Applied Economic Analysis, EC4044

Dr Stephen Kinsella | University of Limerick

Spring 2017

Introduction

The formulation of a problem is often more essential than its solution which may be merely a matter of mathematical or experimental skill. –Albert Einstein

Routine, in an intelligent man, is a sign of ambition.
—WH Auden

Welcome to this module

- This is a *brand new* module, EC4044.
- This is the first time it has been taught. So please, bear with me. It's all under construction.
- Slides will be added to this master version as I get the chance to build them out, shared as both .rmd and .pdf.
- In addition to being a pedagogical experiment, this is also technical experiment.
- We'll be using some open-source tools, your feedback will be crucial as we develop the material.

Brief note on the slides

- The slides and the code that generates them are a part of the course. They will be added to incrementally, so you should expect to see longer and longer slide decks being created.
- That is, I won't be giving out individual slides per lecture. You will see why in a moment.

Learning outcomes for this module

0. Understand principles of data science;
1. Understand where economic data come from;
2. Understand the *politics* of economic data collection and dissemination;
3. Estimate simple economic models
4. Understand the merits of qualitative as well as quantitative economic analysis. Economics is not all ones and zeroes. You do have to talk to real people from time to time.

How you'll learn

0. Hands-on, with your laptops or tablets in class. Laptops are better, but whatever works for you.
1. The idea is to make this module, as far as possible, one 36 hour-long lab.
2. You'll become familiar with a cutting edge statistical language and gain the ability to produce really nice reports, slides, and data analysis using this software. You'll also learn how to interview individuals and groups.
3. Importantly, your work will be open for everyone to view. You'll gain an appreciation of the kind of work that gets social scientists interested in things.

Key Resources

- David Freedman, *Statistical Models, Theory and Practice*, Cambridge University Press, 2009. This is the best book on statistics I have ever read.
- Garrett Grolemund and Hadley Wickham, R for data science, O'Reilly, 2016.
- Gary Koop, *Analysis of Economic Data*, Wiley, 2013. This book is a classic and fun to read.
- *Lots* of online resources with R, especially Datacamp

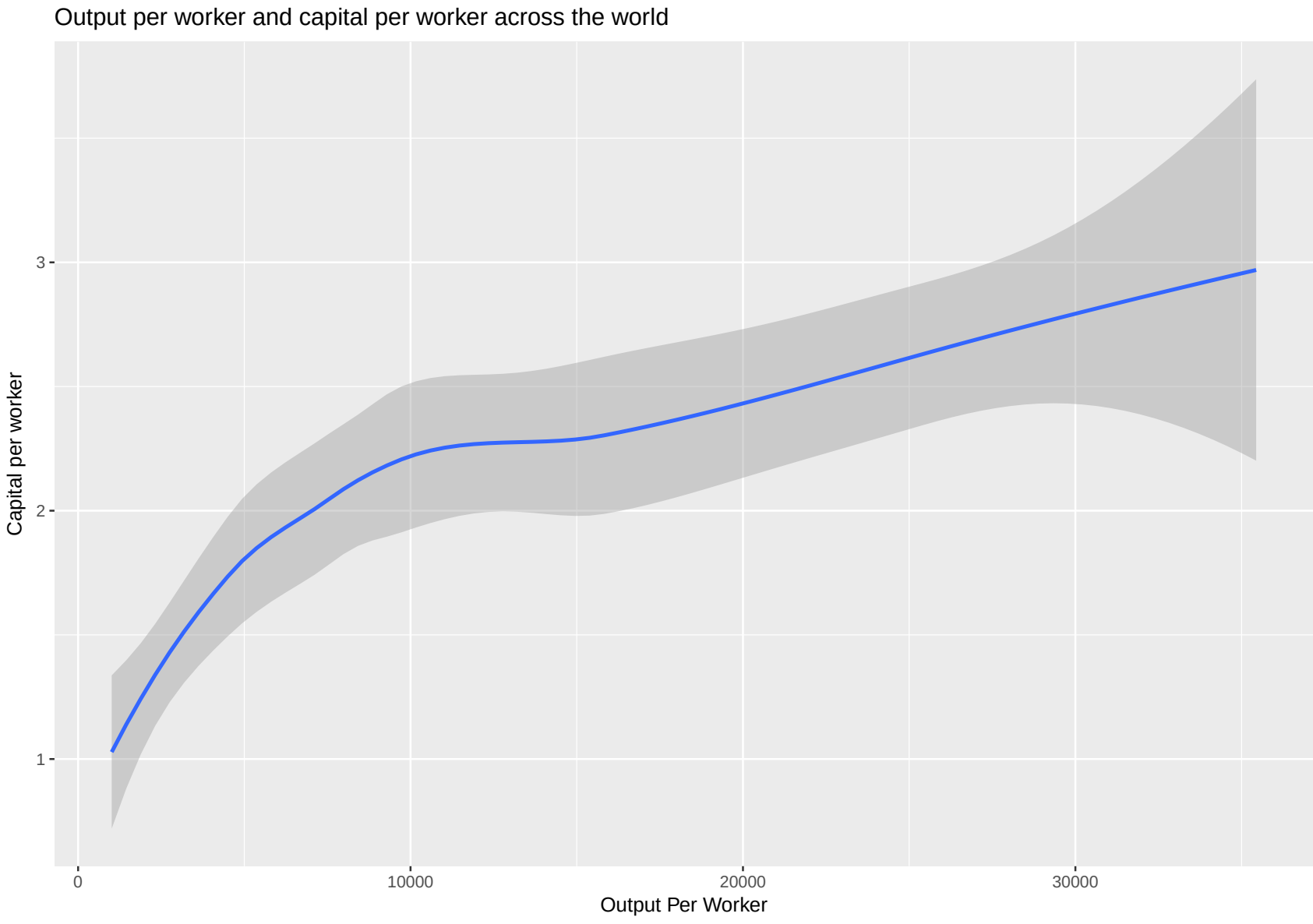
Key Software resources

- R and Rstudio.
- Github, where all the notes, code, and other elements for the course will be.
- Datacamp.com, for the introductions to R.
- SULIS contains the readings.
- Turnitin, for the final data project.

Why R?

- This is not an econometrics class, or a basic statistics class–you are taking one of those.
- This is about using the theories you've learned over the last 3 semesters and learning how to go about investigating the real world and their applicability to that world.
- So you need a tool, but one that can't be too complicated. R allows you enough power (for free) to analyse 50 million datapoints at the same time, but you don't need to know everything about what's going on under the hood to do useful work.
- That is, I want you to be able to drive a car, not tinker with its engine or repair it
- This means we'll be skipping over a lot of detail to get to the important points.

An example of what I mean



Assessment

- 2 Datacamp courses, 'introducing you to R', worth 5% and 'correlation and regression' in R, worth 10%. These are both due by the end of week 6, but you should aim to have the introduction course sorted by week 3 at the latest.
- 1 *optional* Datacamp course, which you choose, worth 10%. Peruse the course catalogue and let us know which one you want. Guide your own learning. You have access to the entire suite of modules for the entire semester. This would cost you 30 euros every month to learn, and you can add the certifications to you linkedin profiles etc for signalling purposes.
- 1 end of term project, due week 13. Details of this will be given in the tutorials, it is worth 75%-85%, depending.
- The objective is not to over-assess you. Rather, there are some basics you need to know to progress in this module, and then to let you play with the data and the tools we give you for 6-8 weeks. The more you use R for economic analysis, the better you'll be at it.

Lecture 1: Movitation, statistical basics and data handling

- Types of economic data: micro and macro
- Observation studies and experiments
- Statisitical inference, probabilit distributions, fitting a model.
- Graphical methods
- Descriptive statistics
- Expected Values and Variances
- Example: Hall and Jones, Growth Accounting, 1999.
- Reading: Koop, Chapter 2, Freedman, Chapter 1

Lab 1: Introduction to R (Teetor, Chapters 1 and 2)

- Installing R + Rstudio
- Getting Github, Quandl, and FRED accounts
- Working through Twotorials

Lecture 2: Modeling using simple regression (Freedman Chapter 1, Koop Chapter 4)

- Understanding correlation
- Why are variables correlated
- Staring at XY plots
- Complexities
- Example: Wage/Salary data from 1985.

Lab 2: Working with data in R (Teetor, Chapter 3)

- Getting data into R
- Simple manipulation
- Your first graphs
- Interpreting your first graphs.

Lecture 3: More on Simple Regression (Koop, Chapter 4, Freedman, Chapter 3)

- Best fitting line
- Interpreting OLS estimates
- Measuring the fit of a regression model
- Nonlinearity in Regression
- Factors affecting β
- Calculating confidence intervals for β
- Example: regression by hand, roll your own betas using R.

Lab 3: Matrix algebra FTW (Freedman Chapter 4)

- Concepts you need to know to get the most out of the rest of the course.
- What is a matrix
- Determinants & Inverses
- Random vectors
- Positive definite matrices

Lecture 4: Multiple Regression (Freedman, Chapter 5)

- Explaining variance in multiple regression
- Statistical aspects
- Interpreting multiple regression
- Biases: multicollinearity/heteroskedasticity/autocorrelation
- Example: education spending and educational attainment

Lab 4: Working with complex data sets in R

- Cleaning and working with data
- Running regressions, outputting tables, interpreting results
- Mashing data sets together
- Writing reports & making slides with Rmarkdown.

Lecture 5: Multiple regression 2 (Freedman, Chapter 5)

- Multiple regression with dummy variables
- Distributed lag models
- Applying theory to data
- Example: Gender pay disparities & producer theory.

Lab 5: Working with complex data sets in R, part deux

- Cleaning and working with data
- Running regressions, outputting tables, interpreting results
- Example: Mashing HUGE data sets together

Lecture 6: Time series analysis (Wickham,)

- Autocorrelation and $AR(1)$ processes
- Stationarity and Unit roots
- Example: Volatility in asset prices
- Example (gapminder): How does life expectancy change over time for each country?

Lab 6: Time series data in R

Lecture 7: Machine learning

- Introduction to machine learning
- Classification and maximum likelihood
- Neural networks
- Example: zip code recognition problems

Lab 7: Your first neural network

Lecture 8: Machine learning 2

- Discovering meaningful patterns in massive data
- Designing models with hidden and observed variables.
- Statistical learning, criticising the model.

Lab 8: More ML

Lecture 9: Big data and public policy

- Big data, what it is, and what it isn't.
- Machine learning and public policy
- Manski vs Minsky

Lab 9: Working on your data-project

Lecture 10: Interviewing & qualitative analysis

- Applied economic analysis is not just thinking about data and numbers. It is also about finding things out about by simply asking people.
- *Why Wages Don't Fall during a Recession*, Truman F. Bewley.
- Structured vs Unstructured interview techniques

Lab 10: Working on your data project

Lecture 11: Interviewing & qualitative analysis

- Survey data vs interview data
- Example: coding and thematic analysis Burnard et al, 2008
- Exercises in interviewing & transcription.

Lecture 12: Recap

Why it is useful to learn these skills in this way and in this order.

My overarcing goal is to help you work as economists.

1. Most problems you'll face that need serious analysis require you to 1. talk to people and figure out what's going on and 2. get data of some kind and see what's going on.
2. Once you have data on your problem, you need to start thinking about cleaning it, visualising it, summarising it, transforming it, and modeling it.
3. Finally, you need to be able to write about it, present it, and more and more, reproduce it so that people can check your work.

R can help you do all of these things.

This is the basic process of applied economic analysis

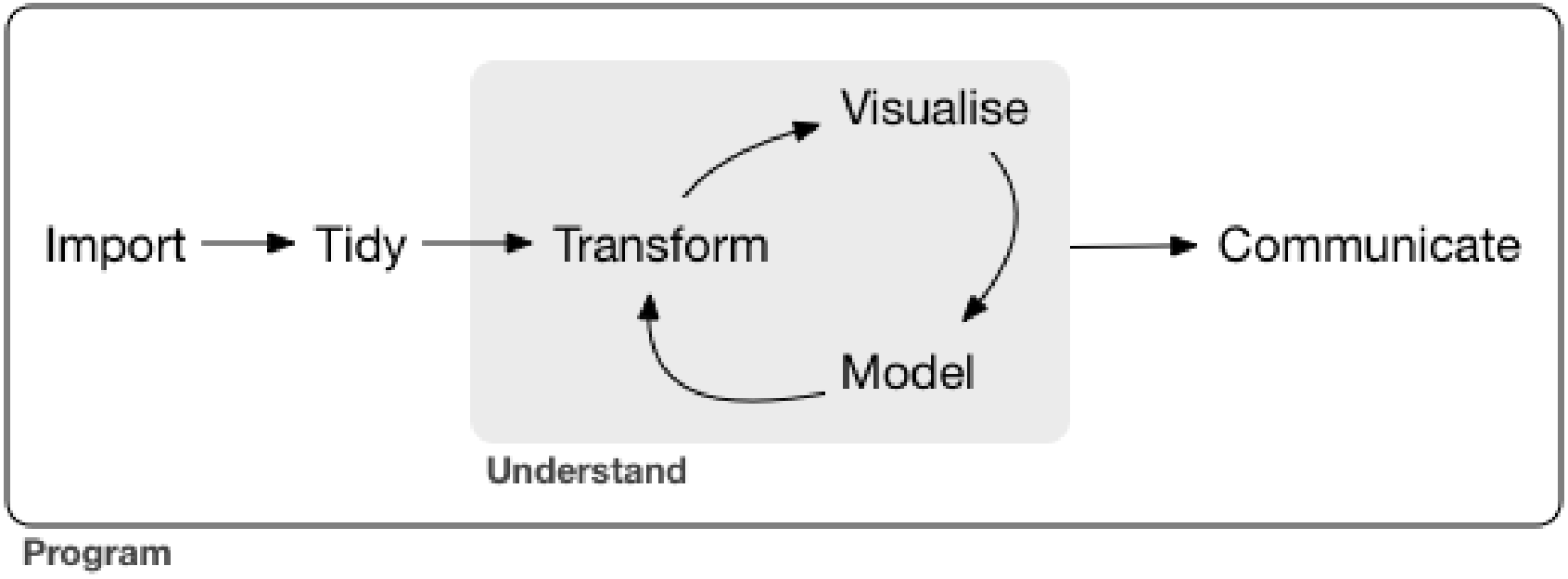


Figure 1: Source: Wickham, 2016

Lecture 1: Movitation, statistical basics and data handling

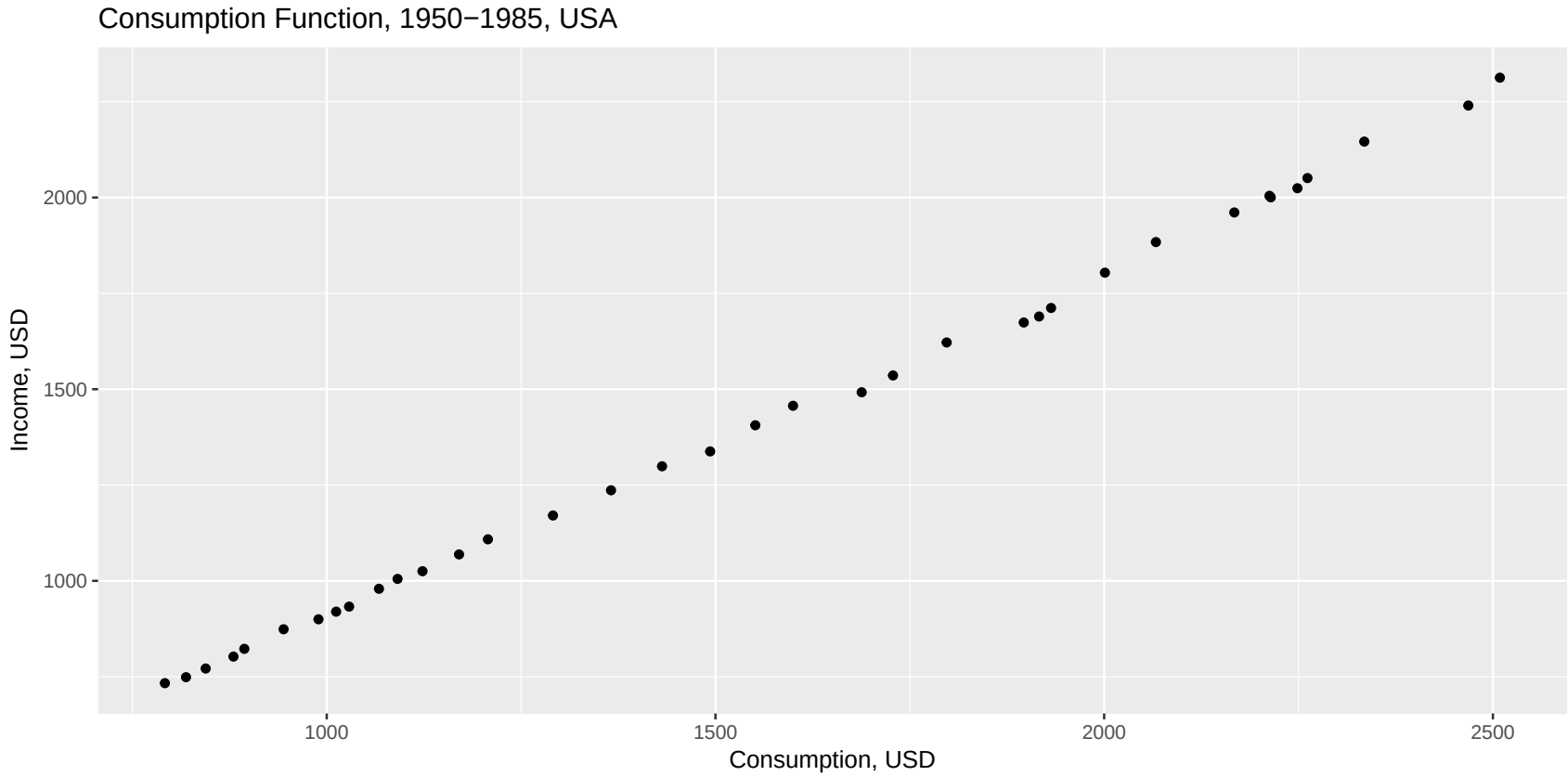
- Types of economic data: micro and macro
- Observation studies and experiments
- Statisitical inference, probabilit distributions, fitting a model.
- Graphical methods
- Descriptive statistics
- Expected Values and Variances
- Example: Hall and Jones, Growth Accounting, 1999.
- Reading: Koop, Chapter 2, Freedman, Chapter 1

Why we use models

- to summarise data
- to predict the future
- to predict the results of interventions

Example: Consumption and Income in the USA, 1950 - 1985 (Note the code that generates the figure is here)

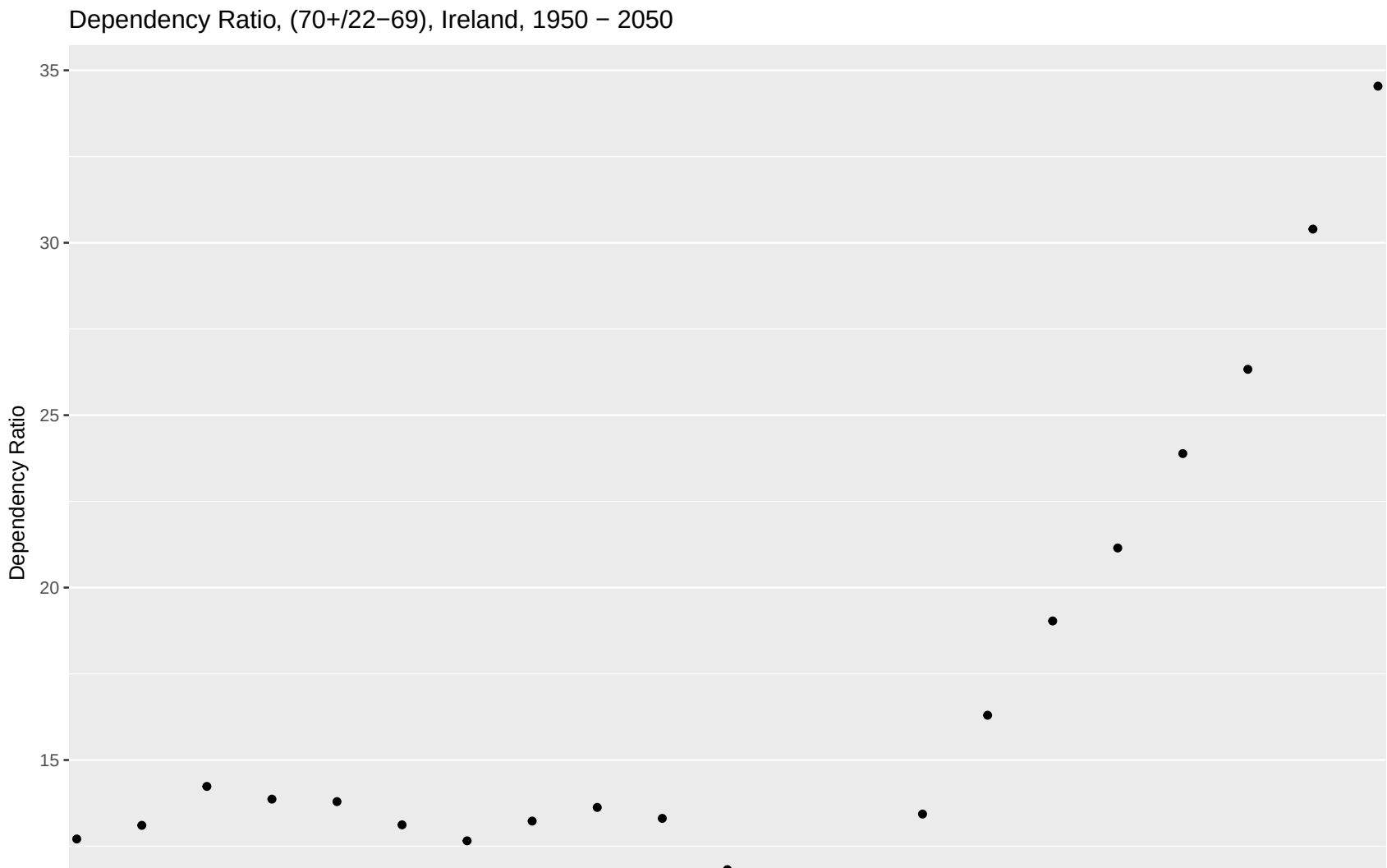
```
cf<-read.delim("http://web.uvic.ca/~dgiles/blog/consump.dat
                sep="", header=TRUE)
ggplot(data=cf)+geom_point(mapping=aes(x = Y, y = CONS))+gg
```




```
REG<-lm(CONS ~ Y, data=cf)
```

Example

```
url<- "https://dl.dropboxusercontent.com/s/1hp05328ek2ws1e/
dep <- read_csv(url)
qplot(Year, Dep, data = dep)+scale_x_discrete(breaks=pretty
```



Digression for a mathematical refresher

- Economists are often interested in the relationship between two (or more) variables.
- A very general way of denoting a relationship is through the concept of a function.
- If the economist is interested in the factors that explain why some houses are worth more than others, he/she may think that the price of a house depends on the size of the house.
- In mathematical terms, he/she would then let Y denote the variable "price of the house" and X denote the variable "size of the house" and the fact that Y depends on X is written using the notation:

$$Y = f(X)$$

This notation should be read "Y is a function of X" and captures the idea that the value for Y *depends* on the value of X.

Thinking in straight lines

The equation of a straight line (what was called a "linear function" above) is

$$Y = \alpha + \beta X$$

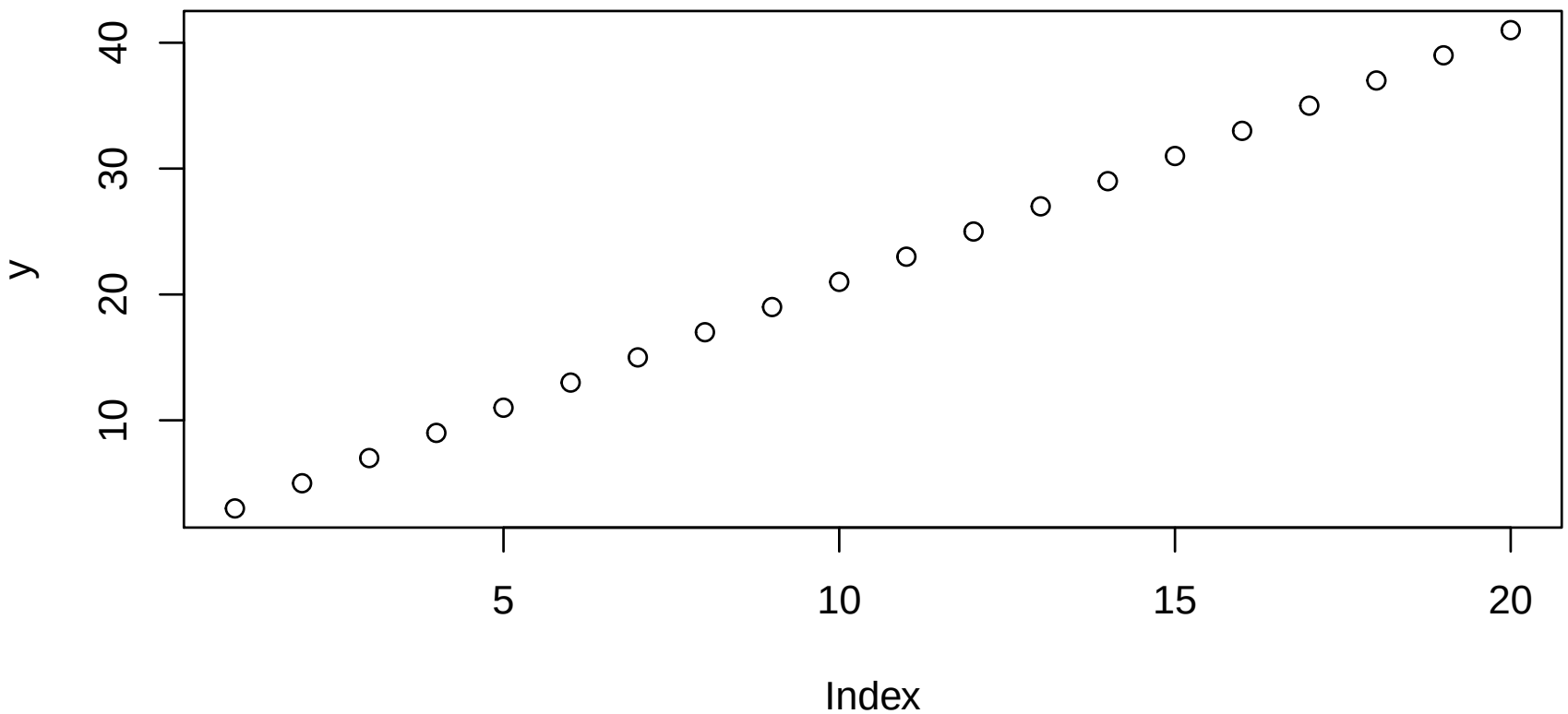
where α and β are coefficients, which determine a particular line. So, for instance, setting $\alpha = 1$ and $\beta = 2$ defines one particular line while $\alpha = 4$ and $\beta =$ -5 defines a different line.

