

Regression III

Lecture 1: Preliminary

Dave Armstrong¹

Until August: Department of Politics & IR and Nuffield College (Oxford)
From September: Department of Political Science (UW-Milwaukee)

e: davearmstrong.ps@gmail.com
w: www.quantoid.net/ICPSR.php

June 23, 2009

¹These slides have been heavily influenced by the previous instructor of this course - Bob Andersen (U of Toronto)

1/51

Outline

Introduction

OLS Basics: Theory and Presentation of Results

Non-Linearity

Outliers and Influential Data

Other Topics

Getting Started with R

2/51

Contact Info

- Instructor: Dave Armstrong
Office Hours: 1-2PM
E-mail: davearmstrong.ps@gmail.com
Course Website: <http://www.quantoid.net/ICPSR.php>
- Teaching Assistant: Matthew Painter
E-mail: painter.63@sociology.osu.edu

3/51

Today's Lecture

1. Show some "highlights" of the course:
 - Applied Regression with attention to *modern* extensions.
 - Explore methods for "problem" data.
 - Emphasis on graphical techniques.
 - Use modern methods to overcome problems with regular linear regression.
2. Prerequisites:
 - 2.1 Regression in matrix form,
 - 2.2 MLE, and
 - 2.3 Statistical Inference.
3. Get started using **R**
 - All analysis will be done using **R**.
 - The course on *Computing in R and S* is compulsory (unless you already know **R**).

4/51

Course Materials

Suggested Texts:

Fox, John. (2008) Applied Regression Analysis and Generalized Linear Models, 2nd ed. Thousand Oaks, CA: Sage Publications, Inc.

Fox, John. (2002) An R and S-PLUS Companion to Applied Regression. Thousand Oaks, CA: Sage Publications, Inc.

A more detailed list is at the back of the course syllabus

- Course materials - lecture slides, syllabus and **R**-scripts will be available at
 - Z:/Armstrong and
 - <http://www.quantoid.net/ICPSR.php>

5/51

Introduction

OLS Basics: Theory and Presentation of Results

Non-Linearity

Outliers and Influential Data

Other Topics

Getting Started with R

6/51

Versatility of Linear Regression

- Simple linear regression summarizes the relationship between a quantitative predictor variable and quantitative response variable with a straight line
- The linear model can be extended to handle:
 - Several explanatory variables (multiple regression)
 - Categorical explanatory variables and interactions between explanatory variables
 - Categorical dependent variables
 - Simple and monotone nonlinear relationships
- The linear model is desirable because it is simple to fit and easy to interpret
- The data must satisfy a number of assumptions, however, for the linear model to be the appropriate model
- Looking at the data graphically allows us to assess whether these assumptions are likely to be met

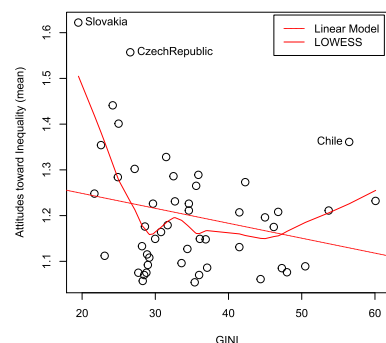
7/51

Regression and Causation

- When observational data are used, statistical models are inherently descriptive. . . they are not causal though we often want to describe a causal process
- Criteria for a causal theory
 - Empirical relationship
 - Cause precedes the effect in time
 - Elimination of rival explanations (i.e., no confounding variables)
 - We must control for possible confounding and intervening variables when using observational data
- The goal is to fit the model that tells the right story
 - We must ensure that the model adequately represents the pattern in the data
 - When making inferences about populations, we must also ensure that the assumptions of the model are met

8/51

Importance of Control Variables



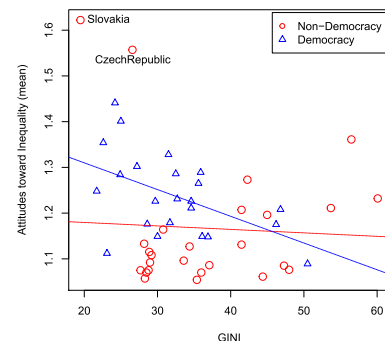
- Relationship is clearly not linear: at high levels of income inequality, attitudes towards inequality are negatively related to the Gini coefficient; at low levels of inequality the trend is in the opposite direction
- There also may be influential outliers
- Next step: Explore outliers, possible control variables and interactions

