is from,
  - What the variables mean,
  - How you have changed it.

Practical Tips

- Save your data as simple columns and rows (no fancy colours, etc.).
- Save it in a text format.
- Document:
  - Where the data is from,
  - What the variables mean,
  - How you have changed it.

Practical Tips

- Save your data as simple columns and rows (no fancy colours, etc.).
- Save it in a text format.
- Document:
    - Where the data is from,
    - What the variables mean,
    - How you have changed it.

Practical Tips

- Save your data as simple columns and rows (no fancy colours, etc.).
- Save it in a text format.
- Document:
  - Where the data is from,
  - What the variables mean,
  - How you have changed it.

Practical Tips

- Save your data as simple columns and rows (no fancy colours, etc.).
- Save it in a text format.
- Document:
  - Where the data is from,
  - What the variables mean,
  - How you have changed it.

Practical Tips

- Save your data as simple columns and rows (no fancy colours, etc.).
- Save it in a text format.
- Document:
  - Where the data is from,
  - What the variables mean,
  - How you have changed it.

Simple Data Set

Data sets should be **simple**.

If you open the data in Microsoft Excel:

- The **first row** should have the variables names
- The variable names should be simple and use the same rules as R object names.
- There should be no extra information or styling.
- If you have missing data cells should be empty or have NA.

Simple Data Set

Data sets should be **simple**.

If you open the data in Microsoft Excel:

- The **first row** should have the variables names
- The variable names should be simple and use the same rules as R object names.
- There should be no extra information or styling.
- If you have missing data cells should be empty or have NA.

Simple Data Set

Data sets should be **simple**.

If you open the data in Microsoft Excel:

- The **first row** should have the variables names
- The variable names should be simple and use the same rules as R object names.
- There should be no extra information or styling.
- If you have missing data cells should be empty or have NA.

Simple Data Set

Data sets should be **simple**.

If you open the data in Microsoft Excel:

- The **first row** should have the variables names
- The variable names should be simple and use the same rules as R object names.
- There should be no extra information or styling.
- If you have missing data cells should be empty or have NA.

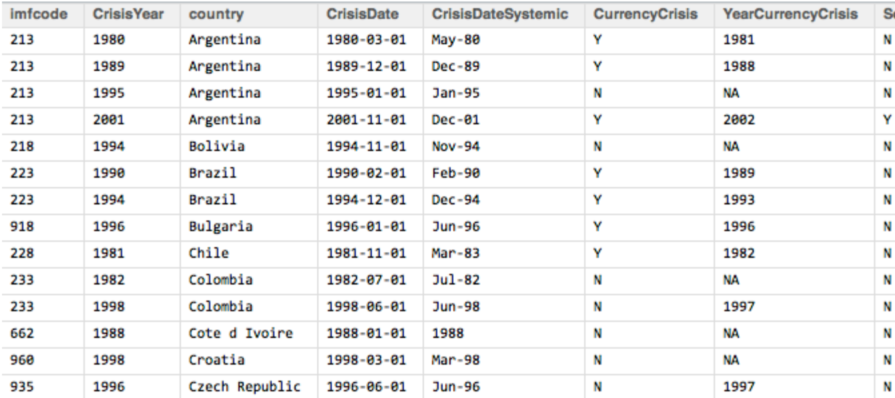
Simple Data Set

Data sets should be **simple**.

If you open the data in Microsoft Excel:

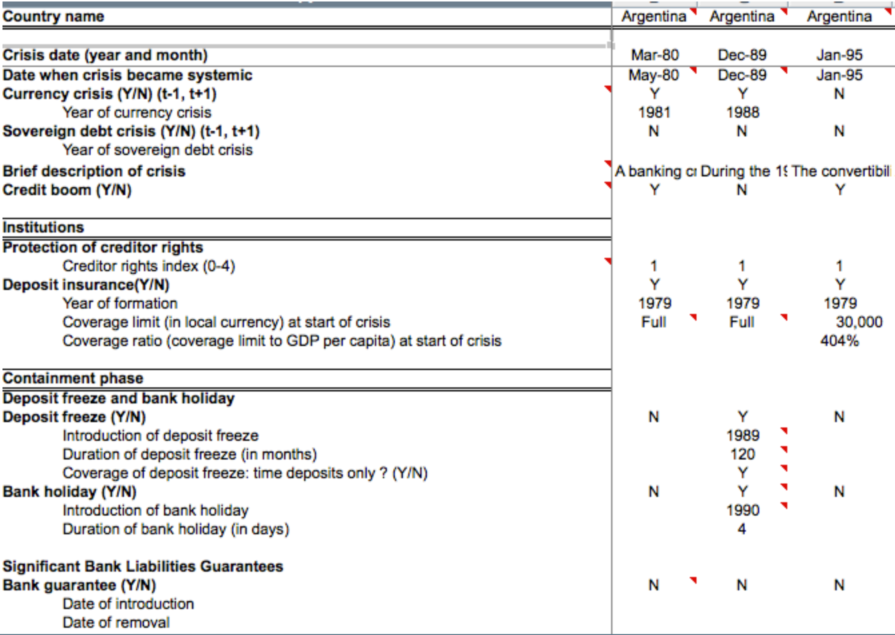
- The **first row** should have the variables names
- The variable names should be simple and use the same rules as R object names.
- There should be no extra information or styling.
- If you have missing data cells should be empty or have NA.