It is probably easiest to understand straight lines by using a graph

- In terms of an XY graph (i.e. one which measures Y on the vertical axis and X on the horizontal axis) any line can be defined by its intercept (α) and slope (β).
- The slope is a measure of how much Y changes when X is changed, or dy/dx.

The XY graph of $Y = \alpha + \beta X$ for $\alpha = 1, \beta = 2$

```
a= 1 ## This is a parameter value.
b= 2 ## This is a parameter value.
x= seq(from = 1, to = 20, by =1) ## This command generates
y = a + b*(x) ## This is the equation showing how y depends
plot(y) ## Plot y.
```



Notation

- Subscripts are used to denote different observations from a variable, so W_1 is the wage of the first individual, W_2 the wage of the second individual, and W_n is the wage of the nth individual.
- We'll use superscripts to denote exponents. So X^2 squares the value of the variable X. X^a raises X to the ath power. a can take a value like 0.1 or 100 or whatever.
- We will often add things up across each other. In many cases we want to add up observations (e.g. when calculating an average you add up all the observations and divide by the number of observations). $\sum$ is an operator like (+) or (-) and the sub and superscripts tell us where to start and stop the summation.

$$\sum_{i=1}^{i=100} (W_i) = W_1 + W_2 + \ldots + W_{100}$$

Question: what would $\sum^{i=29}(W_i)$ do?

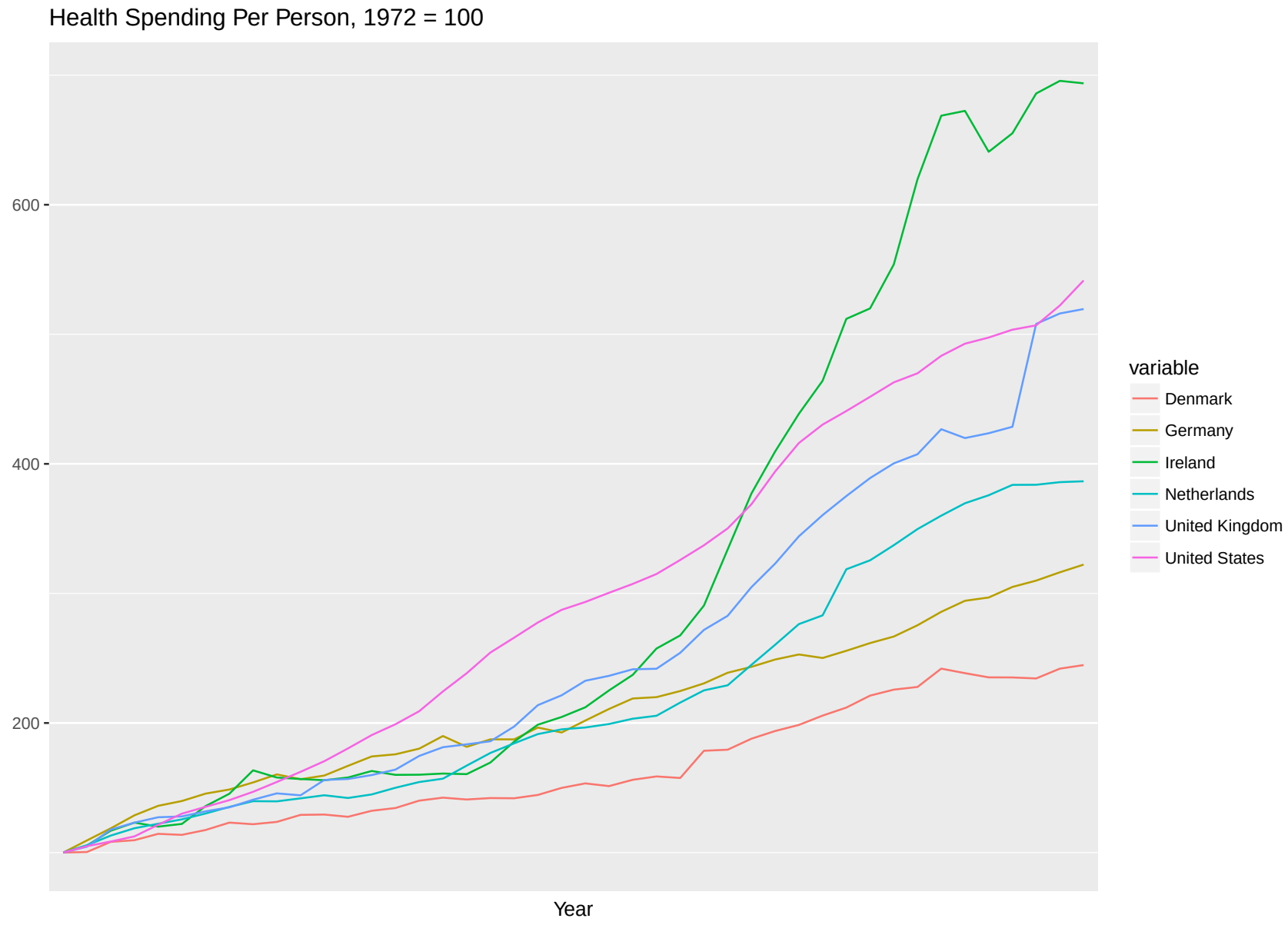
Logarithms

- in some cases the researcher does not work directly with a variable but with a transformed version of this variable
- The logarithm (to the base B) of a number, A, is the power to which B must be raised to give A. The notation for this is: $log_B(A)$.
- So, for instance, if B $= 10$ and A $= 100$ then the logarithm is 2 and we write log $(100) = 2$. This follows since $10^2 = 100$.
- We use logs because they help us truncate data and express growth rates.

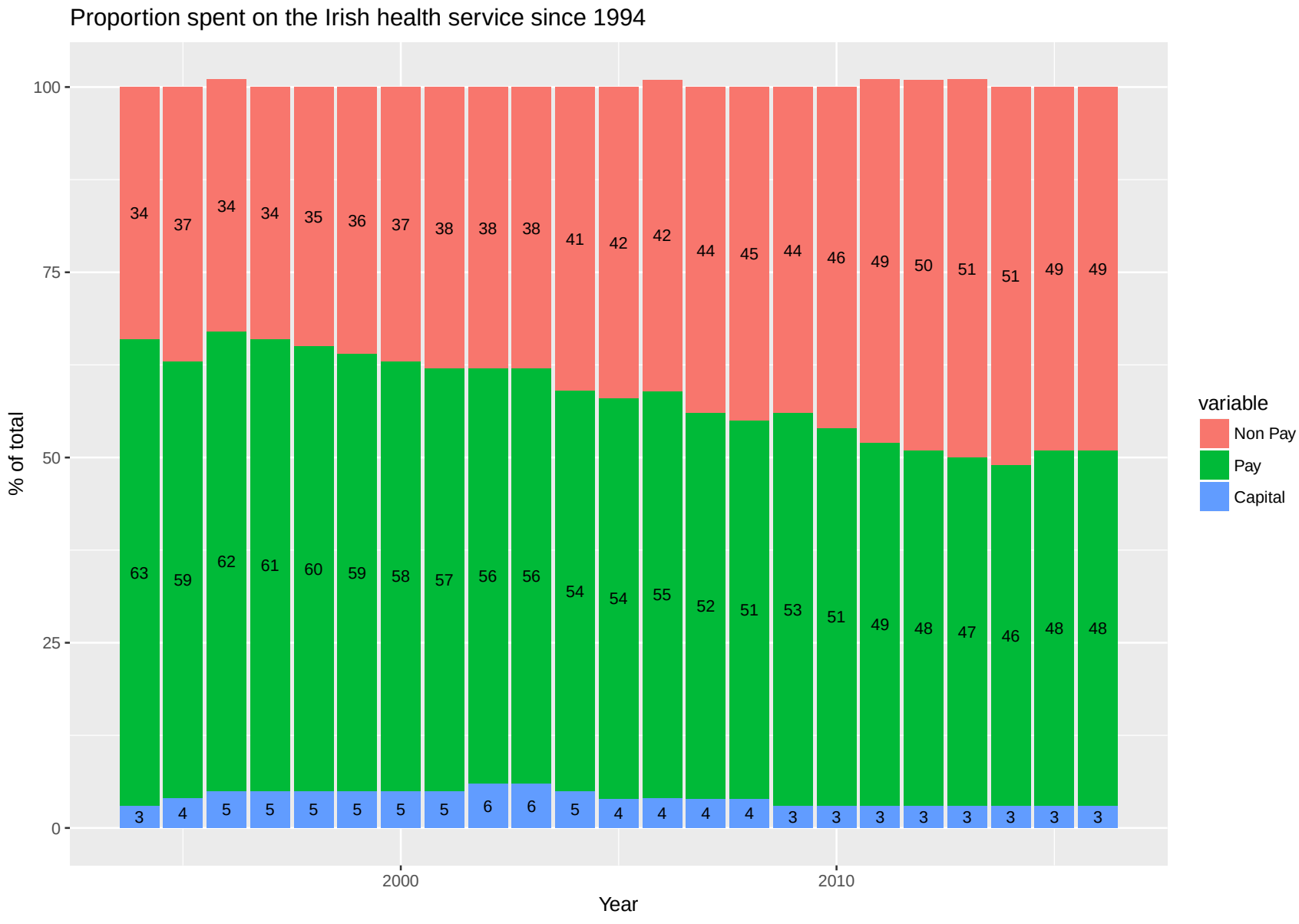
Levels vs rates

- Level: the actual reading. EG nominal GDP for Ireland in 2011 was €173,070 billion. Nominal GDP in 2012 was €175,754 billion.
- Rate: the change from 2011 to 2012 was (175-173)/173*100 $= 1.15\%$, more generally $(Y_{t+1} - Y_t)/(Y_t) * 100$

Index numbers: Very good at making time series data comparable to one another by choosing a base year.



Graphing proportional change over time



We're also interested in:

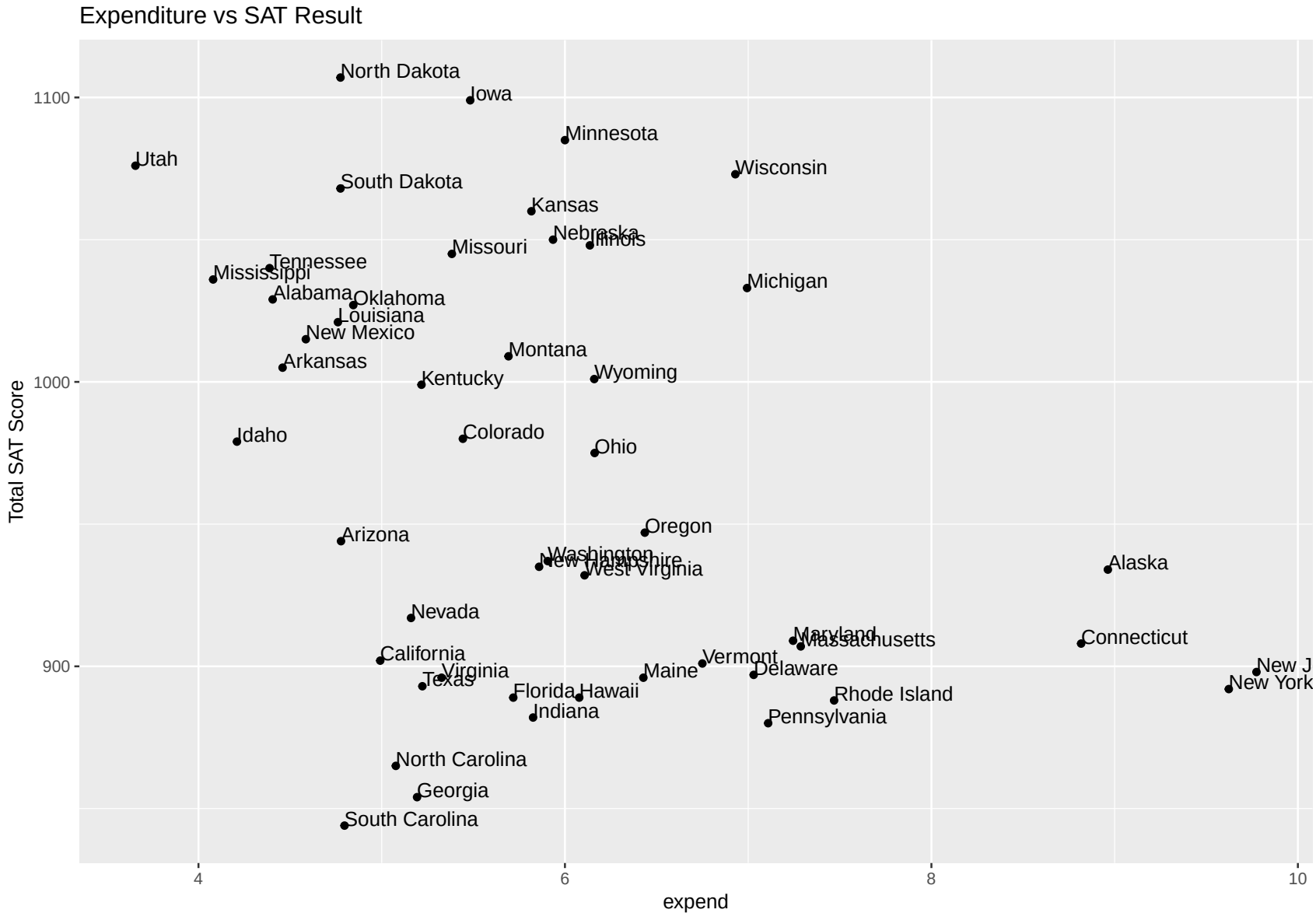
1. Identifying patterns in data
2. Classifying events
3. Untangling multiple causal influences
4. Assessing strength of evidence.

Example: does spending more on education improve outcomes? US Data

Table 1: Table continues below

state	expend	ratio	salary	perc
Alabama : 1	Min. :3.656	Min. :13.80	Min. :25.99	Min. : 4.00
Alaska : 1	1st Qu.:4.882	1st Qu.:15.22	1st Qu.:30.98	1st Qu.: 9.00
Arizona : 1	Median :5.768	Median :16.60	Median :33.29	Median :28.00
Arkansas : 1	Mean :5.905	Mean :16.86	Mean :34.83	Mean :35.24
California: 1	3rd Qu.:6.434	3rd Qu.:17.57	3rd Qu.:38.55	3rd Qu.:63.00
Colorado : 1	Max. :9.774	Max. :24.30	Max. :50.05	Max. :81.00
(Other) :44	NA	NA	NA	NA

Looking at the data

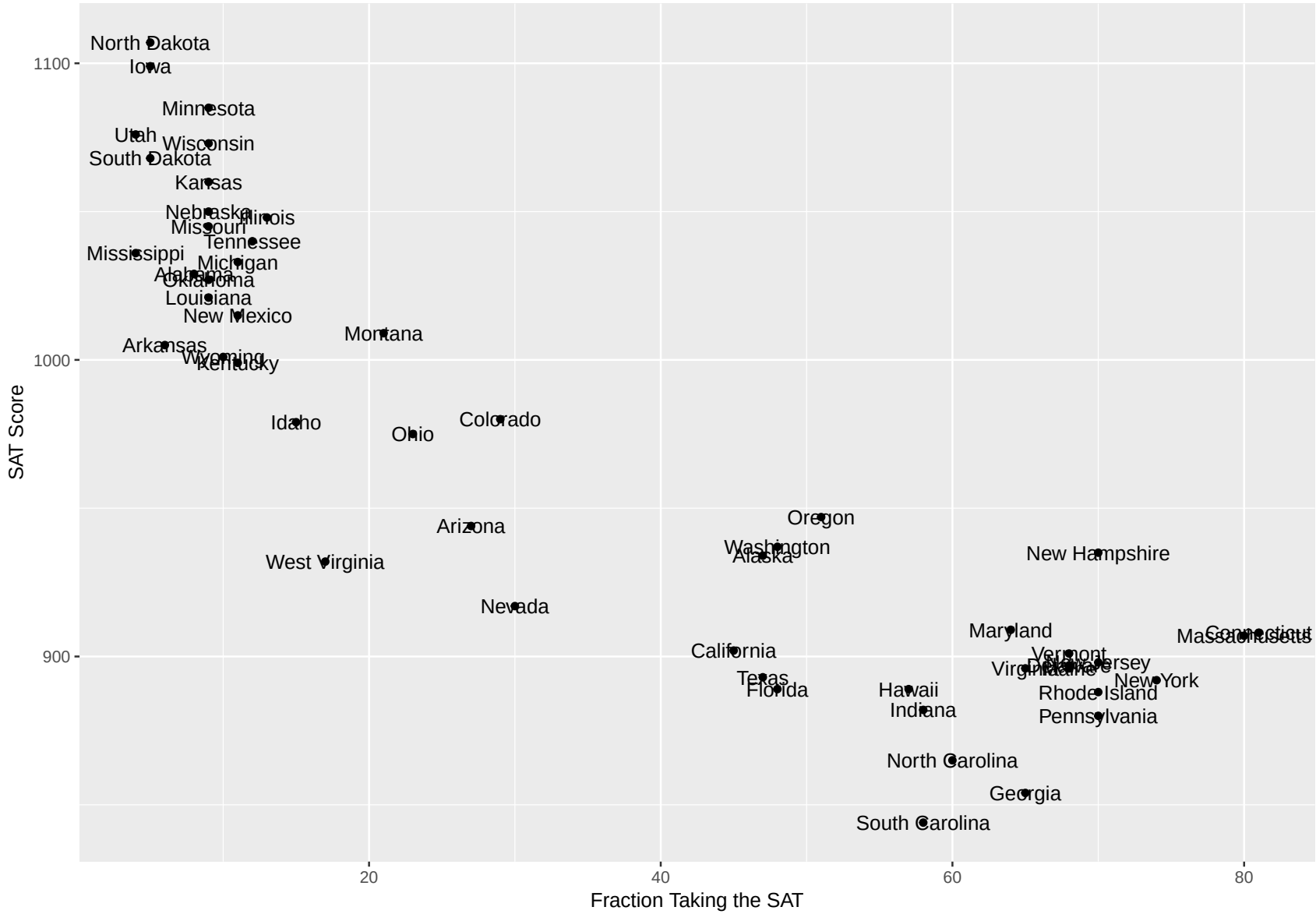


Careful interpreting this dataset!

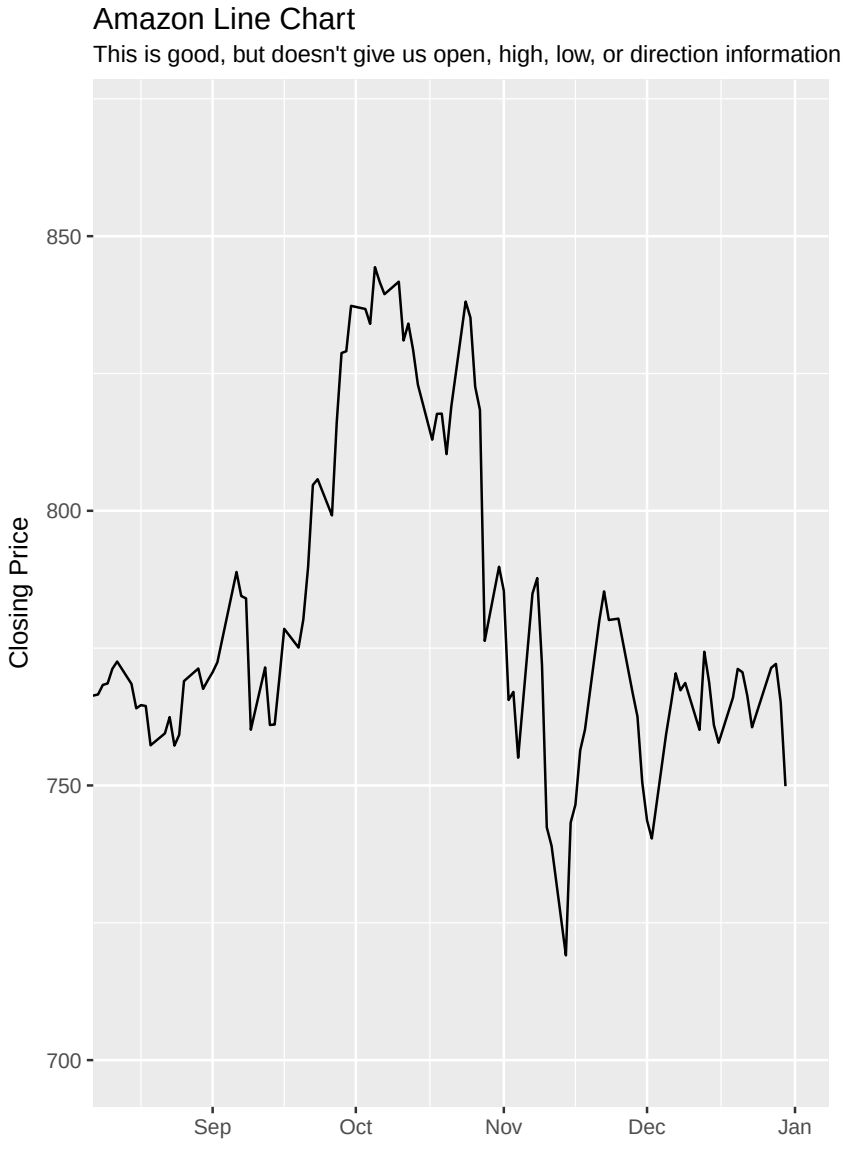
The data are telling us spending more reduces your SAT score. Something is clearly wrong. What?

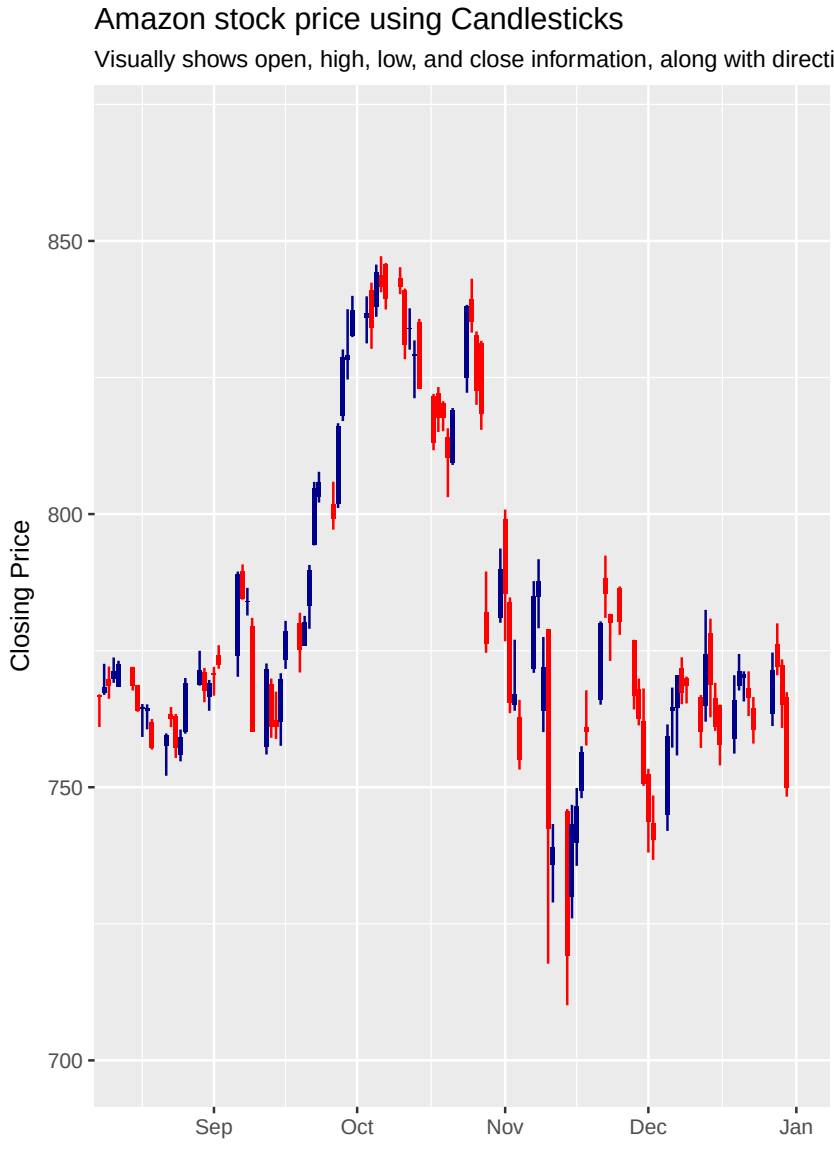
- Teacher salary?
- Religious ethos
- fraction of people taking the test?

Looking graphically at the fraction of the population that take the SAT



Different plots tell different stories. Example using financial data





A lot of the time we care about *causal* inferences

- right now think of causality as an 'if-then' statement
- EG: IF the state spends more on education, will exam results THEN go up?
- Which policies promote reductions in child mortality?
- Economic data often exhibits features not well described by the most basic statistical models – Nonlinear relationships, dependence between observations – Need statistical descriptions which take these features into account

How do you make causal inferences?