	Estimate	S.E.	t-value
Gini	-0.003	0.002	-1.71

$R^2 = 0.059$

9/51

Control Variables and Interactions



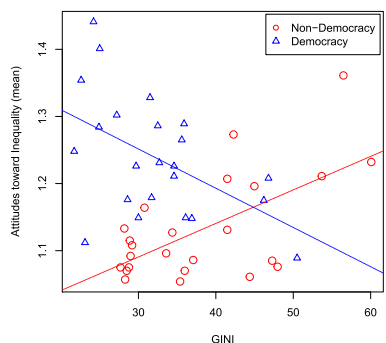
- Addition of democracy and its interaction with Gini coefficient (allowing for different slopes for the two groups) improves the model, but the effects are still not statistically significant
- Two outliers (Slovakia and Czech Republic) are clearly influential for Non-democracy model

	Estimate	S.E.	t-value
Gini	-0.006	0.003	-1.71
Dem	0.234	0.141	1.65
Gini × Dem	-0.005	0.004	-1.25

$R^2 = 0.147$

10/51

Influential Cases



- Czech Republic and Slovakia have unusually high levels of attitudes toward inequality
- When these cases are removed from the model there is a vast improvement in fit
- Both the Gini coefficient and Democracy have significant effects on attitudes and there is a strong interaction between them

	Estimate	S.E.	t-value
Gini	-0.006	0.002	-2.99
Dem	0.478	0.088	5.43
Gini × Dem	-0.011	0.002	-4.30

$R^2 = 0.513$

11/51

Assumptions of Multiple Regression

The multiple regression model takes the following form:

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i$$

The assumptions of the model for ordinary least squares regression (OLS) concern the errors:

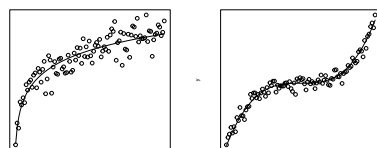
1. Linearity
2. Constant error variance
3. Normally distributed errors
4. Uncorrelated error terms
5. X's are independent of the errors

When these assumptions are met, the OLS estimators are unbiased and efficient estimates of the population parameters

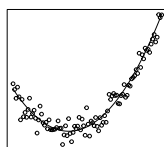
12/51

OLS and Nonlinearity(1): Transformable Nonlinearity

- Transformations of one or both variables can help straighten the relationship between two quantitative variables
- Possible only when the nonlinear relationship is simple and monotone
 - Simple implies that the curvature does not change – there is one curve
 - Monotone implies that the curve is always positive or always negative



(a) (b)

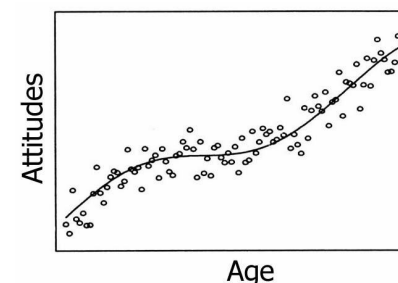


(c)

13/51

OLS and Nonlinearity (2): Polynomial Regression

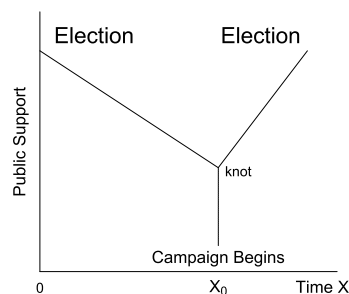
- When the relationship is not monotone or simple, we could try polynomial regression
- If there is only one bend in the curve, we fit a quadratic model – i.e., we could add an X^2 (age^2) term to the model
 - For every bend in the curve, we add another higher term to the model
- The two bends below suggest trying a cubic regression (i.e., include age , age^2 and age^3 as predictors)



14/51

OLS and Nonlinearity(3): Regression Splines

- Regression splines allow the regression line to change direction abruptly
- Piecewise polynomial functions that are constrained to join smoothly at points called knots.
 - These are regression models with restricted dummy regressors
 - Separate regression lines are fit within regions (i.e., the range of X is partitioned) that join at knots



15/51

Handling Complex Nonlinearity: A more general way to think of regression

- Regression traces the conditional distribution of a dependent variable, Y , as a function of one or more explanatory variables, X 's

$$p(y|x_1, \dots, x_k) = f(x_1, \dots, x_k)$$

- Linear regression assumes a linear function but often this is not appropriate
- If the nonlinear relationship is complex, nonparametric regression and generalized additive models provide alternative ways of capturing the relationship
 - These models estimate the functional from the data themselves (i.e., they do not assume linearity)