A Good Data Set



imfcode	CrisisYear	country	CrisisDate	CrisisDateSystemic	CurrencyCrisis	YearCurrencyCrisis	S
213	1980	Argentina	1980-03-01	May-80	Y	1981	N
213	1989	Argentina	1989-12-01	Dec-89	Y	1988	N
213	1995	Argentina	1995-01-01	Jan-95	N	NA	N
213	2001	Argentina	2001-11-01	Dec-01	Y	2002	Y
218	1994	Bolivia	1994-11-01	Nov-94	N	NA	N
223	1990	Brazil	1990-02-01	Feb-90	Y	1989	N
223	1994	Brazil	1994-12-01	Dec-94	Y	1993	N
918	1996	Bulgaria	1996-01-01	Jun-96	Y	1996	N
228	1981	Chile	1981-11-01	Mar-83	Y	1982	N
233	1982	Colombia	1982-07-01	Jul-82	N	NA	N
233	1998	Colombia	1998-06-01	Jun-98	N	1997	N
662	1988	Cote d Ivoire	1988-01-01	1988	N	NA	N
960	1998	Croatia	1998-03-01	Mar-98	N	NA	N
935	1996	Czech Republic	1996-06-01	Jun-96	N	1997	N

A Bad Data Set



Country name	Argentina	Argentina	Argentina
Crisis date (year and month)	Mar-80	Dec-89	Jan-95
Date when crisis became systemic	May-80	Dec-89	Jan-95
Currency crisis (Y/N) (t-1, t+1)	Y	Y	N
Year of currency crisis	1981	1988	
Sovereign debt crisis (Y/N) (t-1, t+1)	N	N	N
Year of sovereign debt crisis			
Brief description of crisis	A banking cı	During the 1¦	The convertibil
Credit boom (Y/N)	Y	N	Y
Institutions			
Protection of creditor rights			
Creditor rights index (0-4)	1	1	1
Deposit insurance(Y/N)	Y	Y	Y
Year of formation	1979	1979	1979
Coverage limit (in local currency) at start of crisis	Full	Full	30,000
Coverage ratio (coverage limit to GDP per capita) at start of crisis			404%
Containment phase			
Deposit freeze and bank holiday			
Deposit freeze (Y/N)	N	Y	N
Introduction of deposit freeze		1989	
Duration of deposit freeze (in months)		120	
Coverage of deposit freeze: time deposits only ? (Y/N)		Y	
Bank holiday (Y/N)	N	Y	N
Introduction of bank holiday		1990	
Duration of bank holiday (in days)		4	
Significant Bank Liabilities Guarantees			
Bank guarantee (Y/N)	N	N	N
Date of introduction			
Date of removal			

Data in Text Files

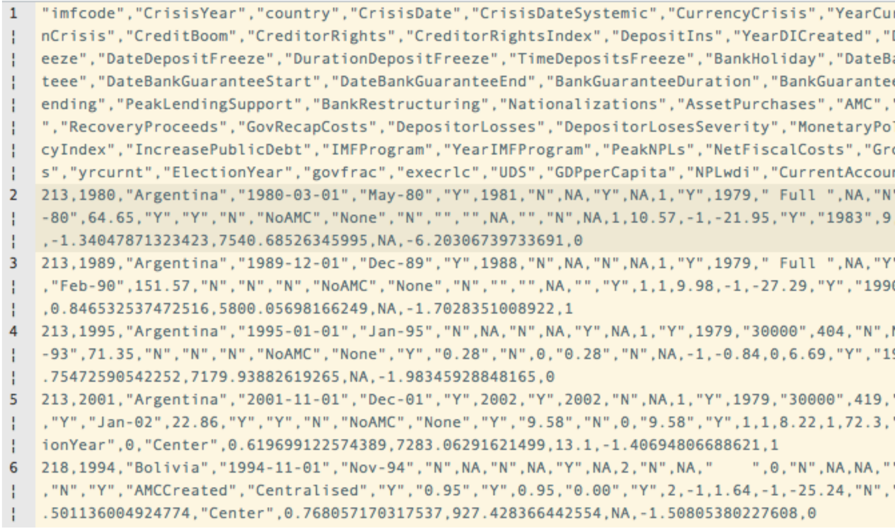
You should save your data in a **plain text file format**.