- They can come form observational studies, natural experiments, randomised controlled experiments, and more.
- Typically economic data come from observational studies. You observe household consumption going up when disposable income goes up.
- Observational data are almost always confounded, meaning there's a difference between the *treatment* and *control* groups. This is because people choose to be in one group or another and you can't control that ex ante, and this affects the response.

Example from Freedman, Statistical Modeling

for school children, shoe size is strongly correlated with reading skills. However, learning new words does not make the feet get bigger. Instead, there is a third factor involved - age. As children get older, they learn to read better and they outgrow their shoes. (According to statistical jargon (. . .), age is a confounder.) In the example, the confounder was easy to spot. Often, this is not so easy. And the arithmetic of the correlation coefficient does not protect you against third factors."

Terminology

- Medical terminology. One group gets a pill with an active chemical, another gets a sugar pill. If the chemical wasn't useful, you should see no difference in outcome between the two groups.
- A control is a subject who didn't get the treatment.
- A controlled experiment is when the experimenter gets to decide who goes in what group.

An early example: how we figured out smoking causes cancer

- Smoking causes heart attacks, lung cancer, and other diseases. How did we figure this out?
- Can we compare female smokers to male smokers? No, because gender is a confounder. So we have to compare male smokers to female smokers.
- Other confounders: age, education, etc.
- So only compare male smokers 55-69 to male non-smokers 55-69,etc.
- Continue to subset by urban/rural/etc
- Eventually confounding effects for smoking seem very, very implausible.

Slight problem

- As you continue to add more and more explanatory variables, you reduce the size of potential study groups, and so room gets bigger for chance effects.
- Randomised control experiments limit the potential for confounding.

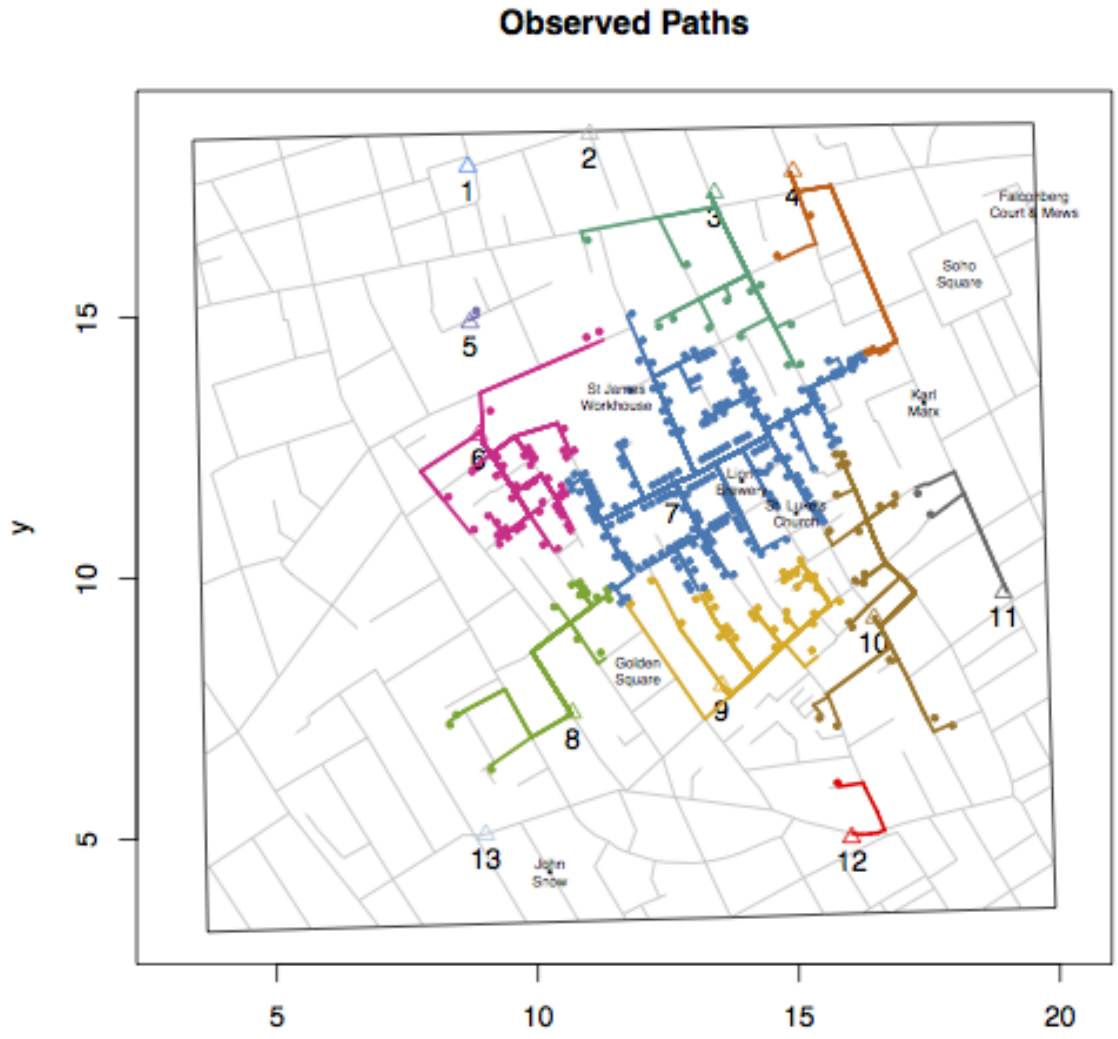
An even earlier example: John Snow and Cholera

Dr John Snow produced a famous map in 1854 showing the deaths caused by a cholera outbreak in Soho, London, and the locations of water pumps in the area.

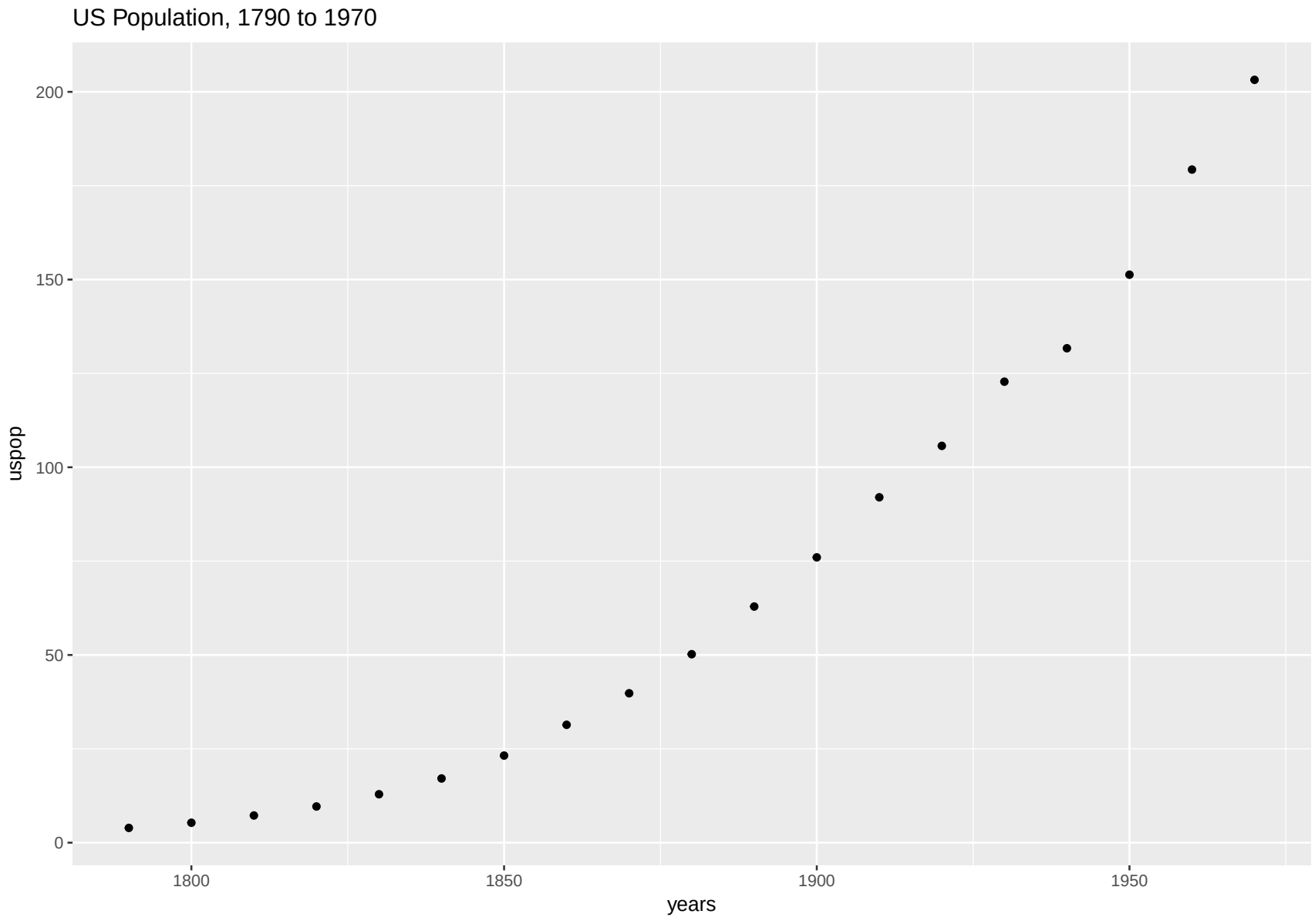
By doing this he found there was a significant clustering of the deaths around a certain pump – and removing the handle of the pump stopped the outbreak and invented epidemiology.

Dr Snow's Map

"The simple graph has brought more information to the data analyst's mind than any other device." — John Tukey



Thinking in terms of economic analysis: Population



Yule: What causes poverty?

- In the late 19th Century, Yule asked: what causes pauperism? Was it policy?
- He gathered data and ran the following regression. (Don't worry if you don't know what a regression is yet)

$$\Delta\text{Pauper} = a + b * \Delta\text{Out} + c * \Delta\text{Old} + d * \Delta Pop + error$$

- Δ means percentage change over time.
- Pauper is the percentage of paupers
- Out is the ratio of those Inside the workhouse to those Outside it.
- Old is the percentage of the population over 65
- Pop is the population

Data

- Yule had data from about 600 districts from 1871, 1881, and 1891.
- There were 4 regions (urban, rural, mixed, metropolitan), giving 8 equations each to be estimated.
- Yule fitted his equations by hand, determining the values of a, b, c, and d by minimising the sum of squared errors

$$\sum(\Delta Paup - a - b * \Delta \text{Out} - c * \Delta \text{Old} - d * \Delta Pop)^2$$

Yule's Results

The table shows some of Yule's 1899 results from table XIX of his classic study.

	Paup (a)	Out (b)	Old (c)	Pop (d)
Kensington	27	5	104	136
Paddington	47	12	115	111
Fulham	31	21	85	174

If you want to mess around with Yule's data in R, go to `https://github.com/jrnold/yule`

Interpreting the results

Consider the metropolitan unions. Fitting the data for 1871-1881, Yule obtained

$$\Delta\text{Paup} = 13.19 + 0.755\Delta\text{Out} - 0.022\Delta\text{Old} - 0.322\Delta\text{Pop} + \textit{error}$$

The interpretation of a coefficient like 0.755 is: other things being held constant, if ΔOut increased by 1 percentage point, meaning the administrative district supports more people outside the poorhouse–then ΔPaup goes up 0.755 percentage points.

This is a quantitative inference.

For 1881 to 1891, his equation was

$$\Delta\text{Paup} = 1.36 + 0.324\Delta\text{Out} + 1.37\Delta\text{Old} - 0.369\Delta\text{Pop} + \textit{error}$$

The coefficient of ΔOut being relatively large and positive, Yule concludes Out-Relief causes poverty. This is a qualitative inference.

Physics envy

Yule's idea was to uncover the 'social physics' of poverty. This is not so easily done. You have to be very, very careful when applying quantitative reasoning to real world problems. These regressions are extra pieces of information to aid decisions. They should not decide for you.

An example: it turned out that Yule's data did not consider the efficiency of the administration of the workhouses.

At best, Yule establishes association rather than causation.

To his great credit, Yule distanced himself from his findings and eventually suggested the authorities drop his measurements all together.

Looking at Yule another way–the modern way, using facets

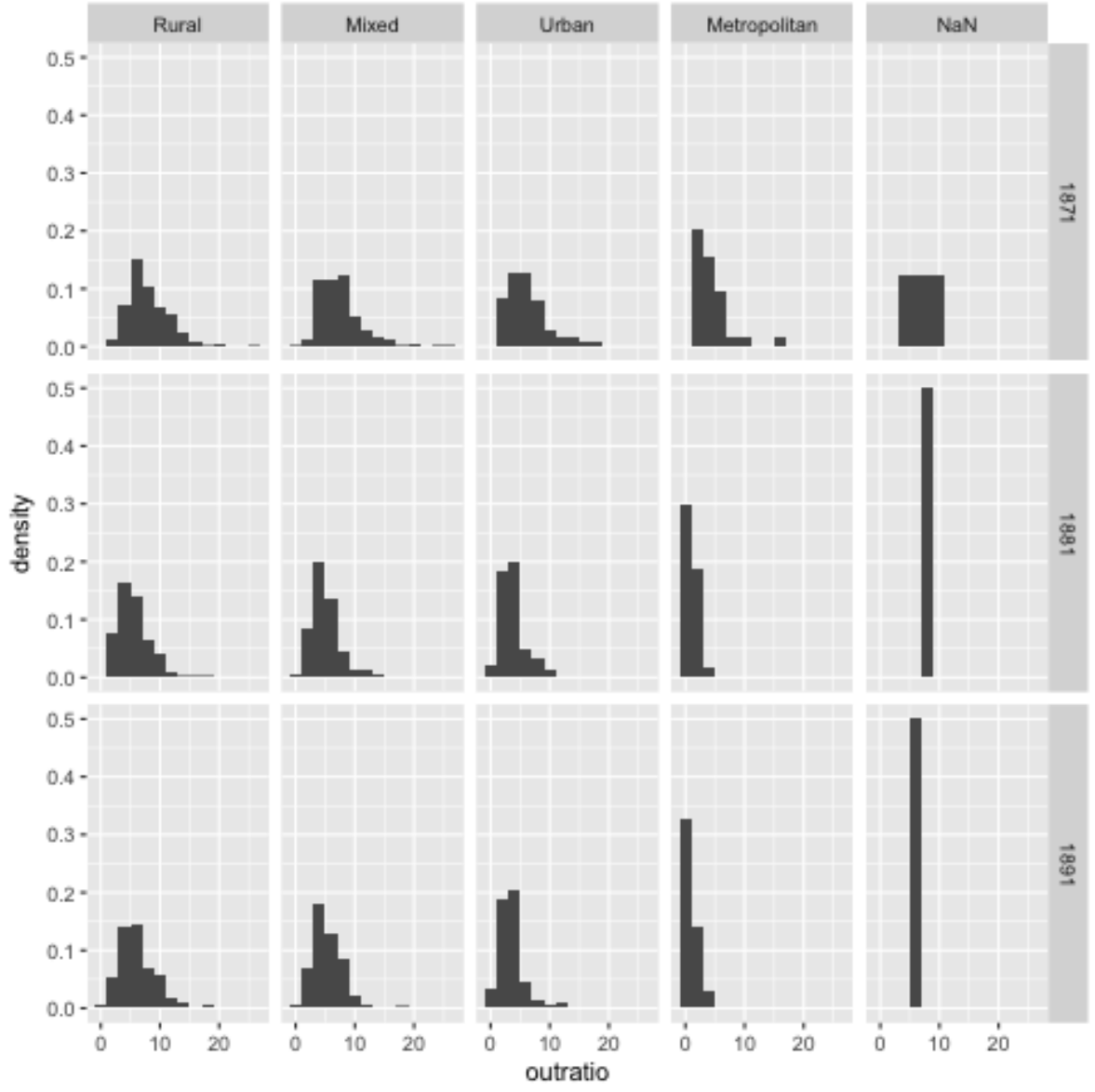


Figure 3: Created using `https://github.com/jrnold/yule`

Why causality matters

When you understand causality, you can start thinking about intervening in the system.

- Descriptive statistics and visualisation tell you about the data you happen to be able to measure.
- Causal models claim to tell you what will happen to some numbers if you change other numbers.
- Something has to remain constant in all the change.

Basic data we handle in economics

- Time series data. Data is collected at specific points in time.
- Cross sectional data. Units across individual data $(W_1, W_2, \ldots, W_n)$
- Categorical data: when answers are Yes/No, Male/Female, etc.
- Panel data (these have both a time series and a cross-sectional component).

Basic plotting/graphing we use to visualise these data

- XY plots/scatter plots
- Line plots (usually for time series)
- Histograms (for frequency)
- Maps
- Network models

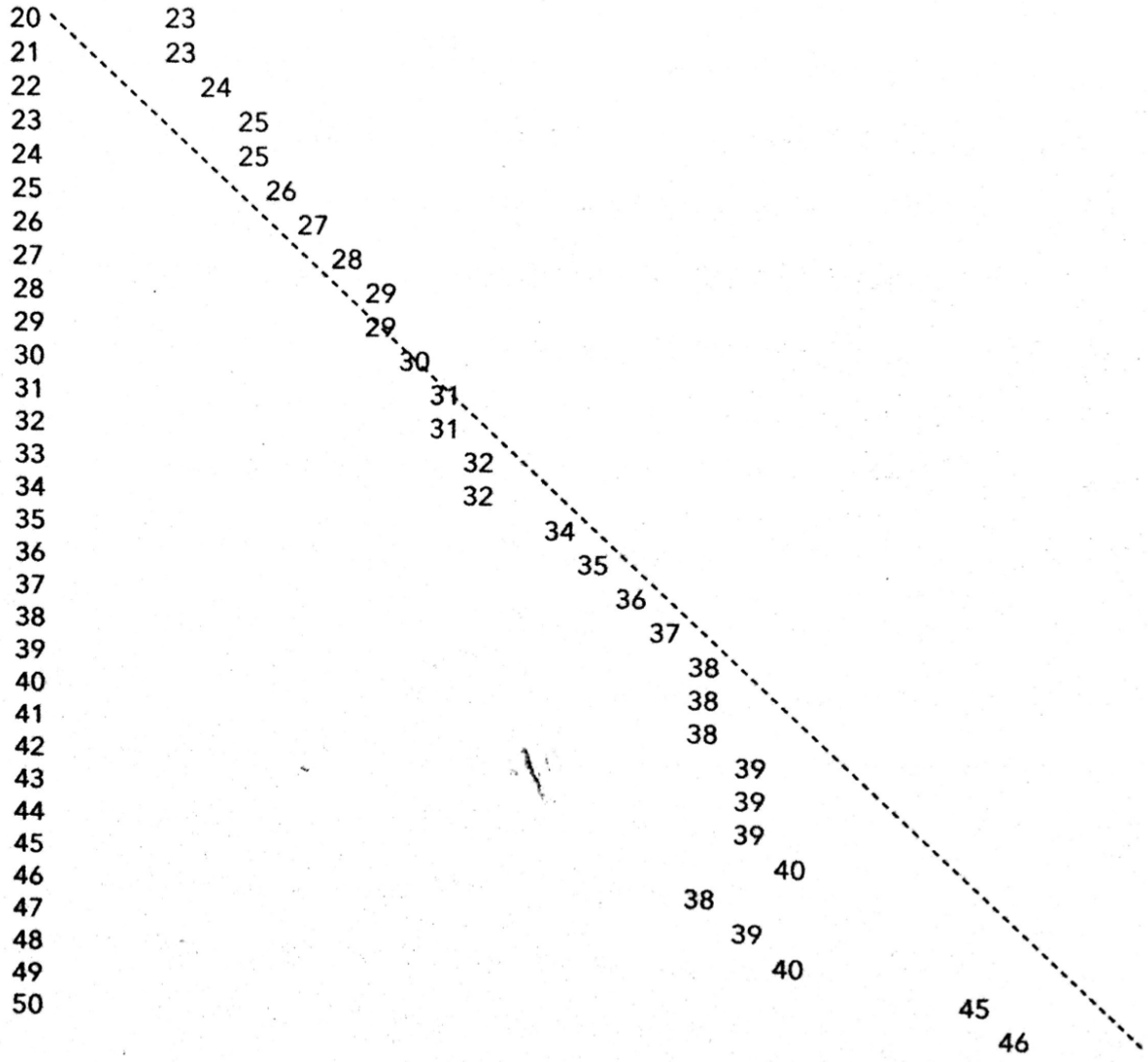
Lecture 2: Modeling (Freedman Chapter 1, Koop Chapter 4)

- Understanding correlation
- Why are variables correlated
- Staring at XY plots: men vs women
- Complexities when thinking about data analysis & modeling
- Example: Wage/Salary data from 1985. Loads of ways to think about why you'd like to be able to do applied economic analysis

Last time:

- Descriptive stats
- Causality
- Graphical models & inference

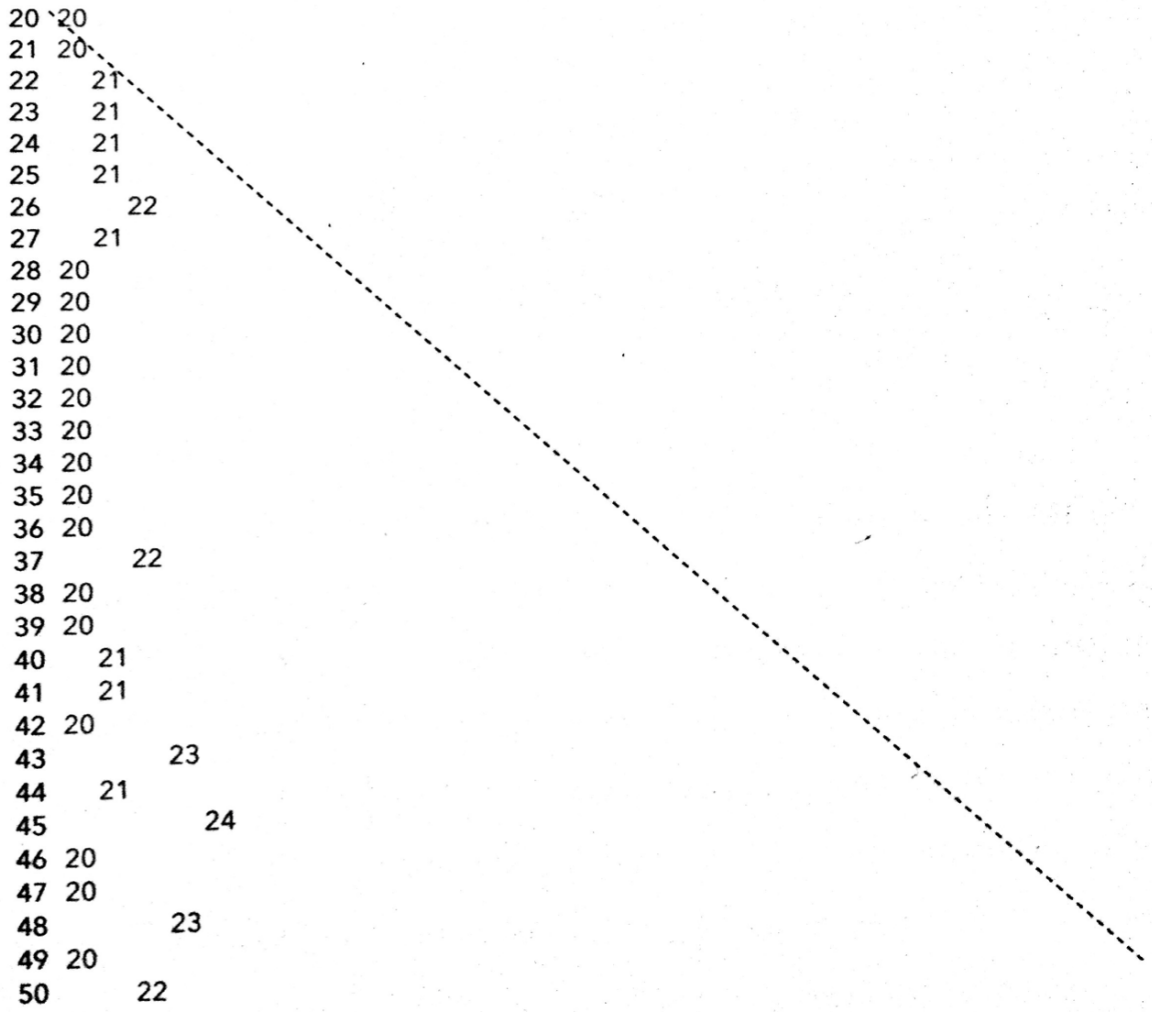
Dataclysm: a woman's age vs the age of the men who look best to her



Source: Christian Rudder, *Dataclysm*

Aaaand from Men

(a man's age vs the age of the wommen who look best to him)



Source: Christian Rudder, *Dataclysm*

Understanding the regression line: computed from five statistics

- the average of x, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- the standard deviation of x, square root of $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- the average of y, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- the correlation between x and y, $r = \frac{1}{n}\sum_{i=1}^{n}(\frac{x_i - \bar{x}}{s_x} * \frac{y_i - \bar{y}}{s)_y})$
- We're tacitly assuming s_x and s_y aren't zero.

An example

Figure 1. Heights of fathers and sons. Pearson and Lee (1903).