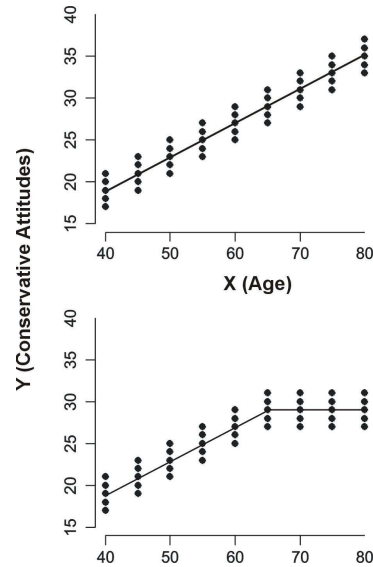
16/51

Modelling the Conditional Mean

- With large samples and when the values of X are discrete, it is possible to estimate the regression by directly examining the conditional distribution of Y
- Here we determine the mean of Y (could also use the median) at each value of X :

$$\mu = E(Y|x) = f(x)$$

- A naïve nonparametric regression line connects the conditional means
- Here, a linear regression would work well for the top graph, but poorly for the bottom graph.



17/51

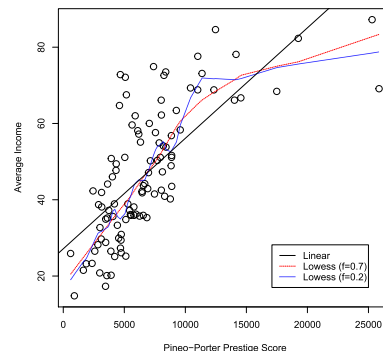
Modelling the Conditional Mean (2)

- With extremely large data sets or when the explanatory variable takes on discrete values, we can easily calculate conditional distributions
- In the 'real world' of social science data, however, we do not often have this luxury
 - If X is continuous, even when the sample size is large, we may not have enough cases at each value of X to calculate precise conditional means
- If we have a large sample we can dissect the range of X into narrow bins that contain many observations, obtaining fairly precise estimates of the conditional mean of Y within them
 - The smaller the sample size, the larger the bin sizes need to be, and thus the fewer the bins. As long as the relationship between X and Y is not too complicated this is fine; If there is complex nonlinearity it can become problematic

18/51

Locally Weighted Scatterplot Smoothing (Lowess)

- Lowess is a form of nonparametric regression that fits a separate weighted least squares regression line to each x_i value and then joins the fitted values together
- We choose a span for the proportion of the data to be included in each local regression that provides a smooth fit
- Especially useful when comparing to a linear regression fit
 - The blue line is $s=.2$; the red line is $s=.7$; the black line is the linear fit



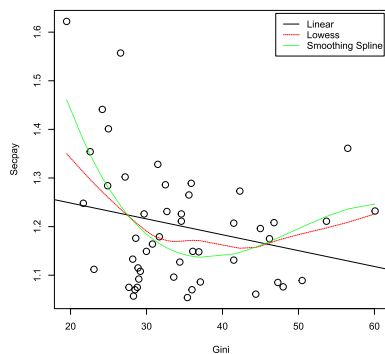
19/51

Smoothing Splines

- Smoothing splines offer a compromise between global polynomial regression and local polynomial regression
 - Different piecewise polynomial trends that are constrained to be joined smoothly at the knots
 - Not as smooth as global polynomial regression, but generally behave much better at the peaks
- Rather than choose a span as for lowess curves, we usually choose the degrees of freedom – low degrees of freedom will fit a smooth curve; high degrees of freedom will give a rough curve
- Smoothing splines play an important role in Generalized Additive Models

20/51

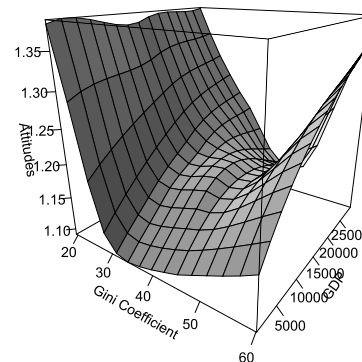
Linear and Nonparametric fits



- The black line is the linear fit, the red line is the lowess smooth from a local linear regression with a span of 0.6
- A clear departure from linearity in these data.
- An F-test comparing the RSS from the linear model with that from the more general trend of the lowess model allows us to assess whether the relationship is linear.