Plain-text file formats:

- Are simple,
- Can easily be loaded into R.

My favourite is **comma seperated values (.csv)**.

Raw Plain-Text Data File



```
"imfcode","CrisisYear","country","CrisisDate","CrisisDateSystemic","CurrencyCrisis","YearCu
nCrisis","CreditBoom","CreditorRights","CreditorRightsIndex","DepositIns","YearDICreated","
eeze","DateDepositFreeze","DurationDepositFreeze","TimeDepositsFreeze","BankHoliday","DateB
teee","DateBankGuaranteeStart","DateBankGuaranteeEnd","BankGuaranteeDuration","BankGuarante
ending","PeakLendingSupport","BankRestructuring","Nationalizations","AssetPurchases","AMC",
","RecoveryProceeds","GovRecapCosts","DepositorLosses","DepositorLosesSeverity","MonetaryPo
cyIndex","IncreasePublicDebt","IMFProgram","YearIMFProgram","PeakNPLs","NetFiscalCosts","Gr
s","yrcurnt","ElectionYear","govfrac","execrlc","UDS","GDPperCapita","NPLwdi","CurrentAccou
213,1980,"Argentina","1980-03-01","May-80","Y",1981,"N",NA,"Y",NA,1,"Y",1979," Full ",NA,"N
-80",64.65,"Y","Y","N","NoAMC","None","N","","",NA,"","N",NA,1,10.57,-1,-21.95,"Y","1983",9
,-1.34047871323423,7540.68526345995,NA,-6.20306739733691,0
213,1989,"Argentina","1989-12-01","Dec-89","Y",1988,"N",NA,"N",NA,1,"Y",1979," Full ",NA,"Y
,"Feb-90",151.57,"N","N","N","NoAMC","None","N","","",NA,"","Y",1,1,9.98,-1,-27.29,"Y","199
,0.846532537472516,5800.05698166249,NA,-1.7028351008922,1
213,1995,"Argentina","1995-01-01","Jan-95","N",NA,"N",NA,"Y",NA,1,"Y",1979,"30000",404,"N",
-93",71.35,"N","N","N","NoAMC","None","Y","0.28","N",0,"0.28","N",NA,-1,-0.84,0,6.69,"Y","1
.75472590542252,7179.93882619265,NA,-1.98345928848165,0
213,2001,"Argentina","2001-11-01","Dec-01","Y",2002,"Y",2002,"N",NA,1,"Y",1979,"30000",419,
,"Y","Jan-02",22.86,"Y","Y","N","NoAMC","None","Y","9.58","N",0,"9.58","Y",1,1,8.22,1,72.3,
ionYear",0,"Center",0.619699122574389,7283.06291621499,13.1,-1.40694806688621,1
218,1994,"Bolivia","1994-11-01","Nov-94","N",NA,"N",NA,"Y",NA,2,"N",NA,"    ",0,"N",NA,NA,"
,"N","Y","AMCCreated","Centralised","Y","0.95","Y",0.95,"0.00","Y",2,-1,1.64,-1,-25.24,"N",
.501136004924774,"Center",0.768057170317537,927.428366442554,NA,-1.50805380227608,0
```

Import Plain-Text Data

Importing Plain-Text Data 1

Use the `read.table` command.

For example, if you have a file on your Desktop called MyFile.csv, load it like this:

```r
# Import plain-text data into R
NewData <- read.table(file = "~/Desktop/Myfile.csv",
    sep = ",")
```

Importing Plain-Text Data 2

```
# Import plain-text data into R
NewData <- read.table(file = "~/Desktop/Myfile.csv",
    sep = ",")
```

Note:

- `"~/Desktop/Myfile.csv"` is the **directory** + **file name**. This will be different depending on what operating system you have and where the file is.
- `sep = ","` tells R that the values in a row are seperated by commas

Importing Plain-Text Data 2

```
# Import plain-text data into R
NewData <- read.table(file = "~/Desktop/Myfile.csv",
    sep = ",")
```

Note:

- `"~/Desktop/Myfile.csv"` is the **directory** + **file name**. This will be different depending on what operating system you have and where the file is.
- `sep = ","` tells R that the values in a row are seperated by commas

Data from Other Statistics Programs

Foreign Package

If you are importing data from another statistics program, such as SPSS or Stata, `foreign` packages.