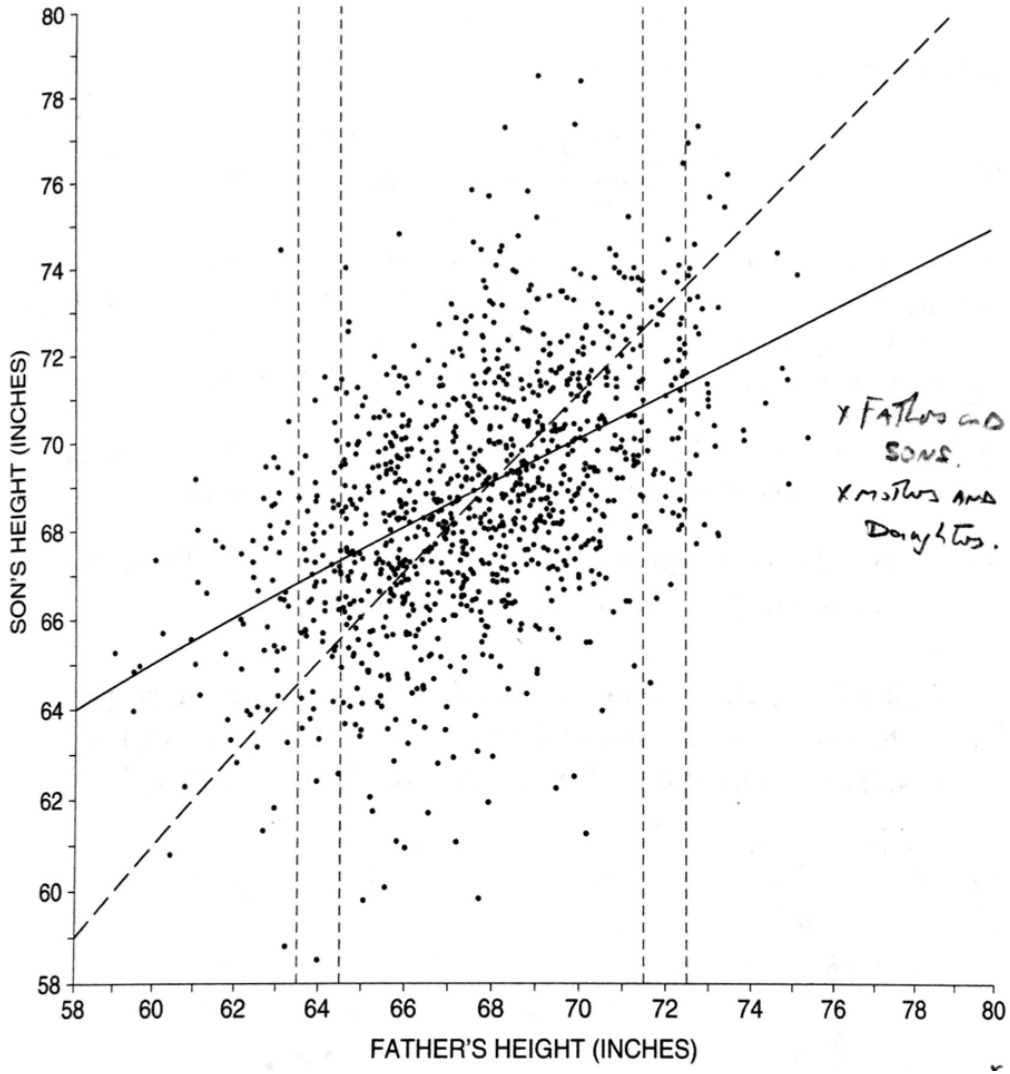


Figure 4: Source: Freedman, 2012

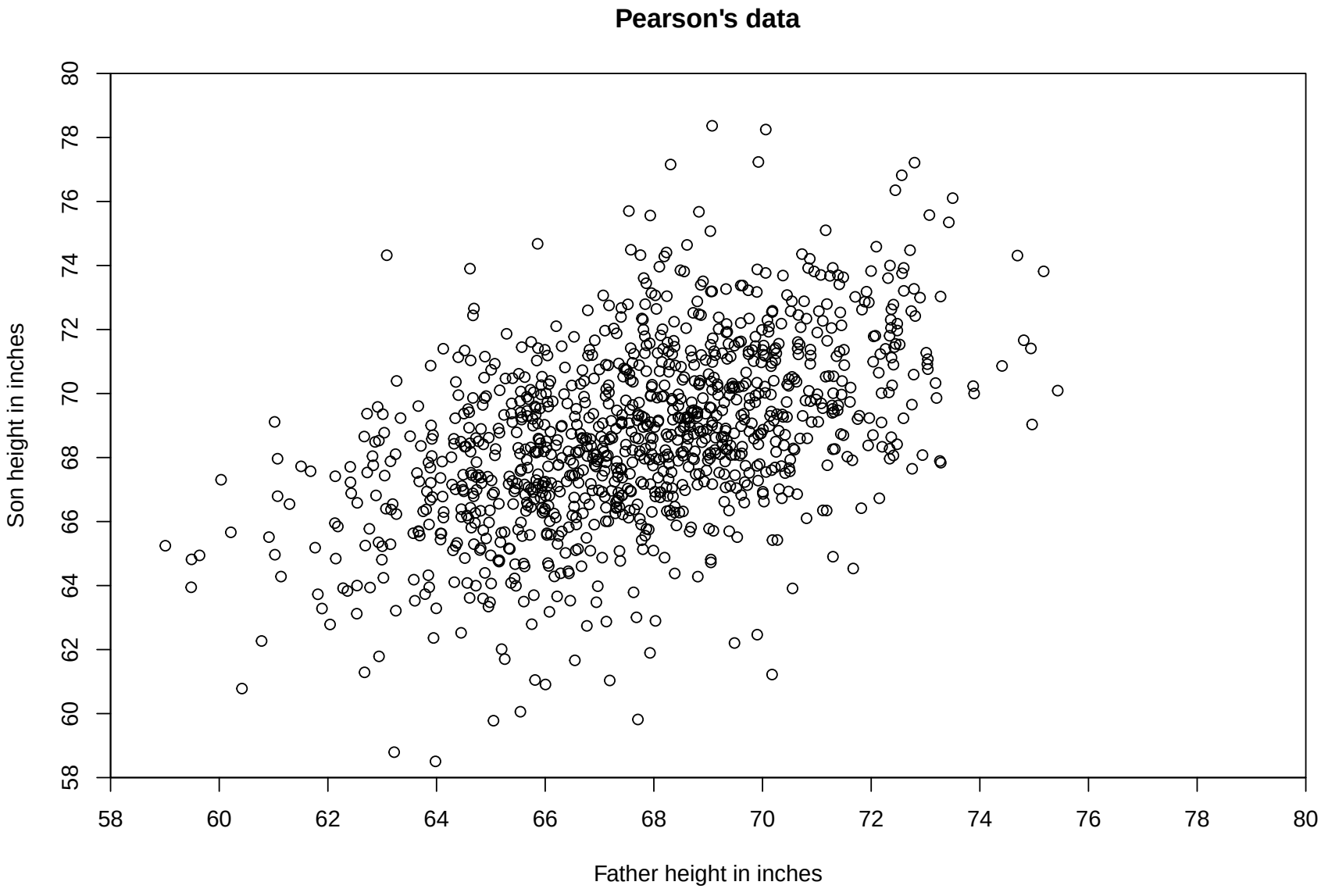
Understanding correlation

- Sir Francis Galton (1822-1911) made some progress on this while thinking about resemblance of parents and sons.
- Galton's student Karl Pearson (1857-1936) measured the heights of 1,078 fathers and their sons at maturity.
- Learn more at Roberto Bertolusso. Next few slides draw heavily on his excellent exposition.

Pearson's data look like this

```
##    fheight  sheight
## 1 65.04851 59.77827
## 2 63.25094 63.21404
## 3 64.95532 63.34242
## 4 65.75250 62.79238
## 5 61.13723 64.28113
## 6 63.02254 64.24221
```

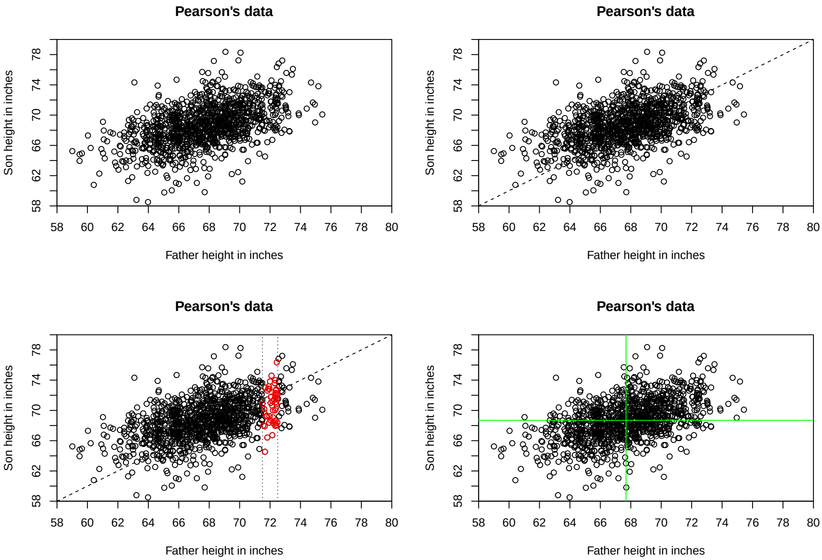
And this


Pearson's data
Son height in inches
Father height in inches
58
60
62
64
66
68
70
72
74
76
78
80

Interpreting this figure

- The scatter diagram above is a cloud shaped something like a rugby ball with points straggling off the edges
- Points in father and son's data slopes upward to the right (y-coordinates tending to increase with their corresponding x-coordinates).
- This is considered a positive linear association between heights of fathers and sons. In general, the data are saying that taller fathers imply taller sons.
- Let's draw a 45-degree line $y = x$ through the cloud of points. (What do you think it represents?)

Pearson with a 45 degree line



Intepreting these figures

- There is still a lot of variability in the heights of the sons within the cloud of points we identified.
- Knowing the father's height still leaves a *lot* of room for error for an individual father in trying to guess the his son's height
- When there is a strong association between two variables, knowing one helps significantly in predicting (guessing) the other. When there is a weak association, knowing one variable does *not* help much in guessing (predicting) the other.

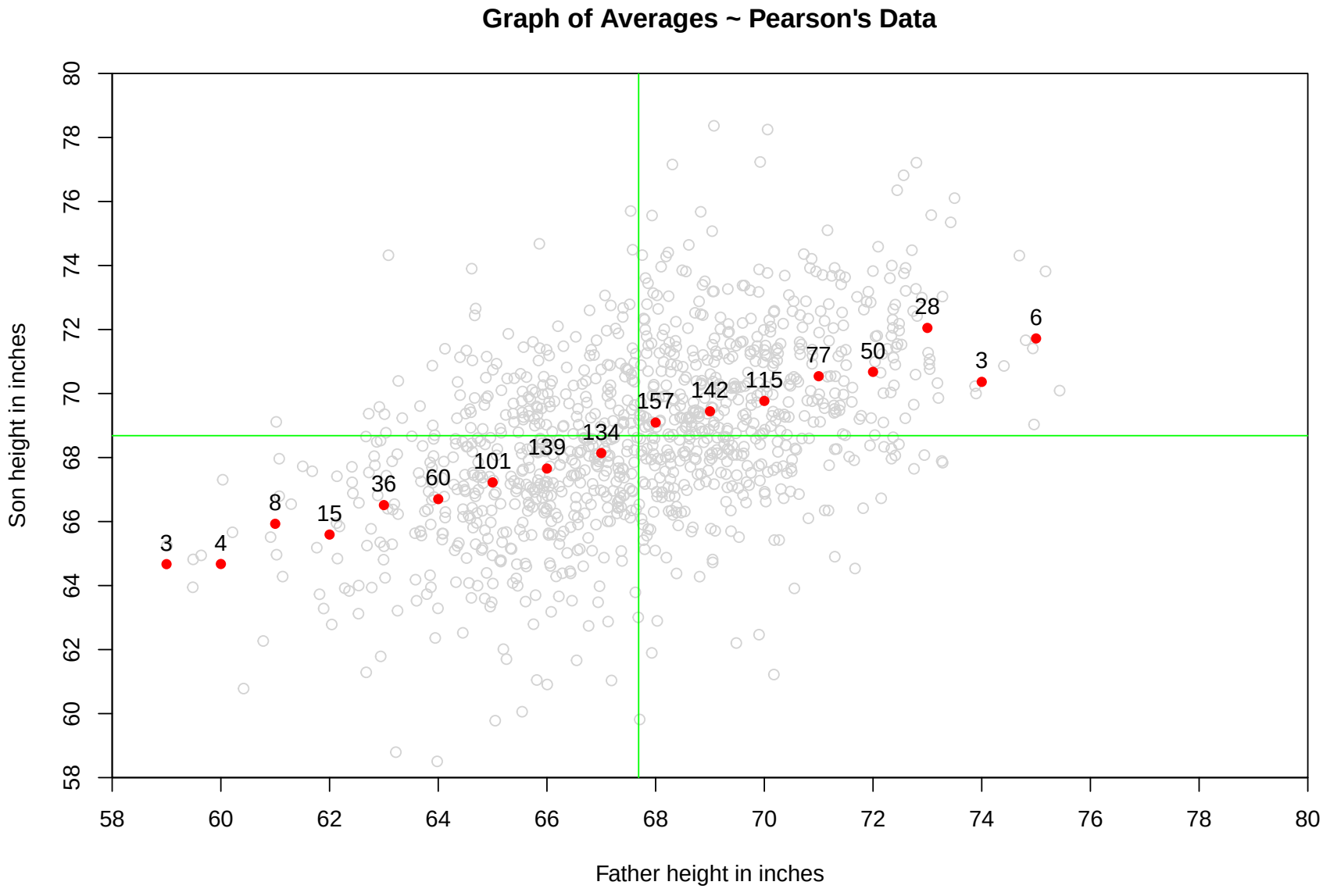
Interpreting these figures 2

- In social science (and other disciplines) studies of relationship between two variables, it is usual to call one *independent* and the other as *dependent*. You will see many different descriptions of these relationships in words. There is only one in maths.
- Usually too, the independent one is thought to influence the dependent one (rather than the other way around).
- In our example, father's height is considered independent, as in we think father's height influences son's height.
- However, we could use son's height as the independent variable. This would be appropriate if the problem were to guess a father's height from his son's height. Do you think this would be useful?

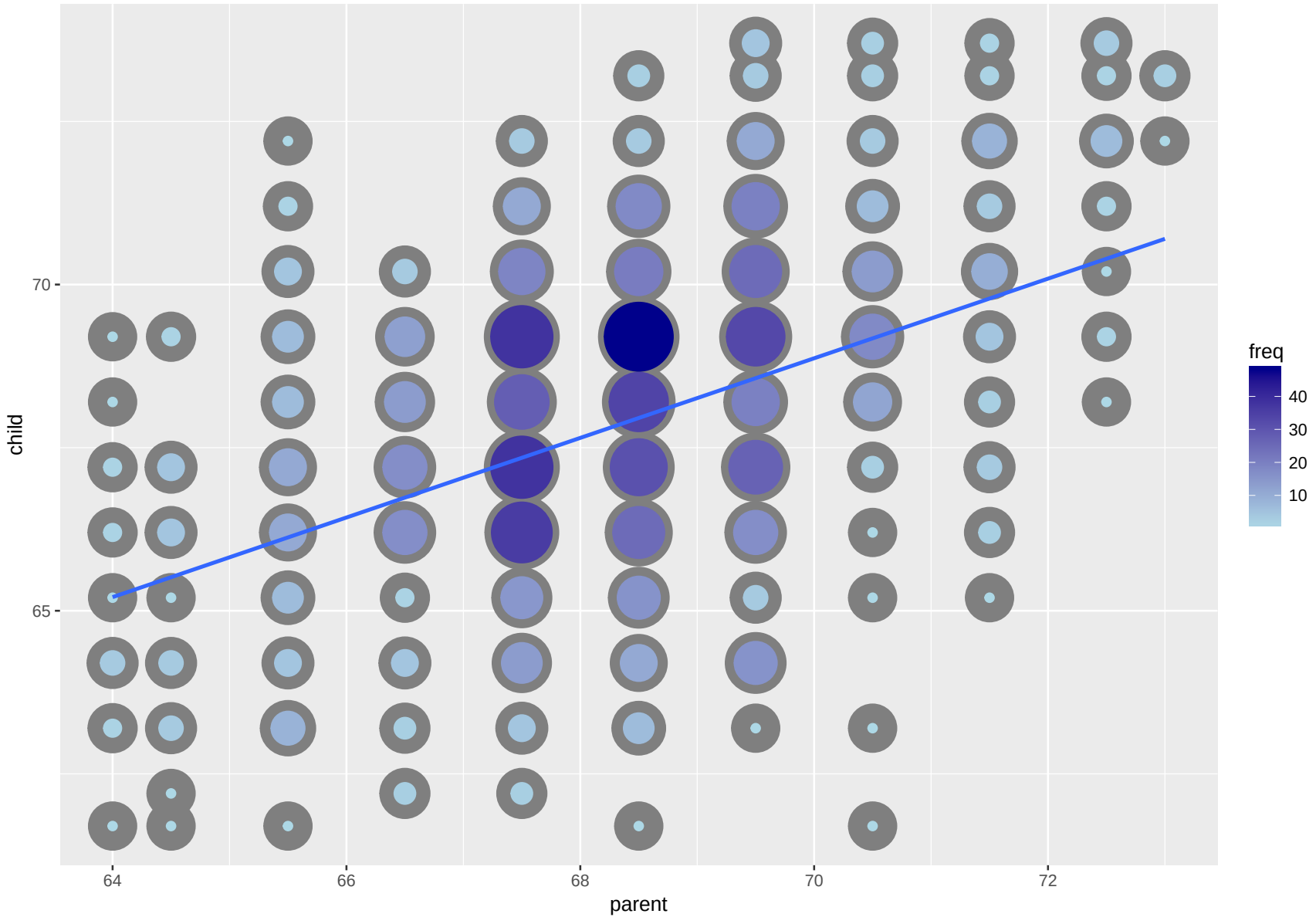
The regression line

- Think of the regression line, for predicting x from y, as a linear approximation to the 'graph of averages'. The graph of averages is the collection of points where the x-coordinate is the center of the vertical strip, and the y-coordinate the mean of all the y-values contained in that strip.

Pearson's regression line (graph of averages)



Another way to look at the data



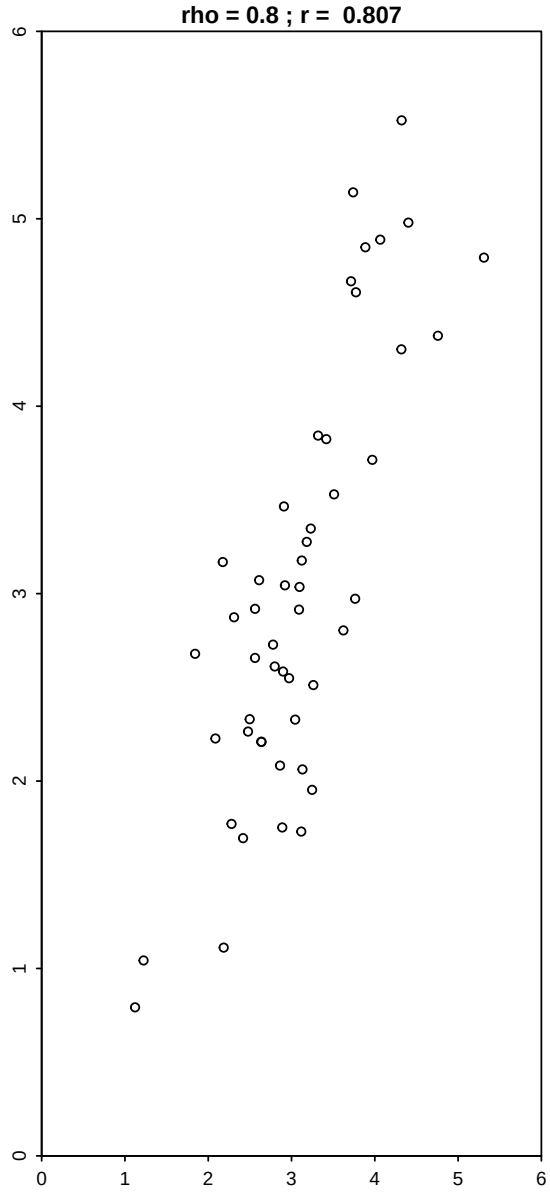
A subtle distinction

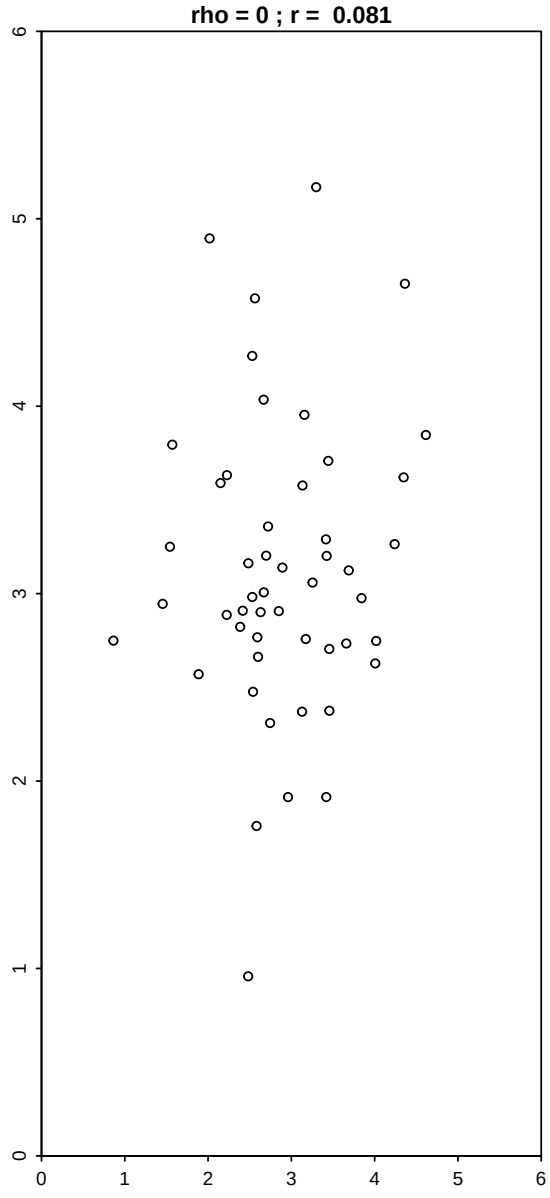
- Recall the distinction between association and causation
- Before graphing models like this was a few clicks, it made a lot of sense to summarise the count data using summary statistics.

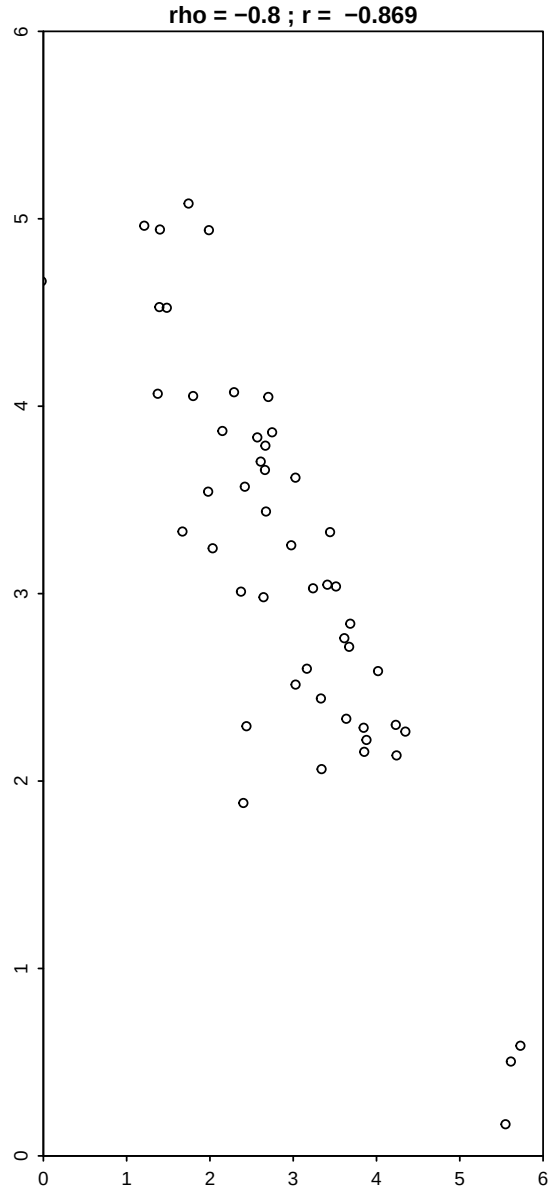
The Pearson correlation concept.

- It measures the extent to which the scatter diagram is packed in around a line.
- If the sign is positive, the line slopes up. If the sign is negative, the line slopes down.
- We measure the coefficient (ρ) as the Covariance of X and Y divided by the variance of X times the variance of Y. This is for a population.
- Important distinction between *population* and *sample*. For a sample, you use a more complicated formula but the interpretation remains the same. In social science we work with the sample coefficient, r.

Have a look at these figures, what do you think the correlation is?







Root Mean Square Error

- If you use the line $y = a + bx$ to predict x from y, the error or *residual* for subject i is $e_i = y_i - a - bx_i$, and
- The MSE is $\frac{1}{n}\sum_{i=1}^{n} e_i^2$.
- Gauss: Among all lines, the regression line has the smallest mean square error.

A regression model for Hooke's law.

A weight is hung on the end of a spring whose length under no load is a. The spring stretches to a new length. According to Hooke's law, the amount of stretch is proportional to the weight. If you hang weight x_i, on the spring, the length.

- $Y_i - a + bx_i = \epsilon_i$ for 1,. . . , n.
- In this equation, a and b are constants that depend on the spring. The values are unkwown and have to be estimated from data.
- The ϵ_i are independent, identically distributed, mean 0, varriance σ^2.
- You choose x_i, the weight on occasion i. The response Y_i is the length of the spring under the load. You do not see a, b, or ϵ_i.

Objects for statistical modeling

You need 3 things to get a model working. Again, you'll typically want to use a model to predict, or account for, some variable.

- Formulas. These relate variables to one another. They are causal statements. EG: WAGE $\sim$ EXPERIENCE $+$ GENDER. This says your wage is explained (we think) by the number of years of experience you have, and your gender. The squiggle yoke is called 'tilde'.
- Data frames–a collection of variables. Each variable gets a column, this column gets a name. The rows are cases (sometimes called elements).
- Functions. These are the building blocks of models and produce the outputs of the models. You need formulas and data frames to make functions work effectively.