21 / 51

Multiple Nonparametric Regression



- The regression surface is clearly nonlinear for gini
- As with the simple model, we could test for nonlinearity using an F-test comparing the RSS of this model with the RSS of the linear model
- If we had only one more predictor, the lowess model would be impossible to interpret - we can't see in more than 3 dimension.

22 / 51

Generalized Additive Models

- Additive Regression Models overcome the curse of dimensionality by applying local regression to low dimensional projections of the data
- The nonparametric additive regression model is:

$$Y_i = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \dots + f_k(x_{ik}) + \varepsilon_i$$

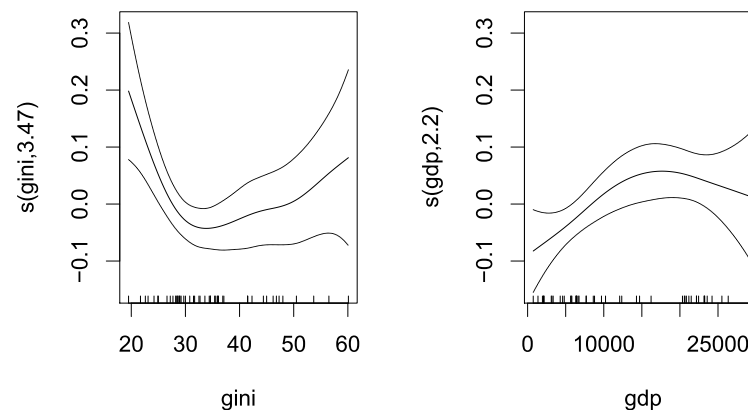
instead of

$$Y_i = \alpha + f_1(x_{i1}, x_{i2}, \dots, x_{ik}) + \varepsilon_i$$

- Additive models create an estimate of the regression surface by a combination of a collection of one-dimensional functions
 - In effect, then, they restrict the nonparametric model by excluding interactions between the predictors
 - An estimation procedure called "backfitting" is used to fit the models.

23 / 51

Generalized Additive Model (2): $secpay \sim s(gdp) + s(gini)$



24 / 51

Introduction

OLS Basics: Theory and Presentation of Results

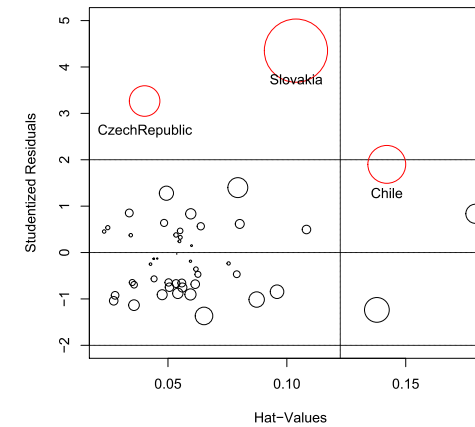
Non-Linearity

Outliers and Influential Data

Other Topics

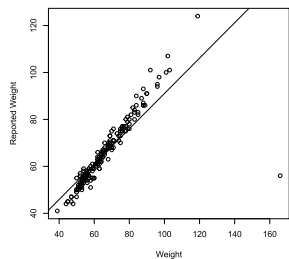
Getting Started with R

Influential Cases: Influence Plots



OLS and Influential Cases

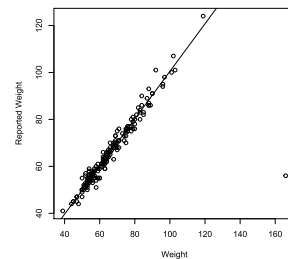
Figure: With Outlier



$\hat{\beta}$	$\sigma_{\hat{\beta}}$	R^2
0.753	0.037	0.699

- Measured weight for case **12** was miscoded.
- Meaningful effect on regression line and R^2 .