For example, to load a Stata file with the **file extension** `.dta`:

```r
# Load foreign package library
library(foreign)
# Load Stata data file MyFile.dta
NewData <- read.dta(file = "~/Desktop/Myfile.dta")
```

Data from Microsoft Excel

Data from Microsoft Excel

Data from Excel, Practical Tips

If your data is in Excel format (i.e. `.xls`)

- In Excel, simplify the data set (columns and rows, take out styling),
- Save it as a text file.
- Load using `read.table`

Data from Excel, Practical Tips

If your data is in Excel format (i.e. `.xls`)

- In Excel, simplify the data set (columns and rows, take out styling),
- Save it as a text file.
- Load using `read.table`

Data from Excel, Practical Tips

If your data is in Excel format (i.e. `.xls`)

- In Excel, simplify the data set (columns and rows, take out styling),
- Save it as a text file.
- Load using `read.table`

Downloading Data from the Internet

Downloading Data from the Internet

You can sometimes download data directly from the internet.

If the data is in a plain-text format on a webpage by itself, use the getURL command from the RCurl package.

```r
# Load required packages
library(RCurl)
library(foreign)
# Create an object for the URL.
url <- "http://myFile.csv"
# Use getURL from RCurl to download the file.
myData <- getURL(url)
# Create a data frame
MyData <- read.csv(myData)
```

More Details

For more details see the Wiki: `http://bit.ly/QemXsI`.

Data APIs

Some packagages use data APIs (application programming interface) to download data from particular sources.

You can download World Bank Development Indicator data directly into R using the WDI package.

- World Bank Development Indicators Website: `http://data.worldbank.org/indicator`

For example to download data on life expectency at birth.

```
# Load package
library(WDI)

# Download data
LifeExpect <- WDI(indicator = "SP.DYN.LE00.FE.IN")

# Show variable names
names(LifeExpect)

## [1] "iso2c"             "country"
## [3] "SP.DYN.LE00.FE.IN" "year"
```

More data sources

You can find **other data API packages**

- `http://bit.ly/Qw16RY`

The **MacroData Guide** is a good source of social science data.

- `http://www.nsd.uib.no/macrodataguide/topic.html`

The Federal Reserve Bank of St. Louis **FRED** database.

- `http://research.stlouisfed.org/fred2/`

Remember: Google is always your friend.

Merging & ID Variables 1

When you want to merge two data sets you should create a **standardised ID variable**.

The ID variable **matches observations** in the two data sets.

Merging & ID Variables 2

ID variable names should be the same. If they are not the same, you can rename them using the `rename` command in the Reshape package.

Merging & ID Variables 3

For example, to rename the variable ”SP.DYN.LE00.FE.IN” from our previous example:

```
# Load package
library(reshape)

# Rename SP.DYN.LE00.FE.IN LifeExpectFemale
LifeExpect <- rename(LifeExpect, c(
  SP.DYN.LE00.FE.IN = "LifeExpectFemale"))

# Show rename results
names(LifeExpect)

## [1] "iso2c"            "country"
## [3] "LifeExpectFemale" "year"
```

Merging & ID Variables 4

Before you merge your data, the **same ID variables** need to have the **same values**.

For example, country names need to be spelled the same.

Merging & ID Variables 5

To recode a variable use subscripts. For example:

```
# Recode Korea, Rep -> SouthKorea
LifeExpect$country[LifeExpect$country
                    == "Korea, Rep."] <- "SouthKorea"
```

A useful package for creating country ID variables is countrycode.

This will turn country names into country codes.

You can use the country codes to merge the data.

countrycode Example

To create IMF country codes with our LifeExpect data:

```
# Load package
library(countrycode)

# Create IMF codes
LifeExpect$IMFCode <- countrycode(LifeExpect$country,
    origin = "country.name", destination = "imf")
# Show end of data set
tail(LifeExpect)

##     iso2c  country LifeExpectFemale year IMFCode
## 979    ZM   Zambia            43.00 2003     754
## 980    ZM   Zambia            42.49 2002     754
## 981    ZW Zimbabwe            42.89 2005     698
## 982    ZW Zimbabwe            42.46 2004     698
## 983    ZW Zimbabwe            42.39 2003     698
## 984    ZW Zimbabwe            42.69 2002     698
```

Saving your Data

To save your data to a plain-text `.csv` file use `write.csv`.

```
write.csv(LifeExpect,
          file = "LifeExpectacyData.csv",
          row.names = FALSE)
```

Finally

Remember:

- Document all of your work!
- Decribe your sources & the variables.
- Always look at your data after merging it. Does it make sense?

Finally

Remember:

- Document all of your work!
- Decribe your sources & the variables.
- Always look at your data after merging it. Does it make sense?

Finally

Remember:

- Document all of your work!
- Decribe your sources & the variables.
- Always look at your data after merging it. Does it make sense?

References I

Diaz, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel. OpenIntro Statistics. 1st ed. http://www.openintro.org/stat/downloads.php.