Example: Wage/Salary relationships

- Let's model the relationship between wage and experience.
- Why would we think about these particular variables affecting the wage?

–We might think that $WAGE = a + b * EXPERIENCE + \epsilon$

or maybe

–$WAGE = a + b * EXPERIENCE + c * EDUCATION + \epsilon$

Data look like this

```
head(CPS85)
```

```
##    wage educ race sex hispanic south married exper union
## 1   9.0   10    W   M       NH    NS Married    27   Not
## 2   5.5   12    W   M       NH    NS Married    20   Not
## 3   3.8   12    W   F       NH    NS  Single     4   Not
## 4  10.5   12    W   F       NH    NS Married    29   Not
## 5  15.0   12    W   M       NH    NS Married    40 Union
## 6   9.0   16    W   F       NH    NS Married    27   Not
```

Model: Fitting WAGE = intercept + b*EXPERIENCE* + *c* EDUCATION + error

	Model 1	Model 2	Model 3
(Intercept)	-4.904***	-4.904***	-4.770
	(1.219)	(1.219)	(7.043)
exper	0.105***	0.105***	0.128
	(0.017)	(0.017)	(1.156)
educ	0.926***	0.926***	0.948
	(0.081)	(0.081)	(1.155)
age			-0.022
			(1.155)
R-squared	0.2	0.2	0.2
N	534	534	534

Thinking about interpreting the formula

The formula is a bit like a sentence. EG WAGE $\sim$ SECTOR is equivalent to

1. WAGE as a function of SECTOR
2. WAGE accounted for by SECTOR
3. WAGE modeled by SECTOR
4. WAGE explained by SECTOR
5. WAGE given by SECTOR
6. WAGE broken down by SECTOR

Conclusion

- Understanding correlation
- Why are variables correlated
- Staring at XY plots
- Complexities
- Example: Wage/Salary data from 1985.

Loads of ways to think about why you'd like to be able to do applied economic analysis

Lecture 3

Last time

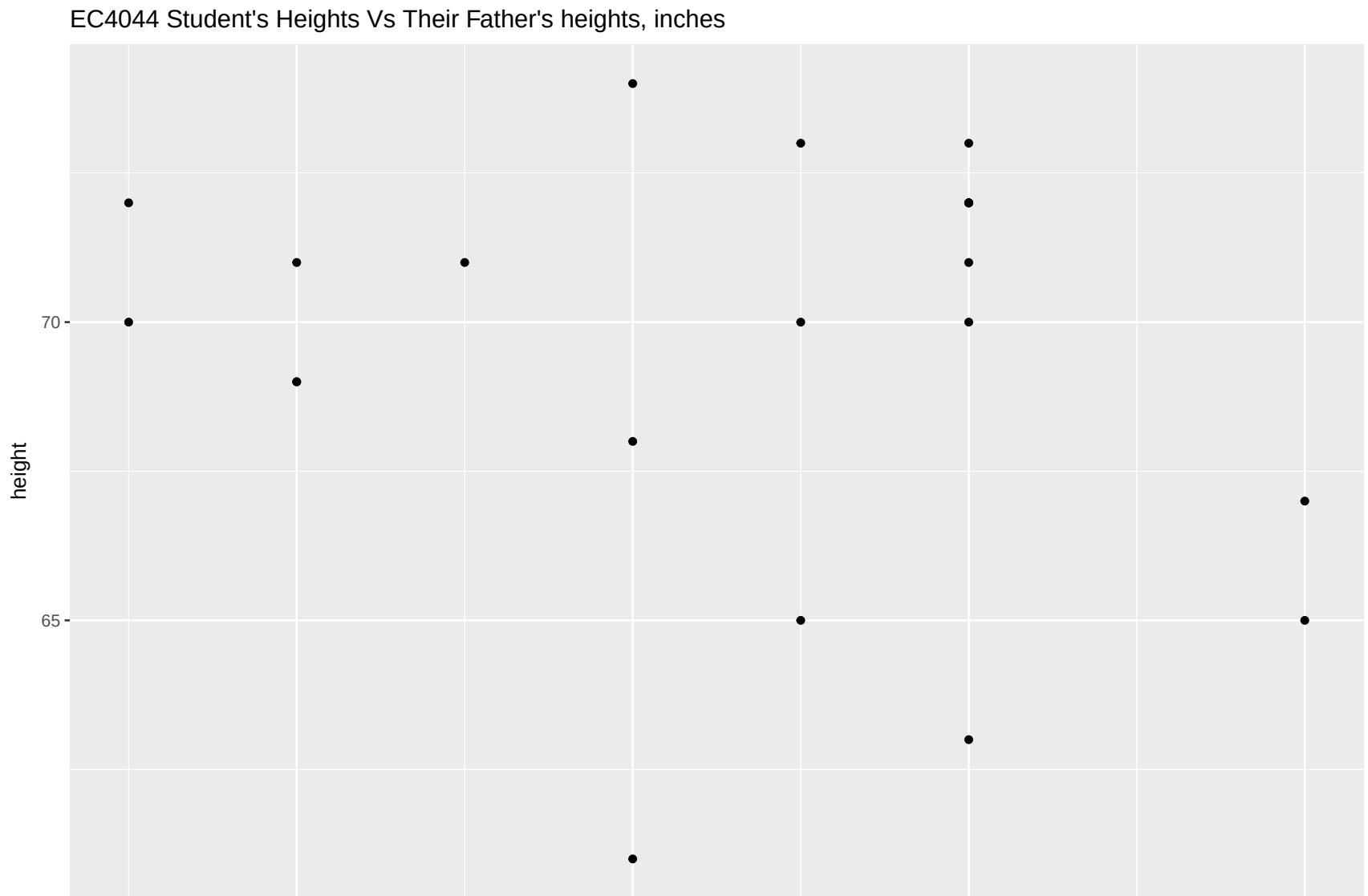
- Correlation and association
- Remember: r measures linear association, not association in general.
- Best fitting line
- Interpreting OLS estimates
- Measuring the fit of a regression model

This time

- Nonlinearity in Regression
- Factors affecting β
- Calculating confidence intervals for β
- Example: regression by hand, roll your own betas using R.
- Readings: Koop cht 4 & Freedman cht 3

EC4044 Galton

You have a factor. You can convert it to a character vector, split on the foot and inch symbols, and then use sapply to do the conversion in an anonymous function.



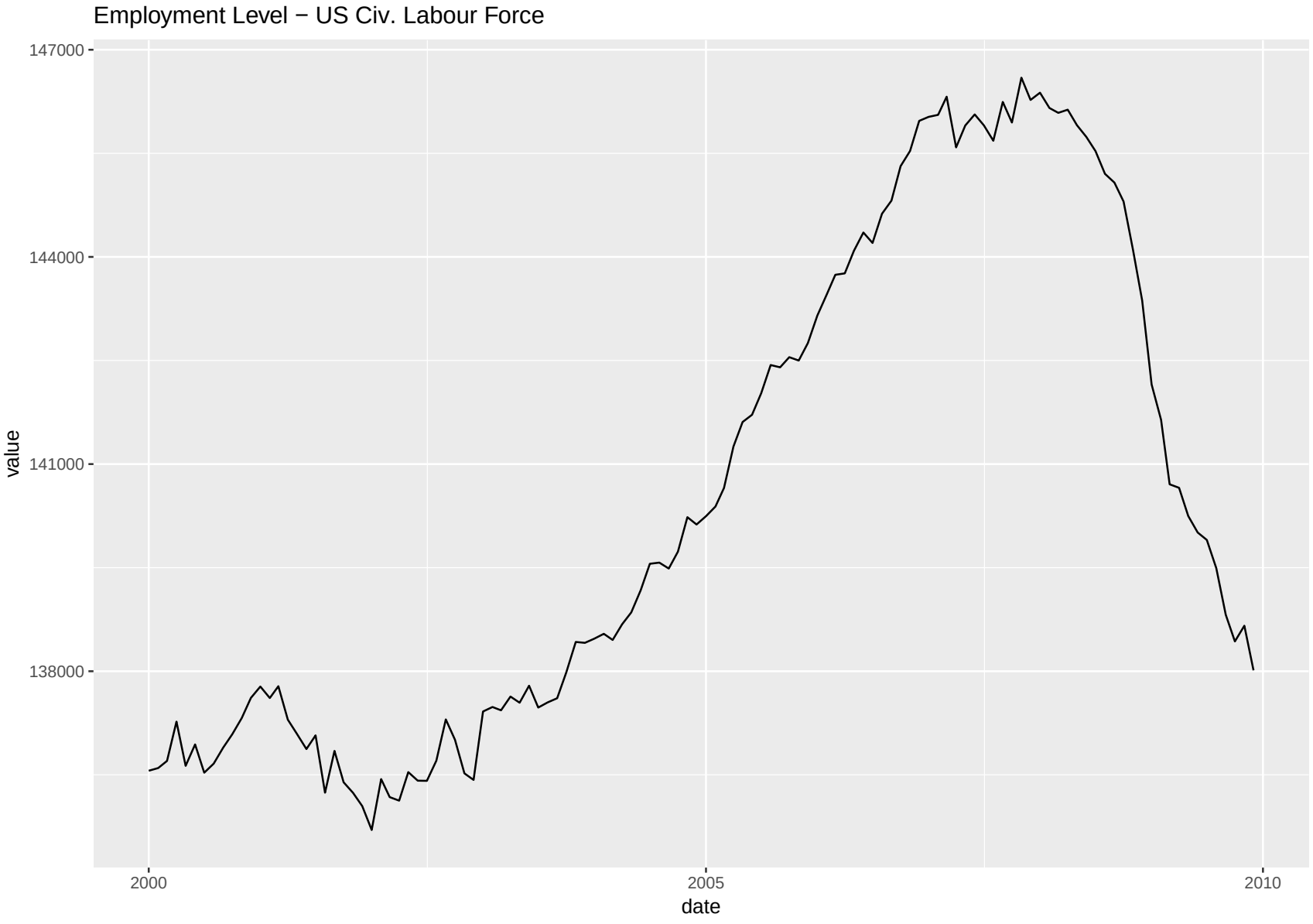
Concept of regression

- You think there's some relation between X and Y like $Y = \alpha + \beta X$. X is the independent variable, Y is the dependent variable, α, β are coefficients.
- It is common to implicitly assume that the explanatory variable "causes" Y, and the coefficient β somehow measures the influence of X on Y.
- Humility is vital here. The linear regression model will always be only an approximation of the true relationship
- The data might be totally crap, introducing model error everywhere.
- In economics, the most probable source of error is due to missing vari- ables, usually because we cannot observe them.

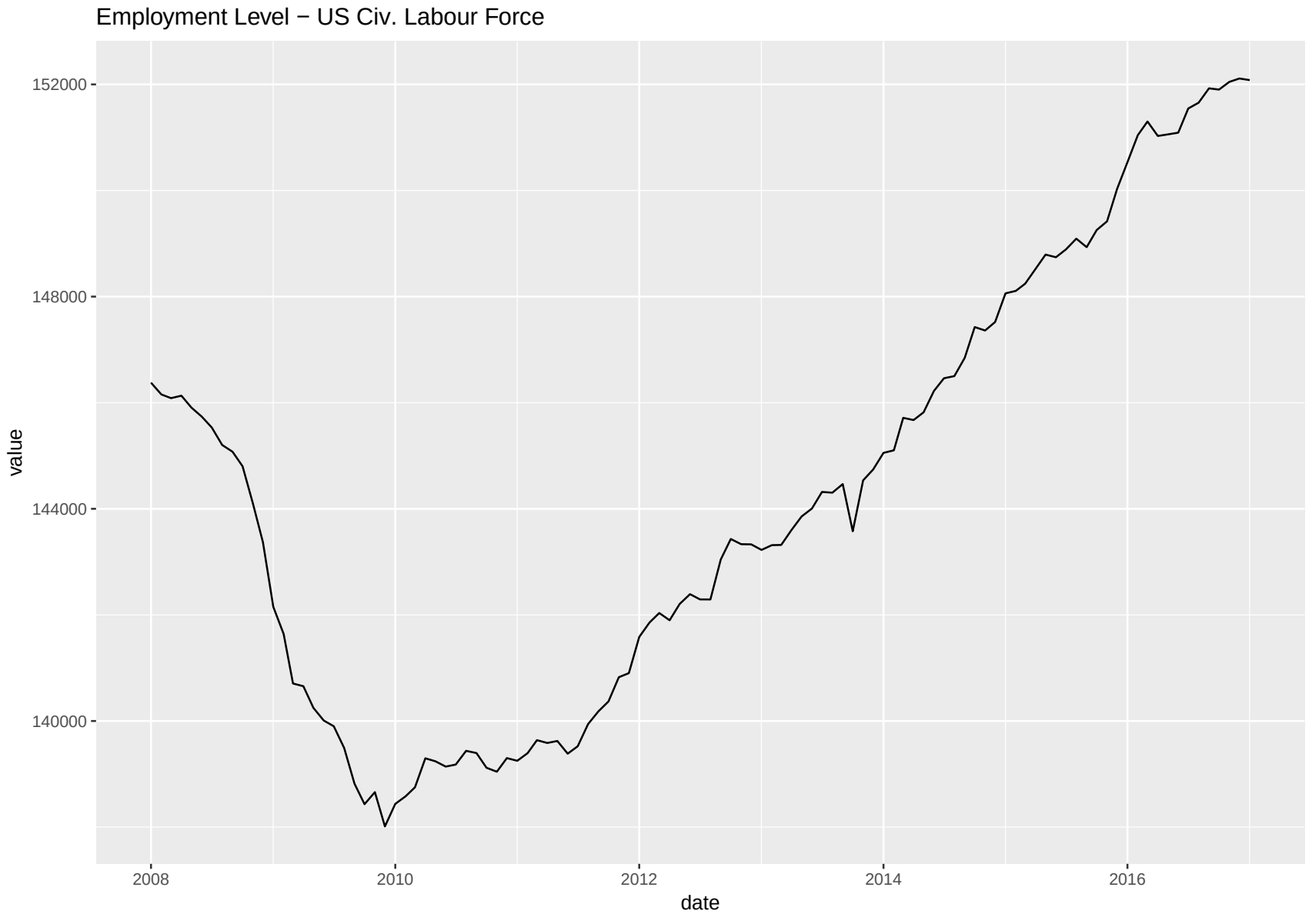
The Econometric Approach

- An economic **model** describes behavior of economic **variables** when world is governed by a specific structure, described by a set of **parameters**
  - Parameters allow us to find causal effects.
- Goal is to use data to learn parameters.
- Separate into steps
  - **Identification**: Supposing we know exactly how certain variables behave, what would we then know about parameters
  - **Estimation**: Find some function of the data that tells us about parameters
  - **Inference**: What values of parameters (if any) are plausibly consistent with the data?

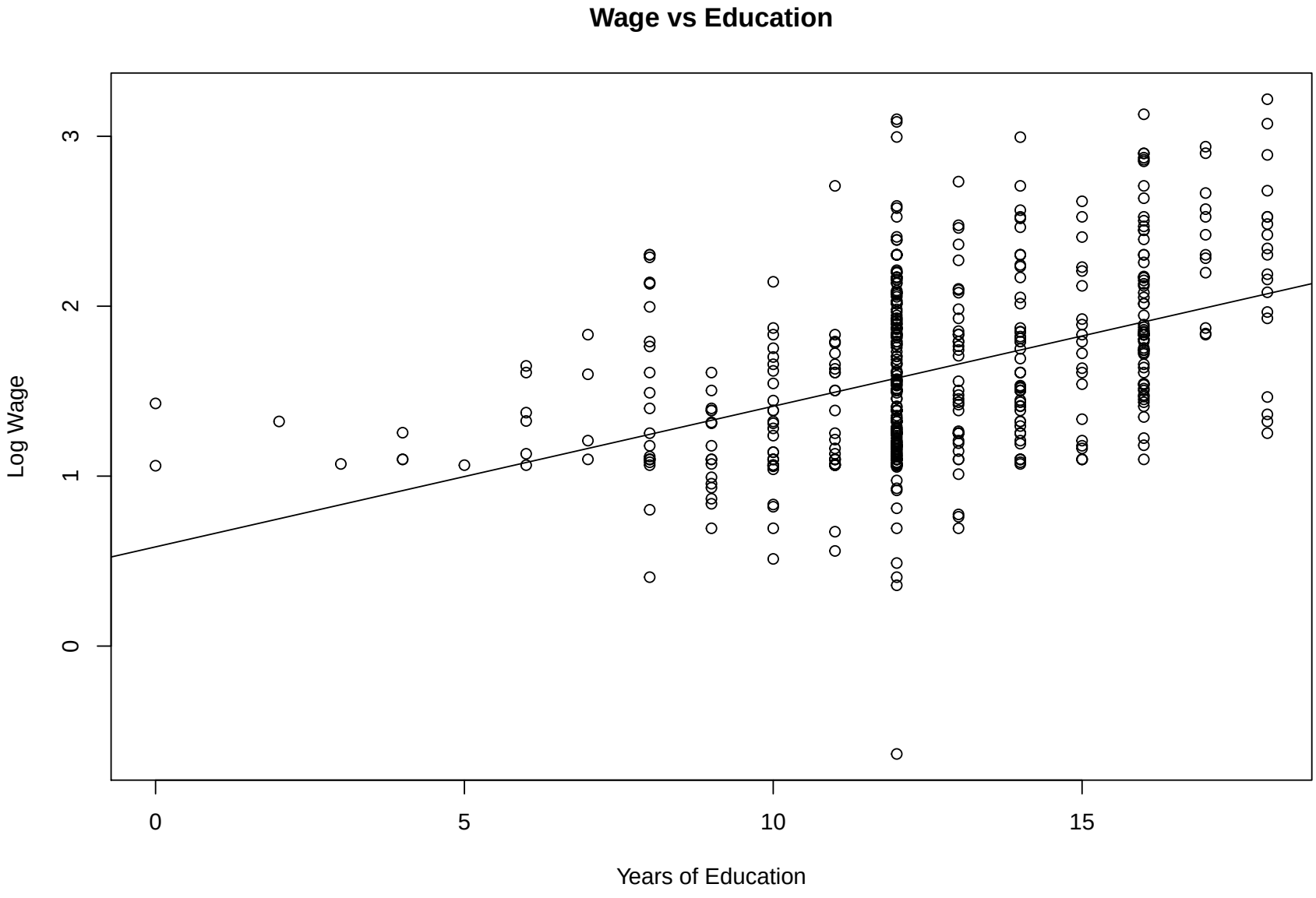
Example: US Employment Data. How would you model this?



What about this? How would you model this?



We could look at wage/education data as before, this time fitting a line to a cloud of points.

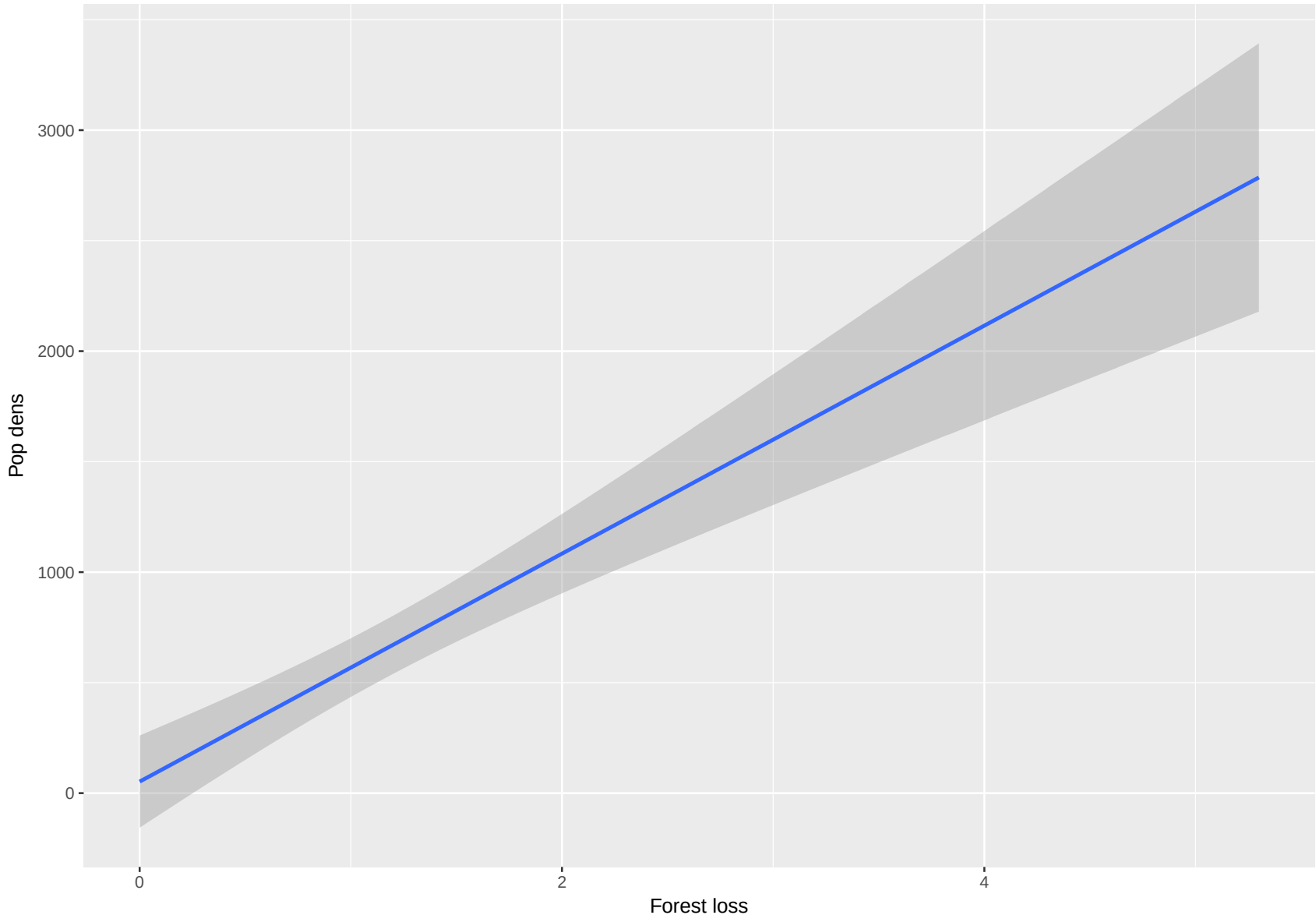


Regression results–what do they mean?

```
wageregoutput <- lm(formula = lwage ~ educ, data = wage1)
wageregoutput$coef
```

```
## (Intercept)        educ
##  0.58377263  0.08274437
```

From Koop: Deforestation vs Population Density



Regression results: how would you interpret these?

```
forestreg <-lm(forest$`Forest loss`~ forest$`Pop dens`, dat
pander(forestreg$coef)
```

(Intercept)	forest$Pop dens
0.6	0.0008423

Interpreting β and $\hat{\beta}$

- This coefficient is the slope of the best fitting straight line through the XY-plot. Mathematically $\beta = dY/dX$.
- $\hat{\beta}$ interpreted as the marginal effect of X on Y and is a measure of how much X influences Y.
- In the deforestation/population density example, $\hat{\beta}$ was positive (0.000842), so Population Density and Deforestation are positively correlated.
- We can interpret β as a measure of how much Y tends to change when X is changed by one unit.
- See this post on a way to create regressions from scratch in R.

Interpreting β and $\hat{\beta}$

- In the deforestation/population density example we obtained $\hat{\beta} = (0.000842)$. This is a measure of how much deforestation tends to change when population density changes by a small amount.
- Since population density is measured in terms of the number of people per 1,000 hectares and deforestation as the percentage forest loss per year, this figure implies that if we add one more person per 1,000 hectares (i.e. a change of one unit in the explanatory variable) defor- estation will tend to increase by 0.000842%.
- Important: regressions measure tendencies in the data.

Regression as statistical model

- You can run OLS on any data set.
- Under some quite strict assumptions (Koop, chapter 4) it tells us true features of the population

1. In population, $y = \beta_0 + \beta_1 x + u$
2. $(x_i, y_i) : i = 1 \ldots n$ are independent random sample of observations following 1
3. $x_i : i = 1 \ldots n$ are not all identical
4. $E(u|x) = 0$
5. $Var(u|x) = \sigma^2$ a constant > 0

Regression properties BLUE

- Under above assumptions, OLS estimator is

1. Consistent: $Pr(|\hat{\beta}_1 - \beta_1| > e) \to 0$ as $n \to \infty$ for any $e > 0$
2. Unbiased: $E(\hat{\beta}_1) = \beta_1$
3. Asymptotically normal

$$Pr\left(\frac{\sqrt{n}(\hat{\beta}_1 - \beta_1)}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} < t\right) \to Pr(Z < t)$$

for any t, where $Z \sim N(0, 1)$

- (In practice can replace σ^2 by $\frac{1}{n-2}\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$)

Interpretation requires humility/skepticism!

- Even when assumptions are satisfied, causal question not answered by above results
- Using the education example, how much would wages change if I stayed in school one more year?
- Maybe education raises wages, or vice versa, or both related to some third factor
- Regression alone won't tell us this
- In practice, above assumptions often dubious even as statistical descriptions
- More powerful statistical methods which better describe relationship can help, but still don't answer causal question.

Test yourself: An exercise with Rstudio

- Get the Forest.XLS data from SULIS. Import it into RStudio.
- Run a regression of deforestation on population density and interpret the results–just the coefficients. The significance tests etc we'll look at later.
- Run a regression of deforestation on change in cropland and one of deforestation on change in pasture land. and interpret the results.
- Create a new variable, V, by dividing population density by 100. What are the units in terms of which V is measured?
- Run a regression of deforestation on V. Compare your results to those for the first regression.