Figure: Without Outlier



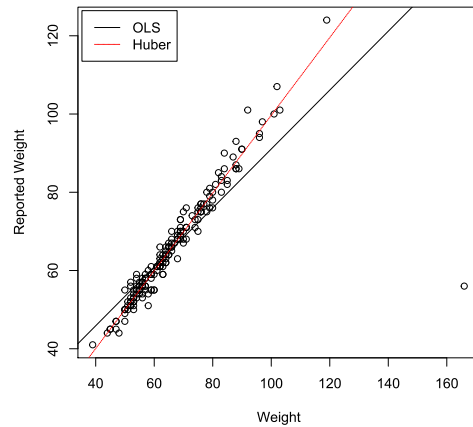
$\hat{\beta}$	$\sigma_{\hat{\beta}}$	R^2
1.014	0.013	0.972

Robust Regression (1)

- Robust regression (e.g., MM-Estimation with Huber weights) can give a significantly better fit to data with influential cases than does OLS.
 - Both slope coefficients and statistical significance can change drastically
 - Residual standard error can also be decreased significantly
 - Not the same as commonly used robust standard errors which are used to compensate for a unknown pattern of heteroskedasticity
- Robust regression assumes large sample sizes, however, so if we have a small one, we may wish to use bootstrapping in order to have greater confidence in our statistical inferences
- Robust regression can also be extended to generalized linear models

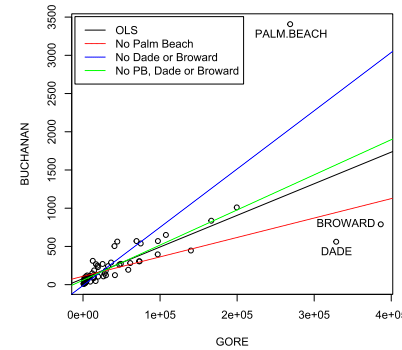
Robust Regression (2)

Figure: OLS and Huber Robust fits to Davis Data



29/51

OLS and Joint Influence: Votes by County 2000 US Presidential Election



- If all three outliers are included, the regression line is similar as to when they are all deleted
- Jointly BROWARD and DADE have nearly equal influence to the influence of PALM BEACH
- If either PALM BEACH or BROWARD/DADE are deleted separately, however, the OLS fit changes dramatically

30/51

Introduction

OLS Basics: Theory and Presentation of Results

Non-Linearity

Outliers and Influential Data

Other Topics

Getting Started with R

Topics Covered in this Course

- Review of the basics of OLS
- Effective presentation of complicated linear models
- Diagnostics and possible “solutions” for problems of nonlinearity, non-normality, nonconstant error variance, outliers, collinearity
- Weighted least squares, robust regression, bootstrapping and cross-validation, robust standard errors, mixed models, latent growth models
- Missing data and multiple imputation

31/51

32/51

Getting Started with R

- What is **R** ?
- A tiny **R** session
- Resources
- Setup under Windows.
- Getting data in
- Graphs & Statistical Models

33/51

What is R?

- A free, open-source implementation of the S language for data analysis and graphics
- Available for various operating systems (including Linux, Mac and Windows)
- A complete programming language
- Supported by a comprehensive help system and a large international community of users
- Increasingly used in advanced social-science research, as well as in many other disciplines
- In constant flux
- Not guaranteed by anyone to be fit for any purpose!

34/51

R Resources

- The main source for everything R-related is the Comprehensive R Archive Network (CRAN): <http://cran.r-project.org>
There one can find:
 - Executable installation files for R
 - The source code
 - Extensive documentation and contributed guides
 - Information about many add-on packages
- RNews - a quarterly(ish) R newsletter containing “news” about R along with a couple of in-depth articles about particular packages or routines.
- The R mailing list - very active, but not especially noob friendly. They’re quite happy to answer questions providing you follow the posting guidelines.
- The Political Methodologist - a twice-yearly publication of the Political Methodology Section of the American Political Science Association. This contains an sections with practical advice on R and \LaTeX

35/51

Add-on Packages

- Add on packages are easily installed from the menus within R (see the course syllabus)
- Once installed, the package must be loaded into the current interactive R session
- Many packages contain datasets. These must also be loaded to be used
- The number symbol “#” is used to insert comments - R will not read anything after it (only works for a single line)

```
> library(car) # Loads the car "car" package
> data(Prestige) # Brings the "Prestige" data to the workspace
> attach(Prestige) # attaches the "Prestige" data
```

36/51

Documentation for R

- Installed as part of the **R** help system are the following documents:
 - *An Introduction to R* (about 100 pages). Gives an introduction to the language and how to use **R** for doing statistical analysis and graphics
 - *R Data Import/Export* (about 35 pages). Describes the import and export facilities available in **R** itself or via the foreign package
 - *Writing R Extensions* (about 75 pages). Covers how to create your own packages, write **R** help files, etc.
- There are also various 'unofficial' guides on CRAN under 'contributed'
- Finally, Fox (2002) provides a great book to start with for regression analysis