Measuring the 'fit' of a regression

- We've already seen ρ and r. Now we need to think about r^2.
- Regression finds the "best fitting" line in the sense that it minimizes the sum of squared errors.
- Sometime the "best fit" is totally rubbish. How can you tell?
- The most common measure of fit is referred to as the R^2. It relates closely to the correlation between Y and X.
- In fact, for the simple regression model, it is the the sample correlation, squared.
- Think about R^2 as: Variation we can explain/All the variation, or more formally Regression sum of squares / Total Sum of Squares.

Example

```
forestreg <-lm(forest$`Forest loss`~ forest$`Pop dens`, dat
pander(summary(forestreg)) # Calls the entire summary comm
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6	0.1123	5.342	1.152e-06
forest$Pop dens	0.0008423	0.0001165	7.228	5.503e-10

Table 7: Fitting linear model: forest‘*Forestloss*‘ *forest*Pop dens

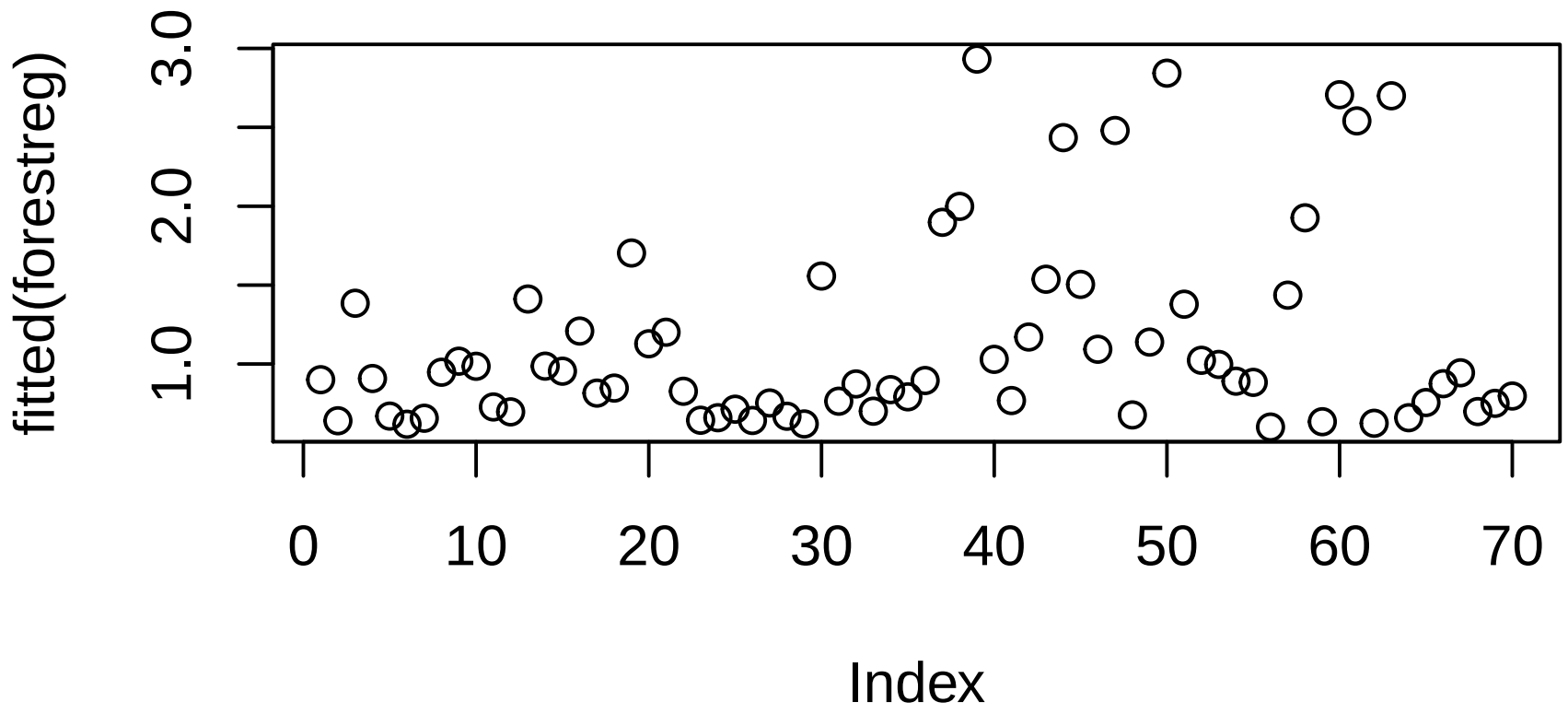
Observations	Residual Std. Error	R^2	Adjusted R^2
70	0.7031	0.4345	0.4262

Pulling some more detail out

```
coef(forestreg) #  weights that minimize the sum of the sq
```

```
##        (Intercept) forest$`Pop dens`
##       0.5999648988       0.0008423268
```

```
plot(fitted(forestreg)) # These are the fitted values of t
```



Back to Wages Example

- what else might be related to wages aside from time spent in school?
- Maybe people with different amounts of work experience also have different wages, at any given level of education
- Regress y = log wage on $\mathbf{x}$ = (constant, years education, experience)

Multivariate regression code

```
library(foreign)
wage1<-read.dta(
  "http://fmwww.bc.edu/ec-p/data/wooldridge/wage1.dta") # .
pander((wageregression2 <- lm(formula =
             lwage ~ educ + exper, data = wage1)))# Regress
```

Multivariate regression code

Table 9: Fitting linear model: lwage $\sim$ educ + exper - At a given level of education, 1 year of experience is associated with 1% higher wages - At a given level of experience, 1 year of education is associated with 9.8% higher wages

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.2169	0.1086	1.997	0.04635
educ	0.09794	0.007622	12.85	4.958e-33
exper	0.01035	0.001555	6.653	7.239e-11

Nonlinearities

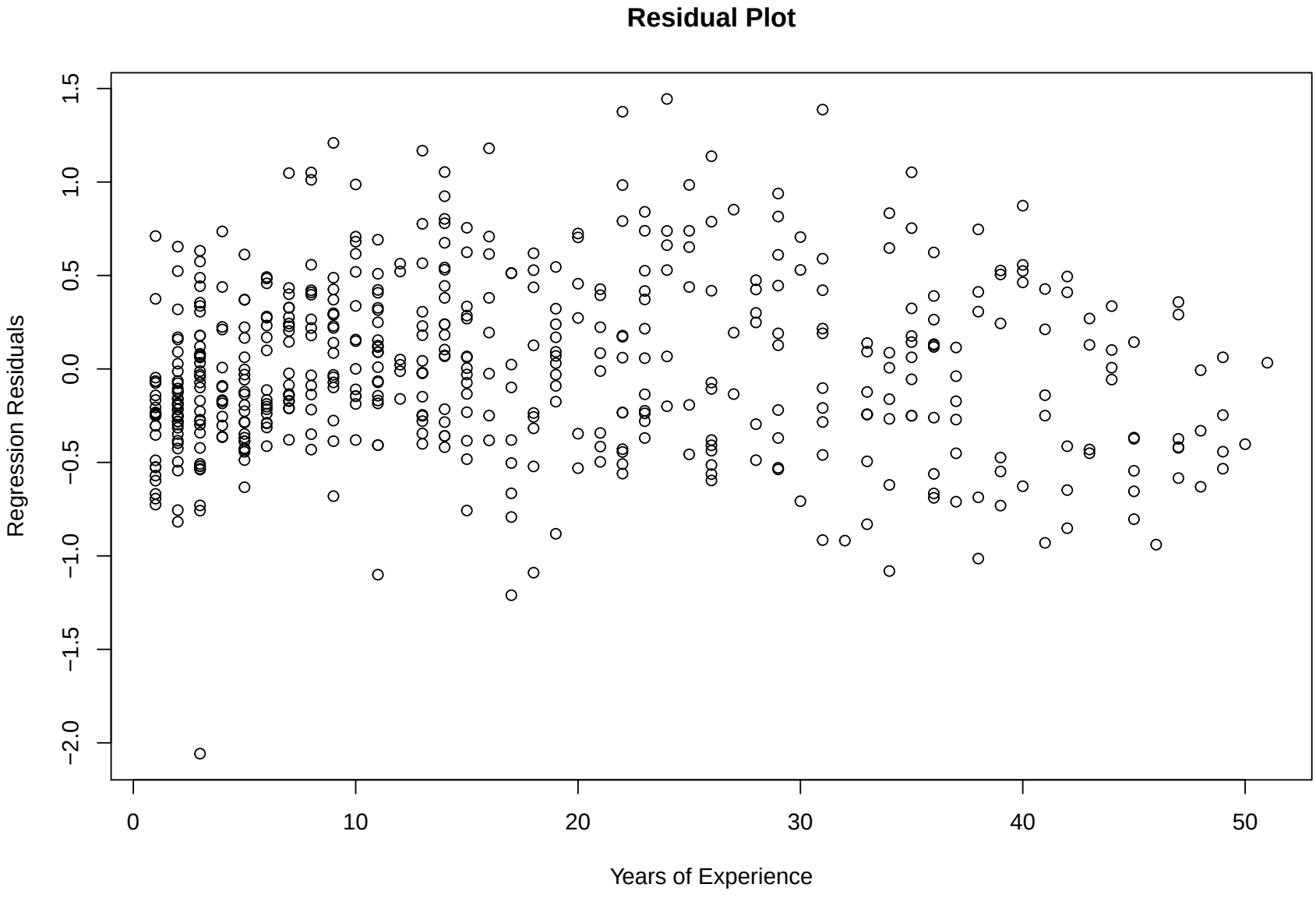
- OLS estimator is linear in β
- But we can model nonlinear functions by allowing $\mathbf{x}$ (or y) to include nonlinear transformations of the data
- For this reason, linearity assumption not as strong as it looks
- Saw this already: use of log wage instead of wage in dollars
- Multiple regression allows formulas like polynomials:
  - $\beta_0 + x_i\beta_1 + x_i^2\beta_2 + \ldots$
- Let's see if this seems like a good idea in our wages case

Residual plot

- Can see if difference of y from predicted value $\mathbf{x}_i'\hat{\beta}$ exhibits systematic patterns by comparing residuals to predictors

```
plot(wage1$exper,wageregression2$residuals,
     ylab="Regression Residuals",
     xlab="Years of Experience", main="Residual Plot")
```

Residuals appear to be predictable from experience



A nonlinear prediction

- Given pattern in the residuals, this suggests we might get a more accurate prediction using a nonlinear function

```
#Add a nonlinear transform of experience to x
wage1$exper2<-(wage1$exper)^2
#Run the augmented regression
wageregression3 <- lm(formula =
        lwage ~ educ + exper + exper2, data = wage1)
```

Output of Regression

```
pander(wageregression3 <- lm(formula =
        lwage ~ educ + exper + exper2, data = wage1))
```

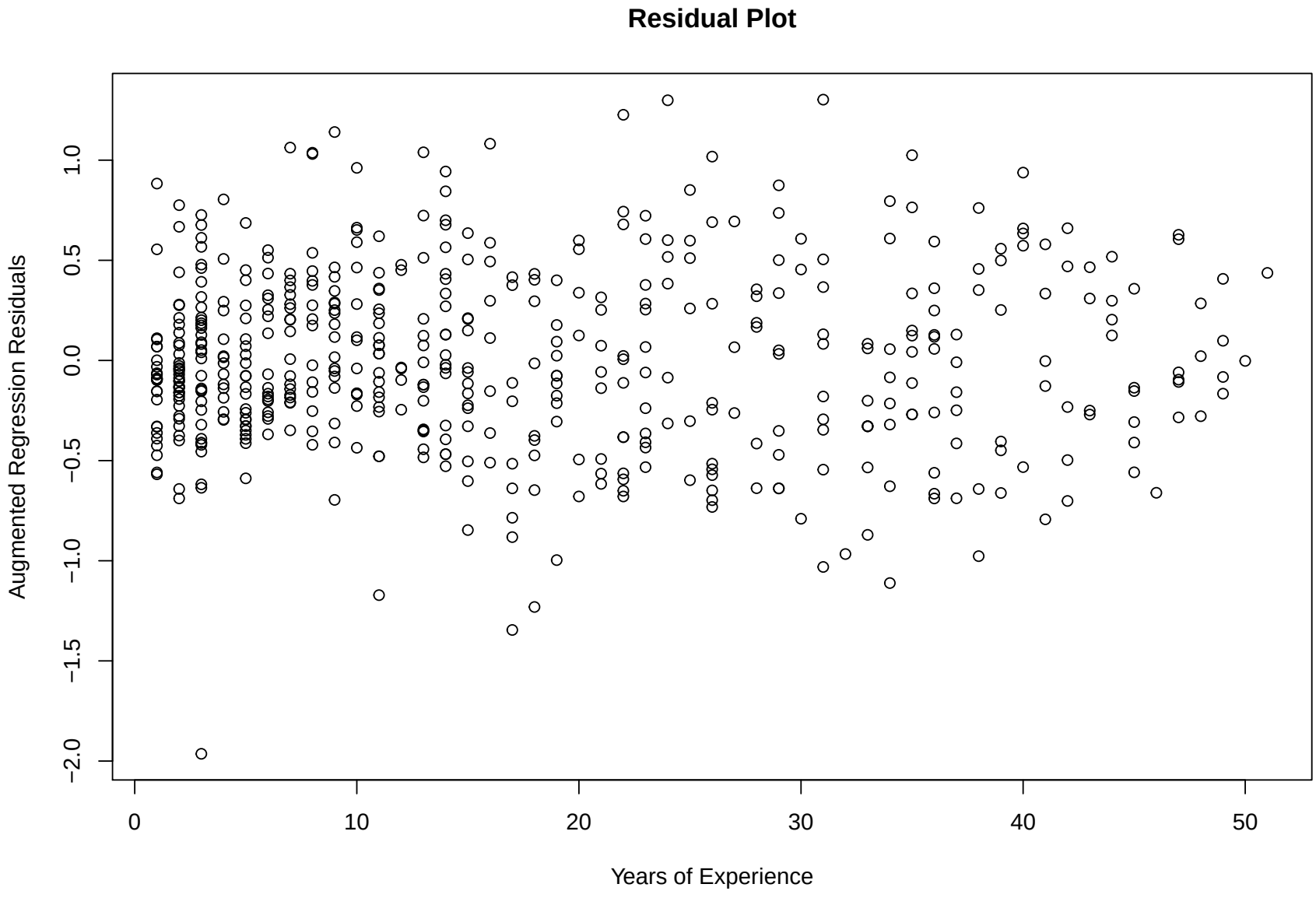
Table 10: Fitting linear model: lwage ~ educ + exper + exper2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.128	0.1059	1.208	0.2275
educ	0.09037	0.007468	12.1	6.979e-30
exper	0.04101	0.005197	7.892	1.765e-14
exper2	-0.0007136	0.0001158	-6.164	1.421e-09

Check residuals again

```
plot(wage1$exper,wageregression3$residuals,
     ylab="Augmented Regression Residuals",
     xlab="Years of Experience",main="Residual Plot")
```

Better now: no easily discernible pattern



Linear models

- If true relationship is linear in $\mathbf{x}$, $\hat{\beta}$ will uncover it
- What assumptions are needed for this to be true? It is very important to know these as in practice they get violated all the time.

1. In population, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$
2. $(y_i, \mathbf{x}_i') : i = 1 \ldots n$ are independent random sample of observations following 1
3. There are no exact linear relationships among the variables $x_1 \ldots x_k$
4. $E(u|\mathbf{x}) = 0$
5. $Var(u|x) = \sigma^2$ a constant > 0

- Sometimes people replace (4) by slightly weaker
- (4'): $E(u_i x_{ij}) = 0$ for $j = 0 \ldots k$
- or add

6. $u \sim N(0, \sigma^2)$

Estimator Properties

- Under Assumptions (1-3) and (4')
  - OLS is consistent: $Pr(\|\hat{\beta} - \beta\| > e) \to 0$ for all $e > 0$
- Under Assumptions (1-4)
  - OLS is unbiased: $E(\hat{\beta}) = \beta$
- Under (1-5), we can derive the sample variance of $\hat{\beta}$ and show its *efficiency*
- Gauss-Markov theorem: Under (1-5), any estimator of β which is unbiased and linear in y has sample variance at least as large as that of $\hat{\beta}$
- Additionally, (1-5) imply $\hat{\beta}$ is asymptotically normal.

Variance and Asymptotic Distribution

- A tedious proof shows under (1-5)

$$Var(\hat{\beta}|\mathbf{x}) = \sigma^2(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i')^{-1}$$

- Under (1-5) a (not so easy) argument via the central limit theorem shows

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Sigma)$$

  where $\Sigma := \sigma^2 E(\mathbf{x}_i \mathbf{x}_i')^{-1}$
- This result is what lets us build confidence intervals and tests

Inference: single parameter

- For any one β_j, the distribution is approximately normal
- We can estimate Σ by $\hat{\Sigma} = \frac{1}{n-k-1} \sum_{i=1}^{n} \hat{u}_i^2 (\frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i \mathbf{x}_i'))^{-1}$ where $\hat{u}_i$ is the sample residual
- Level $1 - \alpha$ confidence interval for β_j is then

$$(\hat{\beta}_j - \frac{z_{1-\alpha/2}}{\sqrt{n}} \hat{\Sigma}_{jj}^{\frac{1}{2}}, \hat{\beta}_j + \frac{z_{1-\alpha/2}}{\sqrt{n}} \hat{\Sigma}_{jj}^{\frac{1}{2}})$$

  where $z_{1-\alpha/2}$ satisfies $Pr(Z < z_{1-\alpha/2}) = 1 - \alpha/2$ when $Z \sim N(0, 1)$
- Common to use quantile of t_{n-k-1} distribution instead, which is exact under (6)
- Doesn't hurt to do this even if (6) false, since for large n approximately the same, and normality is large n approximation anyway

Inference: multiple parameters

- Often want to test hypotheses about multiple coefficients
  - e.g. H_0: $\beta_1 = \beta_2 = 0$, H_1: $\beta_1 \neq 0$ or $\beta_2 \neq 0$
- F test: run regression without restrictions, then run with restriction

$$F = \frac{(\sum_{i=1}^{n} \hat{u}^2_{i,\text{restricted}} - \sum_{i=1}^{n} \hat{u}^2_{i,\text{unrestricted}})/q}{\sum_{i=1}^{n} \hat{u}^2_{i,\text{unrestricted}}/n - k - 1}$$

  - k is number of included variables in unrestricted regression
  - q is number of restrictions (count equal signs in H_0)
- Under (1-5) and H_0, $F \xrightarrow{d} \chi^2_q$ asymptotically
- Under (1-6) and H_0, $F \sim F_{q,n-k-1}$ in finite samples
  - Again, doesn't hurt to use this as approximation

Automatic tests

- t tests of univariate hypothesis $\beta_j = 0$ produced automatically by Summary() command. t-tests determine if two sets of data are significantly different from each other.
- Similarly F test of $\beta_j = 0$ for all $j = 1 \ldots k$ is produced. F-test used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled.

```
summary(wageregression3)
```

Output

```
pander(summary(wageregression3))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.128	0.1059	1.208	0.2275
educ	0.09037	0.007468	12.1	6.979e-30
exper	0.04101	0.005197	7.892	1.765e-14
exper2	-0.0007136	0.0001158	-6.164	1.421e-09

Table 12: Fitting linear model: lwage ~ educ + exper + exper2

Observations	Residual Std. Error	R^2	Adjusted R^2
526	0.4459	0.3003	0.2963

Does experience matter?

- Running tests other than the standard ones requires more work
- Suppose we want to know if experience helps predict wage
- Because we include experience and its square, relevant null hypothesis is $\beta_2 = \beta_3 = 0$

Running the test manually

```
#Run restricted regression
restrictedreg<-lm(formula = lwage ~ educ, data = wage1)
#Restricted residual sum of squares
RSS_r<-sum((restrictedreg$residuals)^2)
#Unrestricted residual sum of squares
RSS_u<-sum((wageregression3$residuals)^2)
#Difference in degrees of freedom
q<-restrictedreg$df-wageregression3$df
#Formula
(Fstat<-((RSS_r-RSS_u)/q)/(RSS_u/wageregression3$df))
```

```
## [1] 42.69616
```

```
#p value: reject H0 if small
(pvalue<-1-pf(Fstat,q,wageregression3$df))
```

```
## [1] 0
```

Variable choice

- In practice, which regressors *should* we include?
- Depends on goal of regression
- If prediction, whatever set yields least error (may not be set leading to least error in sample, due to sampling variability)
- If structure, we want to know particular β_j in context of a model including some "true" set
- Regardless of "truth," can always ask what is difference between estimates when a variable is or is not included

Omitted variables formula

- Consider regression of y on $x_0, x_1, \ldots, x_k$ to get estimate $\hat{\beta}$
- What are results if we instead regress y on $x_0, x_1, \ldots, x_{k-1}$ to get $\tilde{\beta}$, omitting x_k
- Maybe because we don't observe x_k in our data set
- Let $\tilde{\delta}_j$, $j = 0 \ldots k-1$ denote the coefficients in a regression of x_k on $x_0, x_1, \ldots, x_{k-1}$
- Then we can write $\tilde{\beta}_j$ as

$$\tilde{\beta}_j = \hat{\beta}_j + \hat{\beta}_k \tilde{\delta}_j$$

- In words, if a variable is omitted, the coefficients in the "short" regression equal the coefficient in the long regression plus the effect of the omitted variable on the outcome times the partial effect of the omitted variable on the included regressor
- Difference disappears if either excluded regressor had 0 partial correlation with the included regressor or had no partial correlation with the outcome

Bias?

- If (1-3) and (4') hold the long regression, they also hold in the short regression, for different values of β, and so both are consistent for some linear function
- If we have reason to be interested in the linear function corresponding to the long regression, omitted variables mean that we will not get a valid estimator if we are missing some variable and it is linked to the outcome and to the regressor of interest
- Under (1-4) for the long regression, obtain $E(\tilde{\beta}_j|\mathbf{x}) = \beta_j + \beta_k\tilde{\delta}_j$ and so this is called "omitted variables bias". Remember Yule's poverty regression?

Example: experience again

```
#Construct short regression coefficient from formula
deltareg<-lm(formula = exper ~ educ, data = wage1)
delta1<-deltareg$coefficients[2]
betahat1<-wageregression2$coefficients[2]
betahat2<-wageregression2$coefficients[3]
(omittedformula<-betahat1+betahat2*delta1)
```

```
##        educ
## 0.08274437
```

```
#Run short regression without experience directly
wageregression1<-lm(formula = lwage ~ educ, data = wage1)
(betatilde1<-wageregression1$coefficients[2])
```

```
##        educ
## 0.08274437
```

Interpretation

- Omitting experience from the wage regression reduces estimated effect of education on wages
- Reason: people who spend more time in school have less work experience, and work experience is positively associated with wages
- If we want to compare wages of people with similar levels of work experience and different education levels, we get larger differences than if experience not kept constant
- Not clear at all that this is the comparison we want to make
  - If you decide to spend one more year in school rather than working, you will have one more year of education, but will have less work experience than if you hadn't decided to stay in school
  - Much more on this idea next week

More on (3): Multicollinearity

- Finding a single $\hat{\beta}$ requires that system have a unique solution
- This fails if any regressor can be written as linear combination of some other subset of regressors
- E.g. $x_{1i} = a * x_{2i} + b * x_{3i}$ for all i
- Then if $(\beta_1, \beta_2, \beta_3)$ solve the minimization problem, so does $(\beta_1 + c, \beta_2 - c * a, \beta_3 - c * b)$ for any c

Interpreting multicollinearity

- Information in variables is redundant
  - Usually happens if one variable is *defined* in terms of another
  - E.g. $x_1 = 1\{\text{A is true}\}$, $x_2 = 1\{\text{A is false}\}$
  - Logically, always have $x_1 = 1, x_2 = 0$ or $x_1 = 0, x_2 = 1$
  - Not even sensible to ask what would happen if A is both true and false or neither
- First example of *failure of identification*
- Is it a problem?
  - Maybe not: Predicted value $\mathbf{x}_i'\hat{\beta}$ the same no matter which solution chosen
  - Maybe yes: if we want to predict what would happen if $\mathbf{x}$ took on a value not along the combination and this is sensible to ask, we simply have a data set which can't tell us the answer: need better data

Handling multicollinearity in practice

- Let's see how software handles it

```
#Initialize random number generator
set.seed(42)
#Draw 100 standard normal random variables
xa<-rnorm(100)
xb<-rnorm(100) #Draw 100 more
#Define 3rd variable as linear combination of first 2
xc<-3*xa-2*xb
#define y as linear function in all variables + noise
y<-1*xa+2*xb+3*xc+rnorm(100)
#Regress y on our 3 redundant variables
(multireg <-lm(y~xa+xb+xc))
```

Output

```
##
## Call:
## lm(formula = y ~ xa + xb + xc)
##
## Coefficients:
## (Intercept)           xa           xb           xc
##    0.001766     9.856291    -3.914707           NA
```

- We see R simply drops one variable
  - Coefficient set to 0
- In this case the last: choice is arbitrary
- Can always do this: pick one element of identified set
- Sometimes this is reasonable, sometimes not

Ways to derive OLS

- Why did we choose OLS rather than some other estimator to learn from data?
- 3 ways to derive OLS

1. Empirical Risk Minimization
2. Method of Moments
3. Maximum Likelihood Estimation

Interpretations of OLS, 1: Empirical risk minimizer

- Suppose our goal is prediction of y using $\mathbf{x}$
- Suppose we believe a good prediction is one that on average is close to y
- Pick a predictor that minimizes this loss in sample
- If law of large numbers holds, and cases we want to predict drawn from same distribution, in sample loss at a given predictor should be similar to out of sample loss
- Takes more work (and assumptions) to show smallest in sample loss also good out of sample, but idea often works

Interpretations of OLS, 2: Method of moments

- Now suppose we are willing to make some assumptions about distribution
- Assume (1) and (4')
- Method of moments: replace expectation with sample average
- Here gives exactly the first order conditions defining the estimator
- Under (1-3) and (4'), OLS is estimator that satisfies given moment conditions
- Assumes linearity, but only weak conditions on residual