37/51

Getting Help in R

- A number of different types of help are available by clicking the help menu:
 - Documentation on all installed packages is available in a web browser by clicking help → html help
 - The 'official' manuals can be loaded in PDF format by clicking help → manuals
- Help about individual functions and objects can also be obtained within **R** by typing
 - `help(data)` or `?data`, for help on something whose name is known
 - `help.search("ordinal")`, to search all the installed help files for occurrence of a particular text string
 - `aprops("stem")`, to look for 'stem' in the names of objects available in the current **R** session
 - `RSiteSearch("ordinal")` opens a web browser and searches R's mailing list and documentation for matches

38/51

Getting set-up

- Download the installer from CRAN (see directions in course syllabus)
- A good way of working with **R** is:
 1. Erase the 'start in' field in the Rgui.exe shortcut (right-click properties)
 2. For each project on which you use **R** make a directory containing the data etc and put a copy of the Rgui.exe shortcut in that directory.
 3. Then to use **R** for that project, double click on that copy of the shortcut. This will keep all your files in the same place.
- It is also useful to have a good text editor. Notepad will do, but there are much better alternatives
 - The R plug-in for WinEDT is called R-WinEDT (I use this one)
 - An alternative is the Emacs Speaks Statistics (ESS) package
 - Tinn-R is yet another alternative, but there is no development being done on this piece of software

39/51

Getting data in (1): Entering data directly

- The *concatenate* function, `c()`, combines individual cases together into a vector
- The `cbind()` (columns bind) and `rbind()` (rows bind) functions combine vectors together into a matrix
- The `data.frame()` function makes the matrix and a data frame object

```
> Dept <- c("Soc", "Pol", "Psych", "Econ")
> Year <- c(1,2,3,4)
> Dataset <- data.frame(cbind(Dept, Year))
> Dataset
  Dept Year
1  Soc   1
2  Pol   2
3 Psych  3
4  Econ  4
```

40/51

Getting data in (2): External datasets

- For rectangular data in a text file, use `read.table()`:
`Mydata<-read.table("dataname.txt", header=TRUE)`
 - `header=TRUE` signifies that the first row contains variable names
- The foreign library imports data files from and exports data to other formats:

```
> library(foreign)
> GSSdat <- read.spss("C:/regressionIII/data/gss91.sav",
  use.value.labels=T, to.data.frame=T)
```
- `use.value.labels=T` converts value labels to factors, otherwise they will be treated as quantitative
- All SPSS variable names are imported in upper case letters.
- You can also import Stata datasets as follows:

```
> library(foreign)
> GSSdat <- read.dta("C:/regressionIII/data/gss91.dta",
  convert.factors=T)
```
- `convert.factors=T` reads variables with value labels as factors (more on this later).

41/51

A note on `attach()`

- `attach()` adds the argument of the command to the R search path.
- This means, that the names of any variables in the attached data will be accessible by typing their names directly without any mention of the data.
 - This means that `data$var` and `var` will produce the same result if `data` is attached
- This may seem appealing, but beware. If you have multiple data sources attached and there is some intersection in the set of variable names, **R** knows which one it is accessing, but you need to make sure you know, too.
- I will generally not attach data, rather I will refer to variables with `data$var` or `data[, "var"]`. You can do this however you like.

42/51

Aside: Converting from upper-case to lower-case letters

If you do not want the variable names in upper-case letters, you could use the following program to convert them:

```
lower.case <- function(x){
  tmp <- strsplit(x, split="")
  nwords <- length(tmp)
  tmp.out <- NULL
  for(i in 1:nwords){
    tmp.out <- c(tmp.out, paste(letters[ifelse(match(tmp[[i]],
      c(LETTERS, letters)) <=26,
      match(tmp[[i]], c(LETTERS, letters)), match(tmp[[i]],
      c(LETTERS, letters))-26)],collapse=""))
  }
  tmp.out
}
```

The result is:

```
> test <- c("Abc", "DEF")
> test
[1] "Abc" "DEF"
> lower.case(test)
[1] "abc" "def"
```

43/51

Re-specifying variables after importing to R

- To make a numerically coded variable into an *unordered factor* (categorical variable):

```
> VarA <- c(1,2,3,2,3,4,4,2,3,2,1,1,2,3,2)
> VarB <- as.factor(VarA)
> VarA
[1] 1 2 3 2 3 4 4 2 3 2 1 1 2 3 2
> VarB
[1] 1 2 3 2 3 4 4 2 3 2 1 1 2 3 2
Levels: 1 2 3 4
```
- to make `VarA` into an *ordered factor*:

```
> VarC <- ordered(VarA, levels=c("1", "2", "3", "4"))
> VarC
[1] 1 2 3 2 3 4 4 2 3 2 1 1 2 3 2
Levels: 1 < 2 < 3 < 4
```

44/51

Recoding Variables using the `recode` function in the `car` package

- Recoding into a quantitative variable:

```
> library(car)
> VarD <- recode(VarA, "1:2=2; 3:4=1")
> table(VarD, VarA)
      VarA
VarD 1 2 3 4
     1 0 0 4 2
     2 3 6 0 0
```

- Recoding a quantitative variable into an *unordered factor*:

```
> VarE <- recode(VarA, "1:2='Low'; 3:4='High'")
> table(VarE, VarA)
      VarA
VarE  1 2 3 4
     High 0 0 4 2
     Low  3 6 0 0
```

45/51

S Modeling Language

- The S modeling language is convenient in that it has a similar notation for most types of models
- Model specification generally takes the following form:
Outcome \sim Explanatory Variables
- Where the tilde sign (\sim) is interpreted as “regressed on”
- For the general linear model, terms represent additive components as in the regression equation itself
- Some examples of formulas are:

```
y ~ x1 + x2           # Main effects of x1 and x2
y ~ x1 + x2 + x1:x2  # Adds interaction of x1 and x2 to above
y ~ x1*x2            # all main and interactive effects
```

46/51

Graphs in R

- Graphs in R are very flexible.
- They can be saved to many different formats, or simply copied and pasted elsewhere
- All graphs are drawn on a chosen device either until a new device is started, or the device is closed `dev.off()`
- Some commonly used graphics devices are
 - `postscript("mygraphs.ps")`
 - Necessary for \LaTeX
 - `pdf("mygraph.pdf")`
 - Necessary for PDF \LaTeX
 - `windows()`
 - The default graphics device

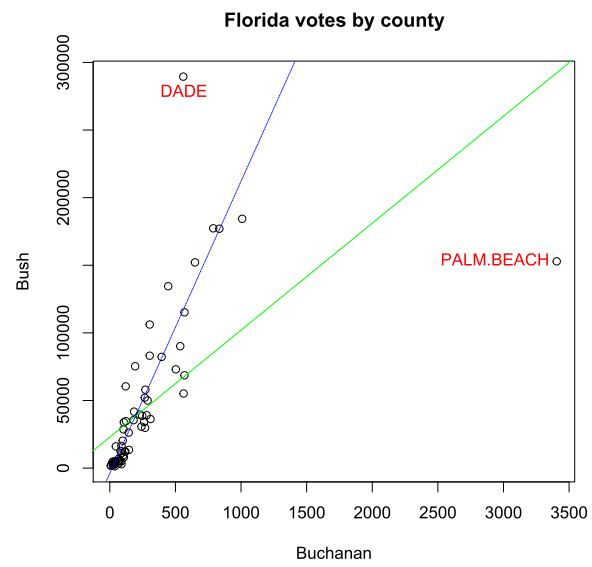
47/51

Graph Example (1)

```
data(Florida)
plot(Florida$BUCHANAN, Florida$BUSH,
     xlab="Buchanan", ylab="Bush")
identify(Florida$BUCHANAN, Florida$BUSH,
         rownames(Florida), col="red")
abline(lm(BUSH~BUCHANAN, data=Florida),
       col="green", lwd=2)
abline(lm(BUSH ~ BUCHANAN, data=Florida,
         subset =BUCHANAN < 3000), col="blue",
       lty=2, lwd=2)
title("Florida votes by county")
```

48/51

Result: Graph Example (1)



49/51

Readings for Today

- * Fox (2008), Chapters 1 & 2
- * Fox (2002), Chapters 1 & 2
- Venables and Ripley (2002), Chapters 1-3

50/51

Tomorrow's class

- Everything you ever wanted to know about OLS

Readings:

- * Fox (2008) Chapters 5, 6, & 9
- * Fox (2002) Chapter 4

51/51