Interpretations of OLS, 3: Maximum likelihood estimator

- MLE idea: estimate distribution by finding parameter values under which the probability density of observing the data set that was actually observed was highest
- Suppose we think $z_i,\ i = 1 \ldots n$ drawn i.i.d. from density $f(z, \theta)$ with unknown parameter θ
- MLE solves

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \Pi_i^n f(z, \theta)$$

- Will see more about MLE later in the class: has very nice properties if we believe our model of the density

MLE view

- Suppose $y_i - \mathbf{x}_i'\beta \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$ for some (β, σ^2)
- Then OLS estimator of β coincides with maximum likelihood estimator
- Tells us that, at least under strong assumptions, OLS should be good estimator

Next time

- Explaining variance in multiple regression
- Statistical aspects
- Interpreting multiple regression
- Biases: multicollinearity/heteroskedasticity/autocorrelation
- Example: education spending and educational attainment

Lecture 4

Last time

- Nonlinearity in Regression
- Factors affecting β
- Calculating confidence intervals for β
- Example: regression by hand, roll your own betas using R.
- Readings: Koop cht 4 & Freedman cht 3

This time

- Explaining variance in multiple regression
- Statistical aspects
- Interpreting multiple regression
- Biases: multicollinearity/heteroskedasticity/autocorrelation
- Example: education spending and educational attainment

Digression on project

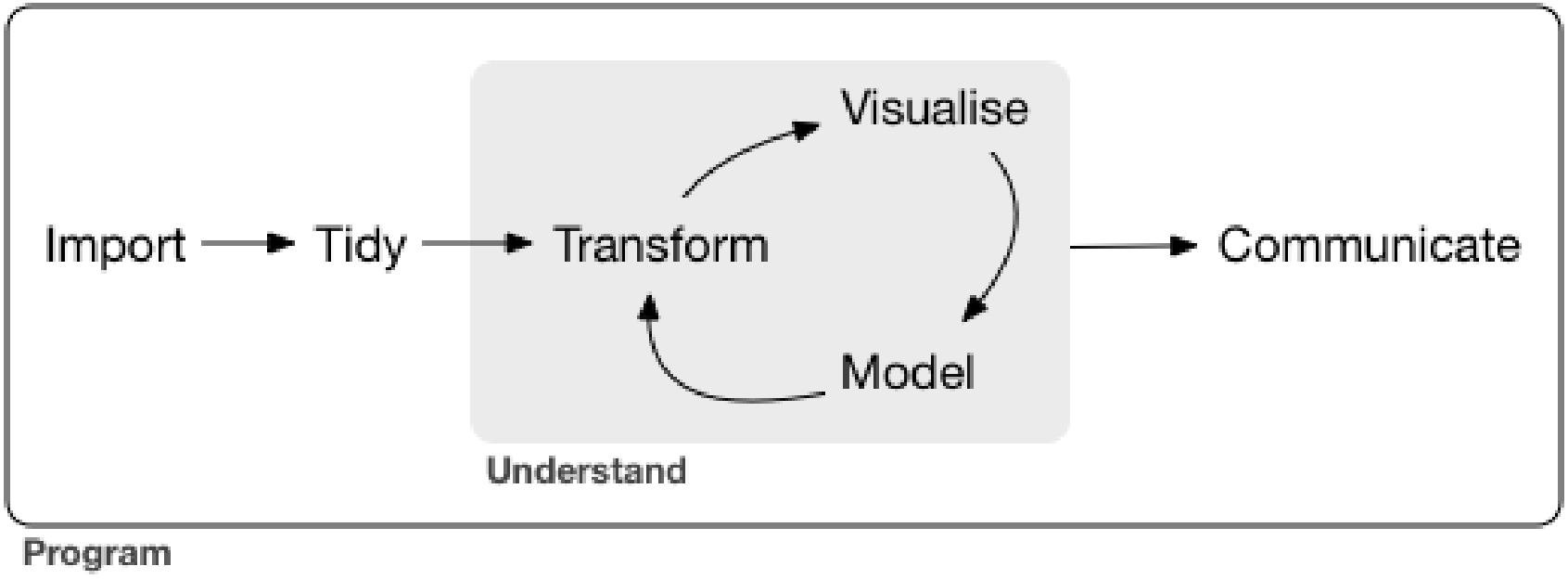


Figure 5: Source: Wickham, 2016

Submit a .RMD file

- You'll be given 2 *default* datasets, so we'll sort the 'tidy' element for you. You can also find your own.
- You do the rest. We want to see you transforming, visualising, modeling and communicating the data.
- Submit a 1 page outline of your project to Niall in week 7/8, meet with him to go through the project outline and make sure it is do-able in the time you have available.

Outline

- Background to the data
- Visualisation/Transformation
- Modeling
- Conclusion
- References, etc.

Multiple regression

- You've seen something like WAGE $\sim$ EDUCATION
- This exists as a relation like WAGE = a + b*EDUCATION + error.
- As code it exists as reg <- lm(Wage $\sim$ EDUCATION, data = somedataset)
- This will pump out values for a and b, as well as giving you some sense of the error, residual, standard errors, etc, and you can then run various tests.
- We do need to think carefully about 'multiple' explanatory variables. This is multiple regression.

Digression: Matrices (freedman, chapter 4)

- Matrix, from 'mother'. Essentially a container that you can manipulate, and while doing so, manipulate the elements of that matrix.
- Express 'size' as m rows by n columns, normally written $n \times n$.
- Multiply through matrices, add, subtract, and divide by scalars easily.
- Multiply by other matrices easily too.
- Here we have two matrices, **A** and **B**, multiplying one another.

$$\begin{pmatrix} 0.8944272 & 0.4472136 \\ -0.4472136 & -0.8944272 \end{pmatrix} \begin{pmatrix} 10 & 0 \\ 0 & 5 \end{pmatrix}$$

Matrices

An $m \times n$ matrix A can be multiplied by a matrix B of sign $n \times P$. This gives a matrix $n \times p$ in shape.

- To get your head around the idea of matrices, do exercises for pages 30–40 of Freeman.

Multiple regression

The model in matrix form is

$$Y = \mathbf{X}\beta + \epsilon$$

- Where Y, the dependent variable, is an n x 1 vector of observable random variables. Y_i is the ith component of Y. - X is an $n \times p$ matrix of observable variables. It's often called the 'design matrix' or 'training matrix'. - β is an $n \times 1$ vector of parameters - ϵ is an $n \times 1$ random vector of errors or disturbances.

Explaining variance in multiple regression

- after fitting to data, we get

$$Y = \mathbf{X}\hat{\beta} + e$$

- The fraction of variance 'explained' by the regression is

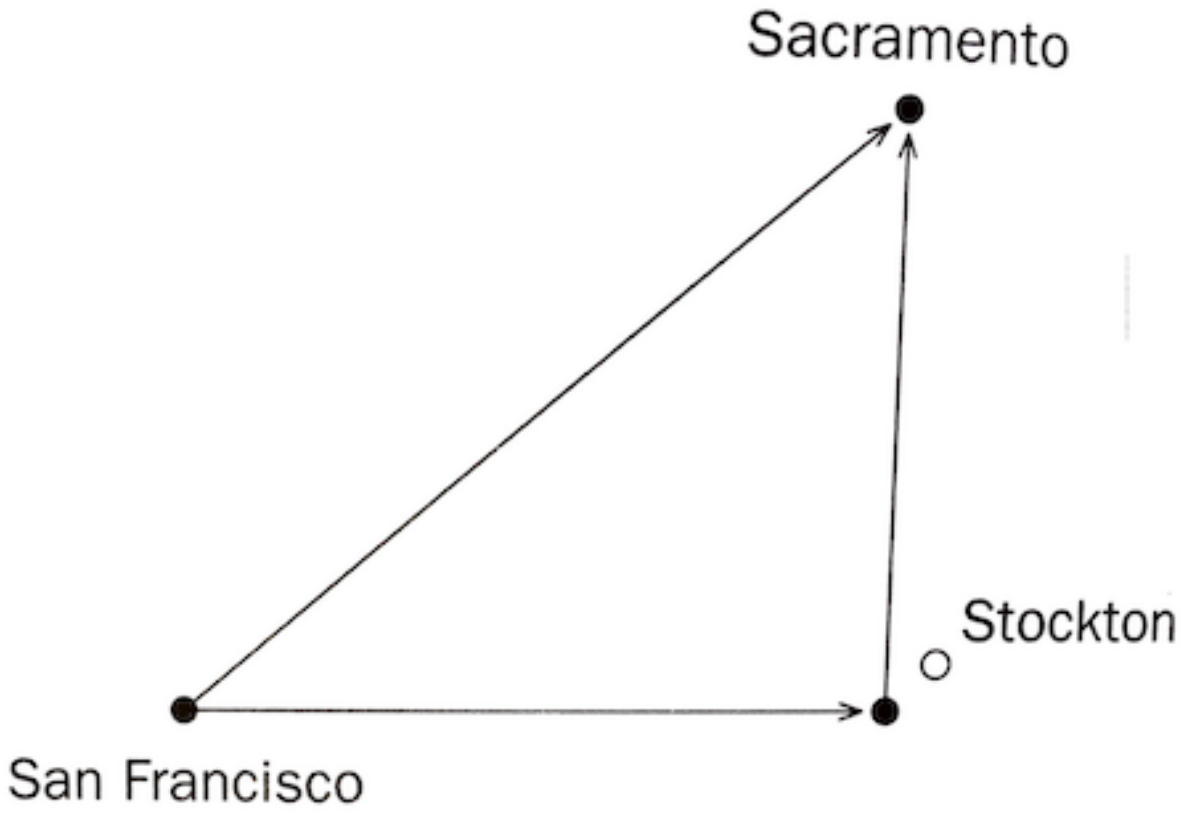
$$R^2 = var(\mathbf{X}\hat{\beta})/var(Y)$$

- Problem: has wrong size and has wrong units. Another problem: there's a difficulty in understanding the idea.

Geometric Intuition. Imagine you're a crow.

- Sacramento is 78 miles from San Francisco (as the crow flies)
- Or crow could fly 60 miles East and 50 miles North, passing Stockton. Pythagoras tells us the area of the triangle is

$$60^2 + 50^2 = 3600 + 2500 = 6100 \text{ miles}^2$$



Ye wha?

- The analogy is: the area between SF and Sacramento is 6100 square miles, of which 3,600 is explained by 'East'.
- Projecting on East stands for projecting Y and X orthogonally to the vector u that is all 1's.
- And then projecting the remainder of Y onto what is left of the column space of X.
- The hypoteneuse of the triangle is $Y = \bar{Y}_u$, which squared in length is $n * var(\mathbf{X}\hat{\beta})$
- The horizontal edge is $X\hat{\beta} - \bar{Y}u$
- The vertical edge is e and $e^2 = n * var(e)$
- The theory of explained variance boils down to Pythagoras theorem and a crow. Very Rubberbandits all together.
- Anyway, a high R^2 is a measure of the goodnewss of fit between the data and the equation: the residuals relative to the standard deviation are small. A low R^2 indicates a poor fit.

Caution (for like the 120th time)

- R^2 measures goodness of fit, not validity of underlying causal model.
- Example: very strong correlation between inflation rate and lung cancer detection rate. But inflation doesn't cause or prevent cancer

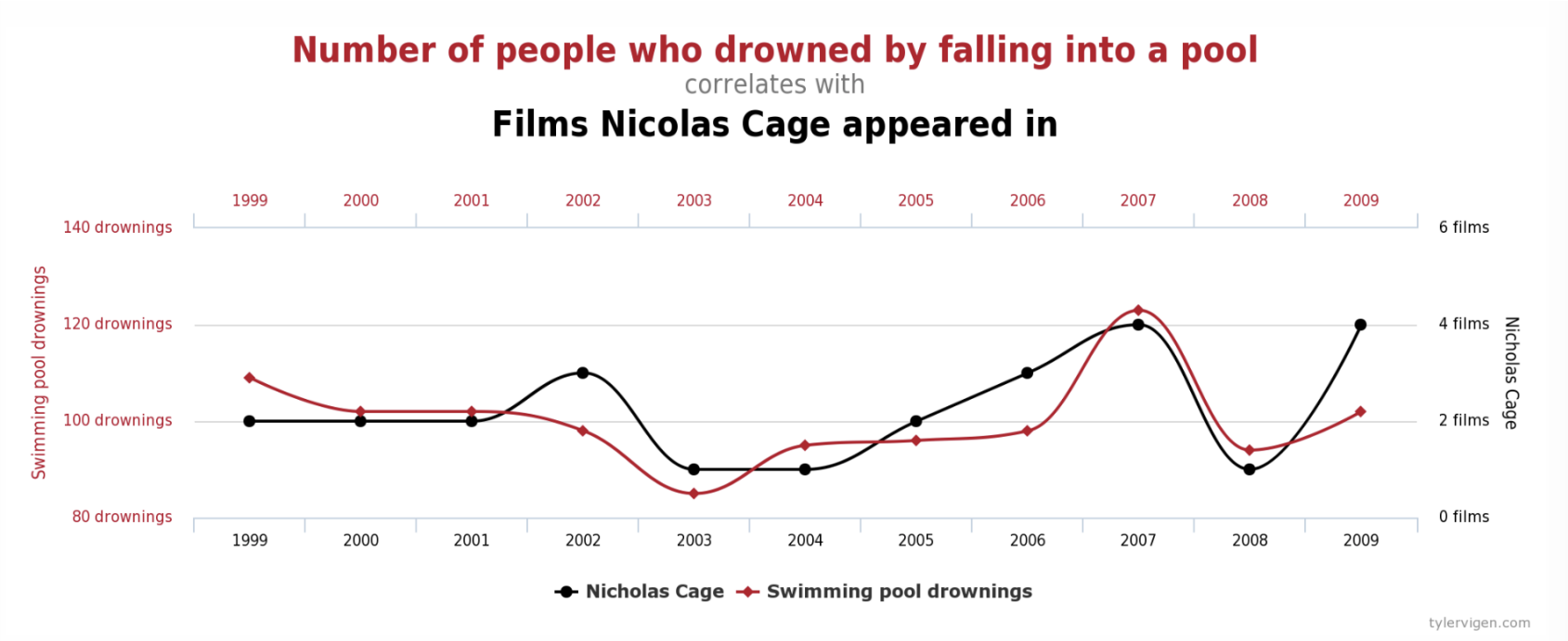


Figure 6: See `http://www.tylervigen.com/spurious-correlations` for more

Application: Wages

- Let's go back to our wages data set: what else might be related to wages aside from time spent in school?
- Maybe people with different amounts of work experience also have different wages, at any given level of education
- Regress y = log wage on $\mathbf{x}$ = (constant, years education, experience)

Multivariate regression code

```
# Obtain access to data sets used in our textbook
library(foreign)
# Import data set of education and wages
wage1<-read.dta(
  "http://fmwww.bc.edu/ec-p/data/wooldridge/wage1.dta")
# Regress log wage on years of education
(wageregression2 <- lm(formula =
            lwage ~ educ + exper, data = wage1))
```

Multivariate regression code

```
##
## Call:
## lm(formula = lwage ~ educ + exper, data = wage1)
##
## Coefficients:
## (Intercept)          educ         exper
##     0.21685       0.09794       0.01035
```

- At a given level of education, 1 year of experience is associated with 1% higher wages
- At a given level of experience, 1 year of education is associated with 9.8% higher wages

Kahneman: Focusing Illusion

"Nothing In Life Is As Important As You Think It Is, While You Are Thinking About It"

Income is determined by:

- Education BUT If everyone had the same education, the inequality of income would be reduced by less than 10%
- Income BUT If everyone had the same income, the differences among people in life satisfaction would be reduced by less than 5%

Kahneman: Focusing Illusion

Happiness is determined by:

- Income BUT If everyone had the same income, the differences among people in life satisfaction would be reduced by less than 5%.
- Inequality BUT On average, individuals with high income are in a better mood than people with lower income, but the difference is about 1/3 as large as most people expect.
- The mismatch in the allocation of attention between thinking about a life condition and actually living it is the cause of the focusing illusion.

What happens to OLS if assumptions break down?

- if the assumptions are wrong, the estimator can be severely biased, or worse, the standard errors computed from the data can be way off.

"Well-specified" linear models

- Recall "standard" assumptions

1. In population, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$
2. $(y_i, \mathbf{x}_i') : i = 1 \ldots n$ are independent random sample of observations following 1
3. There are no exact linear relationships among the variables $x_1 \ldots x_k$
4. $E(u|\mathbf{x}) = 0$
5. $Var(u|x) = \sigma^2$ a constant > 0

- Sometimes people replace (4) by slightly weaker
- (4'): $E(u_i x_{ij}) = 0$ for $j = 0 \ldots k$
- or add

6. $u \sim N(0, \sigma^2)$

Conditional expectation functions

- We call $E[y|\mathbf{x}]$ the conditional expectation function, c.e.f.
- It gives average value of y at any value of $\mathbf{x}$
- The c.e.f. is not required to be linear
- Under assumption (4), we have $E[y|\mathbf{x}] = \mathbf{x}'\beta$
  - Linear c.e.f.
- If we believe this, OLS is not just best linear predictor, but best of any predictor
- Not entirely crazy to think this is reasonable if $\mathbf{x}$ chosen carefully

What we can say under different subsets

- Properties of $\hat{\beta}$ from OLS under different assumptions
- (1-6): finite sample t and F distributions
- (1-5): asymptotic normal distributions, asymptotic efficiency (smallest asymptotic variance), finite sample efficiency (Gauss Markov: best linear unbiased estimator)
- (1-4): unbiasedness
- (1-3), (4'): consistency
- (1-2), (4'): convergence to some (arbitrary) element of identified set

What does this look like?

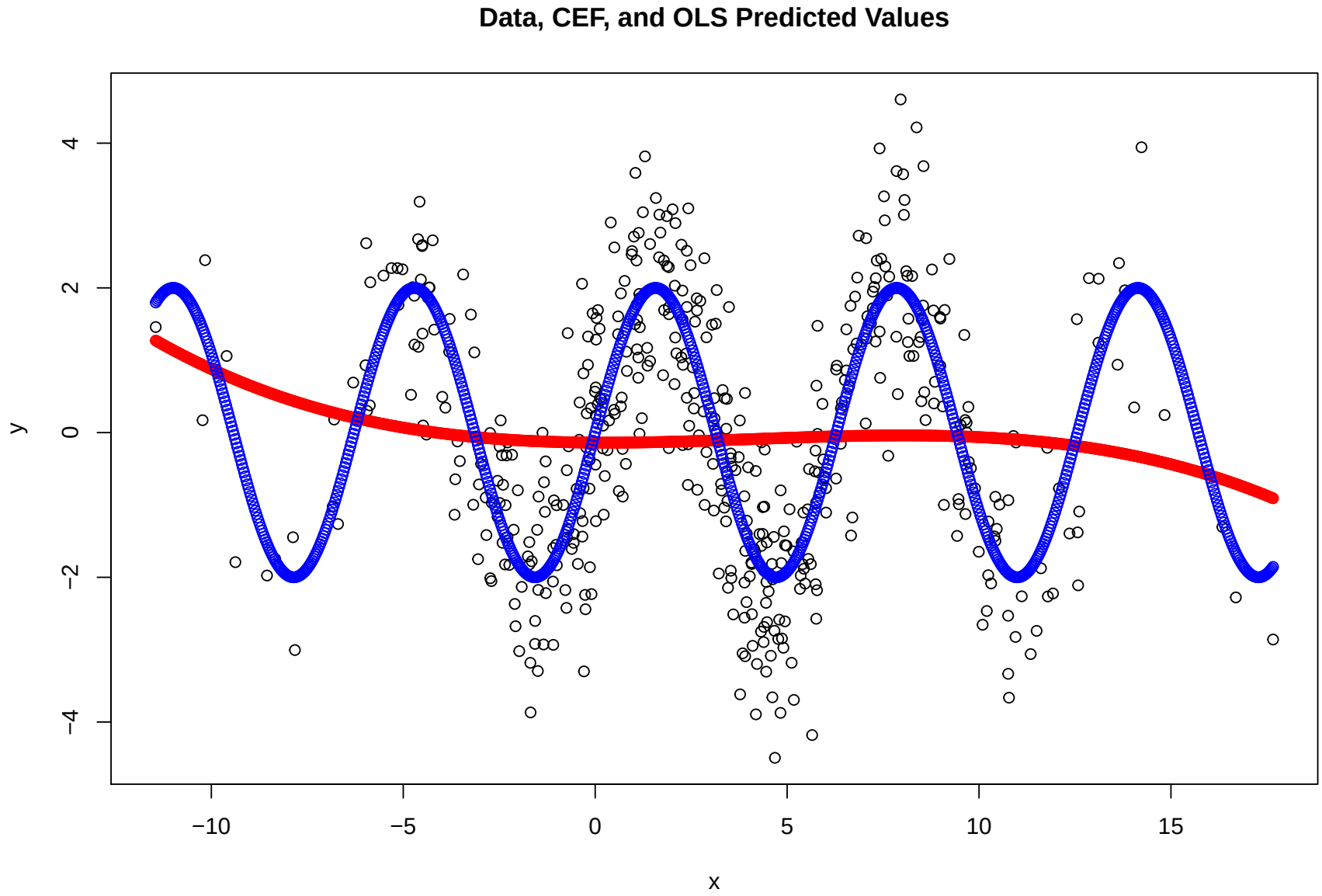
```
#Generate some data from a nonlinear relationship
x1<-rnorm(500,mean = 3, sd=5)
trueerror<-rnorm(500) #Residual from true relationship
#Not same as OLS residual
#A nonlinear relationship: E[y|x]=2*sin(x)
y1<-2*sin(x1)+trueerror
# Run a regression in which c.e.f. not linear in x
#Include a polynomial terms to allow nonlinearity
(misspecifiedregression<-lm(y1 ~ x1 + I(x1^2) + I(x1^3)))
```

```
##
## Call:
## lm(formula = y1 ~ x1 + I(x1^2) + I(x1^3))
##
## Coefficients:
## (Intercept)           x1      I(x1^2)      I(x1^3)
##  -0.1384650   -0.0031855    0.0054605   -0.0004386
```

Plot prediction and c.e.f.

```
#Generate x values at which to evaluate functions
xinterval<-seq(from=min(x1),to = max(x1),length.out=1000)
new<-data.frame(x1=xinterval)
#Calculate $x'\hat{\beta}$ using predict command
pred<-predict(misspecifiedregression,new)
#plot Data, O.L.S. predictions, and c.e.f.
plot(x1,y1, xlab = "x", ylab="y",
     main="Data, CEF, and OLS Predicted Values")
points(xinterval,pred, col = "red") #OLS Predictions
points(xinterval,2*sin(xinterval), col = "blue") #True CEF
```

Plot prediction and c.e.f.



What to make of OLS

- Clearly, doesn't do a great job of recovering a "wiggly" CEF
  - Even with nonlinear terms
- But, does an okay job on average over $\mathbf{x}$
  - By construction, about as likely to be wrong above as to be wrong below
- If this is our goal, or we truly believe c.e.f. not too wiggly, OLS maybe okay

Inference

- Defining the OLS coefficients by (1-3), (4'), how do we do inference?
- Possible to assess sampling variability in $\hat{\beta}$ around population best fit line
- But we need to be able to say more about residuals

Heteroskedasticity

- Inference was made under **Homoskedasticity**

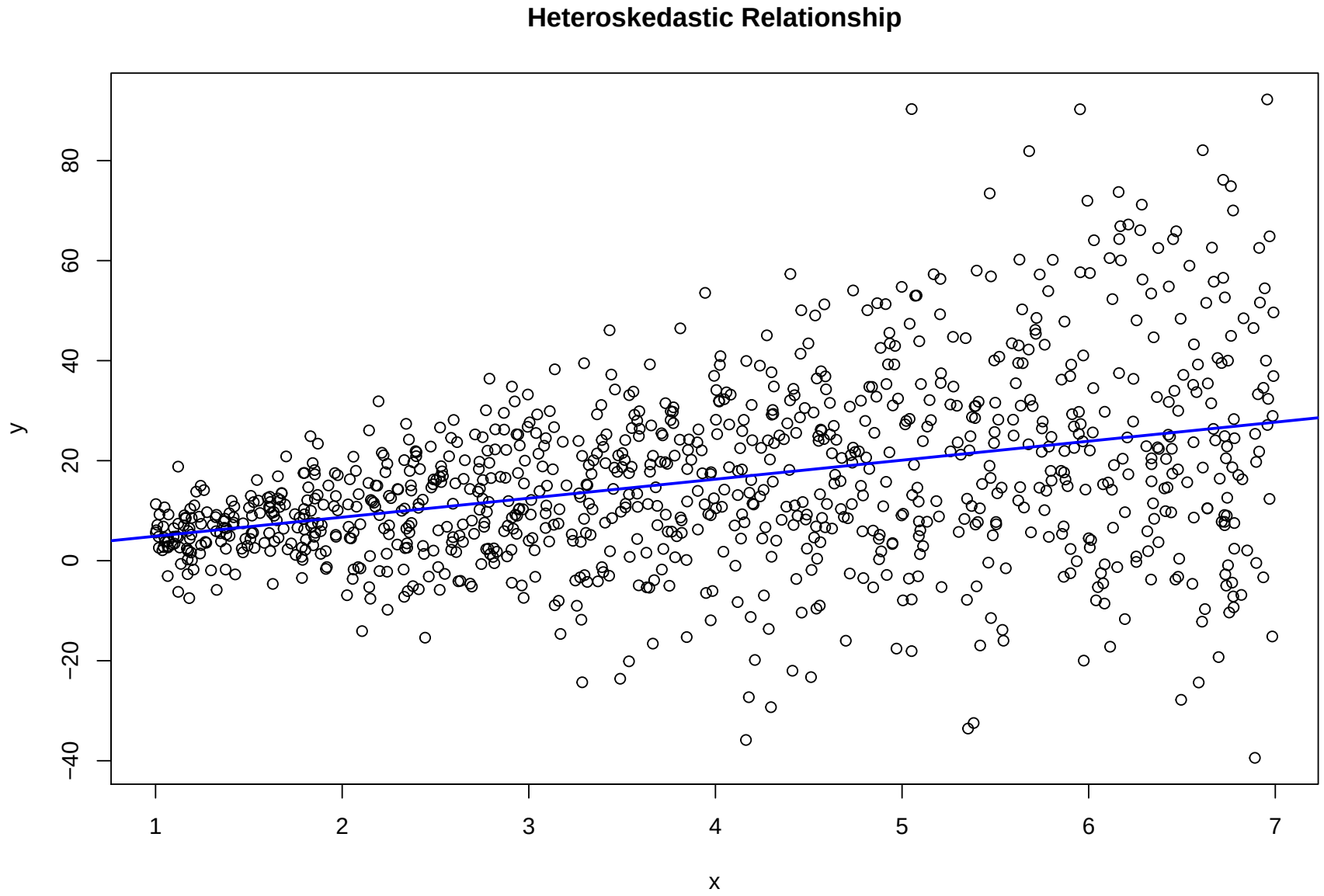
5. $Var(u|x) = \sigma^2$ a constant > 0

- Says dispersion of deviations from best fitting line is equal at all predictor values
- Regardless of what you know, your prediction of y given **x** is equally accurate on average
- Often fails:
- If population c.e.f. nonlinear, distribution around line will depend systematically on **x**
- Even in truly linear case, where (4) holds, variance can depend on **x**
  - This situation is called **heteroskedasticity**

Data exhibiting heteroskedasticity

```
#Generate a data set
x<-runif(1000, min=1, max=7)
u<-rnorm(1000)*(4*x) #u is a function of x
y<-1+4*x+u
#Fit linear regression
hetreg<-lm(y ~ x)
#Plot points and OLS best fit line
plot(x,y,xlab = "x", ylab = "y",
     main = "Heteroskedastic Relationship")
abline(hetreg, col = "blue", lwd=2)
```

Data exhibiting heteroskedasticity



Handling heteroskedasticity

- Conditional distribution of y given x now varies
- $Var(u|\mathbf{x}) := \sigma^2(\mathbf{x})$
- Gauss Markov assumptions fail, OLS is not MLE
- OLS no longer efficient
- But method of moments interpretation still works
- Two options:
  - Stick with OLS, but do inference some other way
  - Find a new estimator

Inference under heteroskedasticity

- Finite sample properties of OLS will depend on form of relationship
- Asymptotics for $\hat{\beta}$ still work fine
- By CLT, get asymptotic normal distribution, with new variance formula that depends on $\sigma^2(\mathbf{x})$
- Do we need to estimate $\sigma^2(\mathbf{x})$?
  - Seems hard: don't know functional form
  - Don't quite need it for asymptotics

t-tests under heteroskedasticity

- t-tests use same formula, but robust standard errors
- Since justification is asymptotic, use Normal instead of t critical values

```
library(sandwich)
# Calculate heteroskedasticity-consistent (HC) estimate
Sigmahat<-vcovHC(hetreg,type="HC0")
library(lmtest) #Implements tests
# Test coefs using robust s.e.s & normal critical values
(robusttests<-coeftest(hetreg,df=Inf, vcov=Sigmahat))
#Compare to tests using non-robust S.E.s
sumhetreg<-summary(hetreg)
(nonrobust<-sumhetreg$coefficients)
```

Results: Robust

```
##
## z test of coefficients:
##
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.09753    1.01330  1.0831   0.2788
## x            3.79983    0.32953 11.5310   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Results: nonrobust

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 1.097526  1.3336360  0.8229576 4.107287e-01
## x           3.799828  0.3095017 12.2772451 2.278663e-32
```

Testing multiple parameters

- Usual formula for F test not valid under heteroskedasticity
- Can use Wald test: $H_0 : R\beta = r$, $R\beta \neq r$ for R a $q \times k+1$ matrix consisting of q linear combinations of coefficients
- Under H_0, $\sqrt{n}(R\hat{\beta} - r) \sim N(0, R\Sigma R')$
- Test stat squares this and standardizes by variance $n(R\hat{\beta} - r)'(R\hat{\Sigma}R')^{-1}(R\hat{\beta} - r)$
- $\hat{\beta}$ from an unrestricted regression is used
- Distribution is χ^2_q asymptotically
- Can also implement using **waldtest()** command in **lmtest** library

Robust errors for returns to education

```
library(foreign)
wage1<-read.dta(
  "http://fmwww.bc.edu/ec-p/data/wooldridge/wage1.dta")
wreg <- lm(formula =
        lwage ~ educ + exper + exper2, data = wage1)
#Build robust standard error matrix: note, this command
#divides sandwich formula by n, to get s.e.s instead
#of asymptotic variance, so no "n" needed in Wald stat
WhiteErrors<-vcovHC(wreg,type="HC0")
# Null hypothesis sets b2, b3 to 0
r1<-(c(0,0,1,0); r2<-c(0,0,0,1)
R<-rbind(r1,r2) #Matrix of linear restrictions
# Test joint significance of experience terms
Waldstat<-t(R%*%wreg$coef)
      %*%solve(R%*%WhiteErrors%*%t(R))
      %*%(R%*%wreg$coef) # Test Statistic
pval<-1-pchisq(Waldstat,2) #P-value
```

Robust errors for returns to education

```
## [1] "Wald statistic value is "

## [1] 87.06502

## [1] "p-value is "

## [1] 0
```

Generalized Least Squares

- Can we come up with a better estimator than OLS?
- If form of heteroskedasticity known, yes
- Weight sum of squares by inverse of conditional variance, minimize weighted sum
- Use **weight** option in **lm()**
- Asymptotically more efficient (lower variance) than OLS
- But when do you know form of heteroskedasticity?
  - Rarely: whole point is you don't have correct specification
- If form not exactly known, can try to estimate a model
- But may be wrong, and even if right, is noisy in finite samples
- Empirical researchers rarely bother with GLS

Testing for heteroskedasticity

- Don't bother
- Screws up coverage of confidence intervals
- Robust standard error formula works fine even if homoskedastic
  - Most economists use it by default
- Read about Breusch-Pagan LM test or White test online if you really want to test for it

Conclusions

- OLS works fine for any data distribution
  - So long as what you want to know is best linear fit to conditional expectation
- In general, error terms will depend on $\mathbf{x}$
  - Includes misspecification and noise
  - Inevitable in some situations: discrete outcomes
- Inference for OLS still possible when homoskedasticity fails
  - Use sandwich covariance matrix
  - Replace F test with Wald test
- Tests tell you about relationship between y and $\mathbf{x}$: is there a linear relationship on average
- Reminder: Conditional expectation is just another feature of joint distribution: whether it is what you want depends on structure and design

Lecture 5

- Autocorrelation and $AR(1)$ processes
- Stationarity and Unit roots
- Example: Prices
- Example (gapminder): How does life expectancy change over time for each country?

Last time we looked at multiple regression

- Notion that we can test multiple vectors against one another, one at a time, to discern the effects of several dependent variables on one another.
- Then we look at situations where the underlying assumptions break down–heteroskedasticity.
- This time it makes sense to look at situations where you have to model variables over time.

Time series Review

- Time series present another design for data
  - Series (Y_t, X_t) $t = 1 \ldots T$ of ordered observations
- Serial correlation $Cov(Y_t, Y_{t+h}) \neq 0$ makes usual inference difficult
- But can also model serial correlation to describe properties of data
- With strong exogeneity, can do regressions using time series data and rely on finite sample properties
- If time series are weak dependent, have modified LLN and CLT
- Use this to show consistency and asymptotic normality of regression with dependent data

Weakly Or Serially Dependent Data

- When the future is "like" the past and historical events are like present ones, we can use past to learn about future
- A time series Y_t is stationary if $E[Y_t] = E[Y_{t+h}]$, $Cov(Y_t, Y_{t+h})$ depends only on h, not t
- Y_t is weakly dependent if $Cov(Y_t, Y_{t+h}) \to 0$ as $h \to \infty$ and $\sum_{h=-\infty}^{\infty} |Cov(Y_t, Y_{t+h})| < \infty$
- Can fail if future fundamentally different from past

Regression with dependent data

- When X_t and u_t satisfy weak dependence, can obtain consistency, asymptotic normality for regression
- **Strict exogeneity** $E[u_t|X_1 \dots X_T] = 0 \ \forall t$ can be weakened to
- **contemporaneous exogeneity** $E[u_t|X_t] = 0 \forall t$
- Or as far as $Cov(X_t, u_t) = 0$
- This permits X_t to be related to u_{t-1}, for example
- Allows case where $X_t = Y_{t-1}$: autoregression can be estimated by linear regression
- Unbiasedness fails: usually have substantial bias in finite samples
- If u_t homoskedastic and has no serial correlation, asymptotic variance has standard formula
- This can be true if X_t contains all historical information relevant to predicting Y_t, possibly including lags of Y_t and X_t
- Otherwise, limit depends on long run variance

Time Series Regression: Assumptions

- (TS.1') Linear Model (X_t, Y_t) drawn from model $Y_t = X_t'\beta + u_t$ and satisfy stationarity and weak dependence
- (TS.2') No perfect multicollinearity of X_t
- (TS.3') Contemporaneous exogeneity $E[u_t|X_t] = 0 \forall t$
- or (TS 3") $Cov(X_t, u_t) = 0$
- (TS.4') Comtemporaneous homoskedasticity: $Var(u_t|X_t) = \sigma^2$ $\forall\ t$
- (TS.5') No serial correlation $E[u_t u_s|X_t, X_s] = 0\ \forall\ t \neq s$

Time Series Regression: Results

- Under (TS.1'),(TS.2'),(TS.3'') (or (TS.3'), which implies (TS.3'')), have consistency:

$$\hat{\beta} \xrightarrow{p} \beta$$

- Inference is exactly standard: t-statistics, Wald tests, CIs, etc
- With serial correlation or heteroskedasticity, consistent but need robust standard errors

Example: Monetary Policy Reaction Function ("Taylor Rule")

- Fed or ECB sets monetary policy by changing Fed Funds Rate r_t
- Want to know how Fed responds to macroeconomic variables
- Suggested that good description is that Fed responds to inflation inf_t and output gap (measure of difference in GDP from "potential" level) gap_t
- r_t evolves more slowly than inflation or output, so current value also predictable from past: r_{t-1}
- Regression equation for reaction function is

$$r_t = \beta_0 + \beta_1 r_{t-1} + \beta_2 inf_t + \beta_3 gap_t + u_t$$

- Run regression using data from **FRED**: macroeconomic database for US

```
taylorrule<-dynlm(ffr~L(ffr)+inf+gap)
```

Estimates

Table 13: Fed Funds Rate Policy Rule

	Dependent variable:
	ffr
L(ffr)	0.926***
	(0.023)
inf	0.079**
	(0.036)
gap	0.100***
	(0.024)
Constant	0.189*
	(0.101)
Observations	248
R^2	0.945
F Statistic	1,403.876*** (df = 3; 244)
Note:	*p<0.1; **p<0.05; ***p<0.01

Inference in Regression with dependent errors

- Homoskedasticity assumption not critical
  - Heteroskedasticity can be handled by using usual sandwich variance formula
- Zero serial correlation in errors $E[u_t u_s | X_s, X_t] = 0$ still needed
- Strong assumption: given predictor variables, future values of outcome completely uncorrelated with past values
- This might be true if we have included any past variables that are related to outcome
- If errors serially correlated, future predictable from past, so should include more if forecasting
- May not want to or be able to include all relevant predictors
- Don't need 0 serial correlation for consistency, asymptotic normality
- If u_t stationary and weakly dependent, OLS asymptototically normal, with a variance that depends on correlation in errors
- Follows from weakly dependent CLT
- If we can estimate the variance, can perform t and Wald tests with new variance estimator

Modeling Residual Serial Correlation

- Can estimate long run variance if we have a model of serial correlation in errors
- E.g. $u_t = \rho u_{t-1} + e_t$, $e_t \sim i.i.d.(0, \sigma_e^2)$ independent of X then $Cov(u_t, u_{t+h}) = \rho^h \sigma_e^2$
- One parameter, can be estimated by two step procedure
  - Regress Y_t on X_t to get residuals $\hat{u}_t$
  - Regress $\hat{u}_t$ on $\hat{u}_{t-1}$ to get $\hat{\rho}$
- With $\hat{\rho}$, can use it to run weighted regression which eliminates serial correlation
- Generalized least squares procedure is **Cochrane-Orcutt** (or, with slight modifications, **Prais-Winsten**) method
- Gives efficient estimates if model of serial correlation is correct
- Can extend to AR(2) model, etc
- If model not correct, still consistent, but inference not accurate

Example continued: Robust inference for Taylor Rule

```
coeftest(taylorrule,vcovHAC)
```

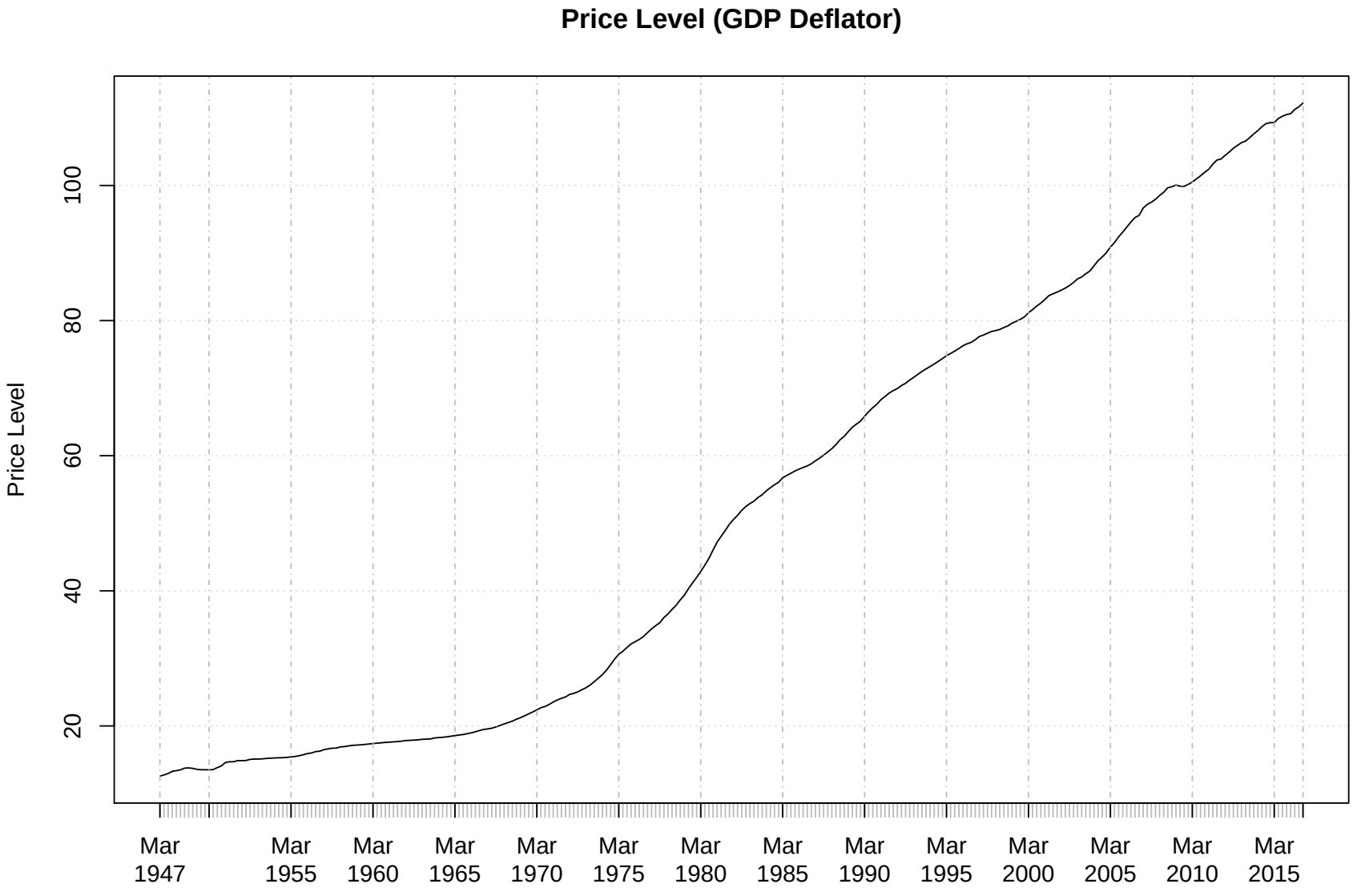
Table 14: Taylor Rule, Usual vs HAC standard errors

	Dependent variable:	
	ffr	
	dynamic linear	*coefficient test*
	(1)	(2)
L(ffr)	0.926***	0.926***
	(0.023)	(0.028)
inf	0.079**	0.079
	(0.036)	(0.064)
gap	0.100***	0.100***
	(0.024)	(0.029)
Constant	0.189*	0.189
	(0.101)	(0.140)
Observations	248	
R^2	0.945	
F Statistic	1,403.876*** (df = 3; 244)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Sources of nonstationarity

- Data nonstationary if $E[Y_t] \neq E[Y_s]$ or $Cov(Y_t, Y_{t+h}) \neq Cov(Y_s, Y_{s+h})$ for $s \neq t$
- Fails when value is deterministic function of time
- Trends: many series grow (or shrink) over time
  - Mean not a constant
  - GDP much higher now than in past
- Breaks: behavior of series differs after some time period
  - Dollar price of gold constant at \$35/oz 1945-1971, then increasing and highly variable after
- Seasonality: behavior of series different in different quarters/months/day of the week
  - Retail sales always largest in December
- Fails when dynamics unstable
  - Series has no well-defined mean or variance to return to
  - E.g. Random walk: $Y_t = Y_{t-1} + e_t$, $E[e_t|Y_{t-1}] = 0 \; \forall \; t$
  - Has conditional mean $E[Y_{t+h}|Y_t] = Y_t$ for $h > 0$, but no time invariant distribution can satisfy above
  - Same true of any unstable ARMA process

A trending time series: Prices, 1947-2016



Dealing with nonstationary data

- Running a regression using a nonstationary time series won't produce a consistent estimate
- Tough luck for macroeconomists and forecasters: no reason future has to be like the past
- But there is hope, if source of nonstationarity is known
  - Transform series so that transformed series is stationary
- When mean and variance known deterministic functions of time μ_t and σ_t^2, can just normalize
  - $\frac{Y_t - \mu_t}{\sigma_t^2}$ is stationary
- In any model built out of stationary components, can solve for the stationary variable
- For random walk $Y_t = Y_{t-1} + e_t$, Y_t nonstationary but $\Delta Y_t = e_t$ is stationary
- Transformed variables are stationary but regression equation may have different interpretation
  - GDP not stationary but GDP growth might be
- Differences, ratios, etc can be used

Trend stationarity

- Simplest case is deterministic trend
  - $Y_t = \beta_0 + \beta_1 t + u_t$ for some stationary (not necessarily independent) u_t
- Can also do quadratic trend $\beta_0 + \beta_1 t + \beta_2 t^2$, exponential trend etc
- (Usually better to take logs and model as linear trend for exponentially growing series like many macroeconomic variables)
- If we knew coefficients, we could just subtract trend
- Solution: put trend as covariate in regression

Inference with time trends

- With time trends, inference is nonstandard: book says not to trust R^2, in fact also don't trust SEs for these variables
- It's not stationary, so strictly speaking consistency/normality results for stationary data don't apply, but they do still hold
- In fact, coefficient on trend converges faster than usual, so asymptotically can ignore estimation error
- Including trend as regressor equivalent to regressing both Y and X on t, running regression on residuals
- Usual (or robust, if errors heteroskedastic or serially correlated) standard errors from this equivalent procedure are asymptotically valid

Making data stationary

- Strategy of including deterministic function of time in regression generalizes
- Seasonality
  - Include dummy for month/quarter/day of week in regression to account for seasonal changes in mean
- Breaks
  - Include dummy for after/before break date
- Other approach is transforming series directly
  - Most government statistics reported already de-seasonalized
  - Can also do something like $X_t - L^4 X_t$ change in series over one year
  - Since both in same season, seasonal mean removed

Integrated Series

- Series where source of nonstationarity is random much harder to deal with
- Random walk $Y_t = Y_{t-1} + e_t$ typical example of series with **stochastic trend**
- Series drifts far away from any particular value, but this drift is not on preset path
- With constant, $Y_t = b_0 + Y_{t-1} + e_t$, expected to increase by b_0 each period, but doesn't stay close to line
- Remove both trends by differencing: $\Delta Y_t = b_0 + e_t$ stationary
- Any series Y_t so that Y_t nonstationary, ΔY_t stationary, called **integrated** (of order 1)
- Asset prices, exchange rates, many macro variables seem to look like this
- Big concern because standard inference invalid if variable nonstationary

Removing integration

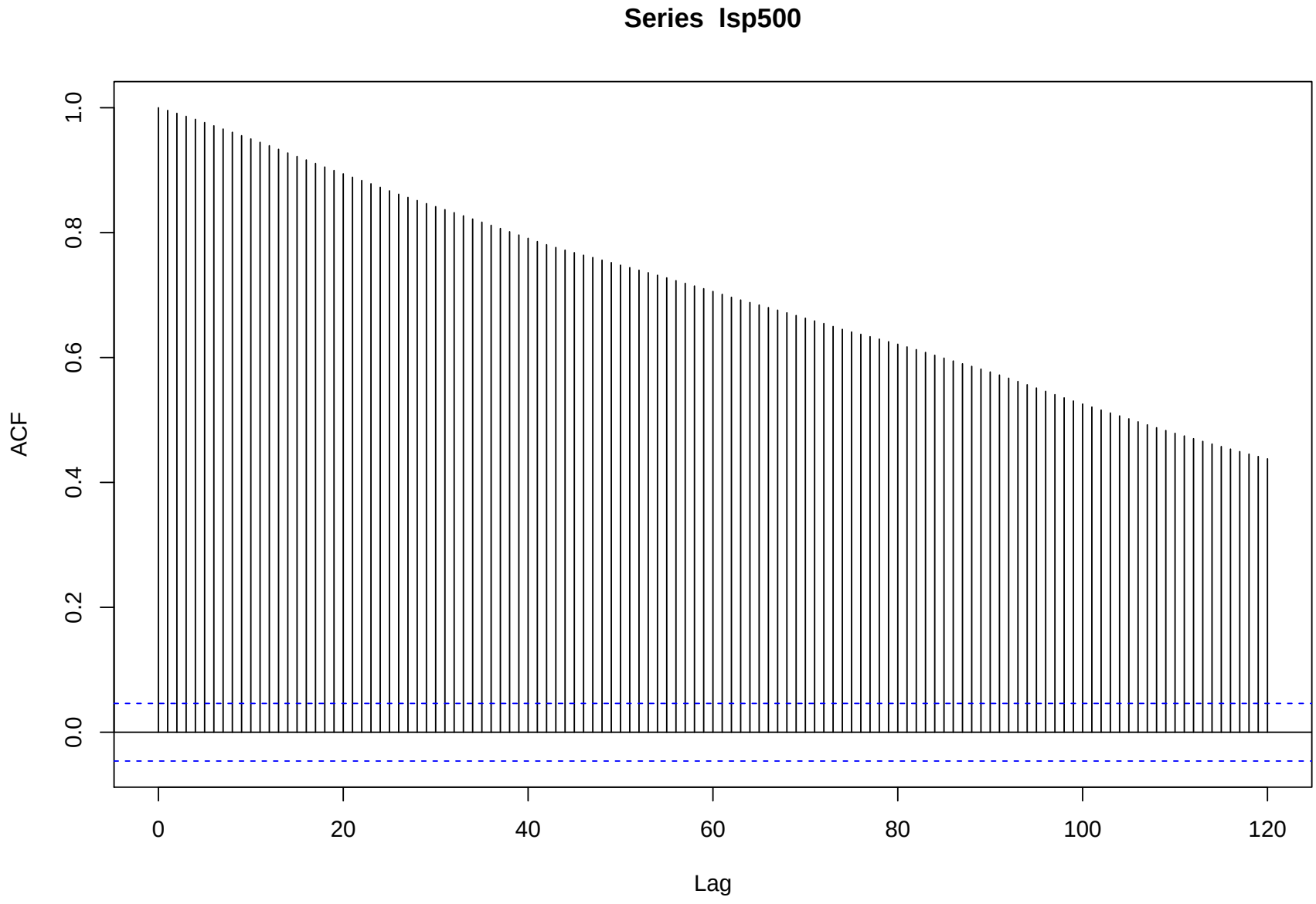
- Failing to include trend if needed, or account for integration, creates bias in regression
  - Will get strong relationship (small p value, high R^2) if both series trending
  - Called "spurious" or "nonsense" correlation
- Two trending series regressed on eachother will be correlated, even if no structural relationship
  - "Time" is omitted variable
- Simplest solution is to detrend: differencing removes stochastic and linear trends
- Other detrending methods may have different properties
- Many macroeconomists swear by something called the HP filter
- Recent paper title: "Why you should never use the HP filter"

A tool for detecting dependence: the Autocorrelation Function (ACF)

- ACF measures dependence between time periods
- Autocovariance is $V(h) = Cov(Y_t, Y_{t+h})$
- Should go to 0 and rapidly if series is weakly dependent
- Traditional to divide by variance to get correlation instead of covariance
  - $\gamma(h) = Corr(Y_t, Y_{t+h}) = \frac{V(h)}{V(0)}$
- Estimate by sample correlations, display chart
- If it goes towards 0 quickly consistent with weak dependence
- If not, may need to adjust series before using in regression
- Shape may also be informative about structure
  - e.g. AR(1) $Y_t = \rho Y_{t-1} + e_t$ has ACF $\gamma(h) = \rho^h$
  - For random walk, acf declines linearly
  - For MA(q), drops to 0 after q periods
- *acf*() in *R*

Autocorrelation function of daily log price, S&P 500 (1-04-2010-Last Week)

Staring at the ACF



Gapminder

For each of 142 countries, the package provides values for life expectancy, GDP per capita, and population, every five years, from 1952 to 2007.

Working with the data

```
gapminder %>%
  filter(year == 2007) %>%
  group_by(continent) %>%
  summarise(lifeExp = median(lifeExp))
```

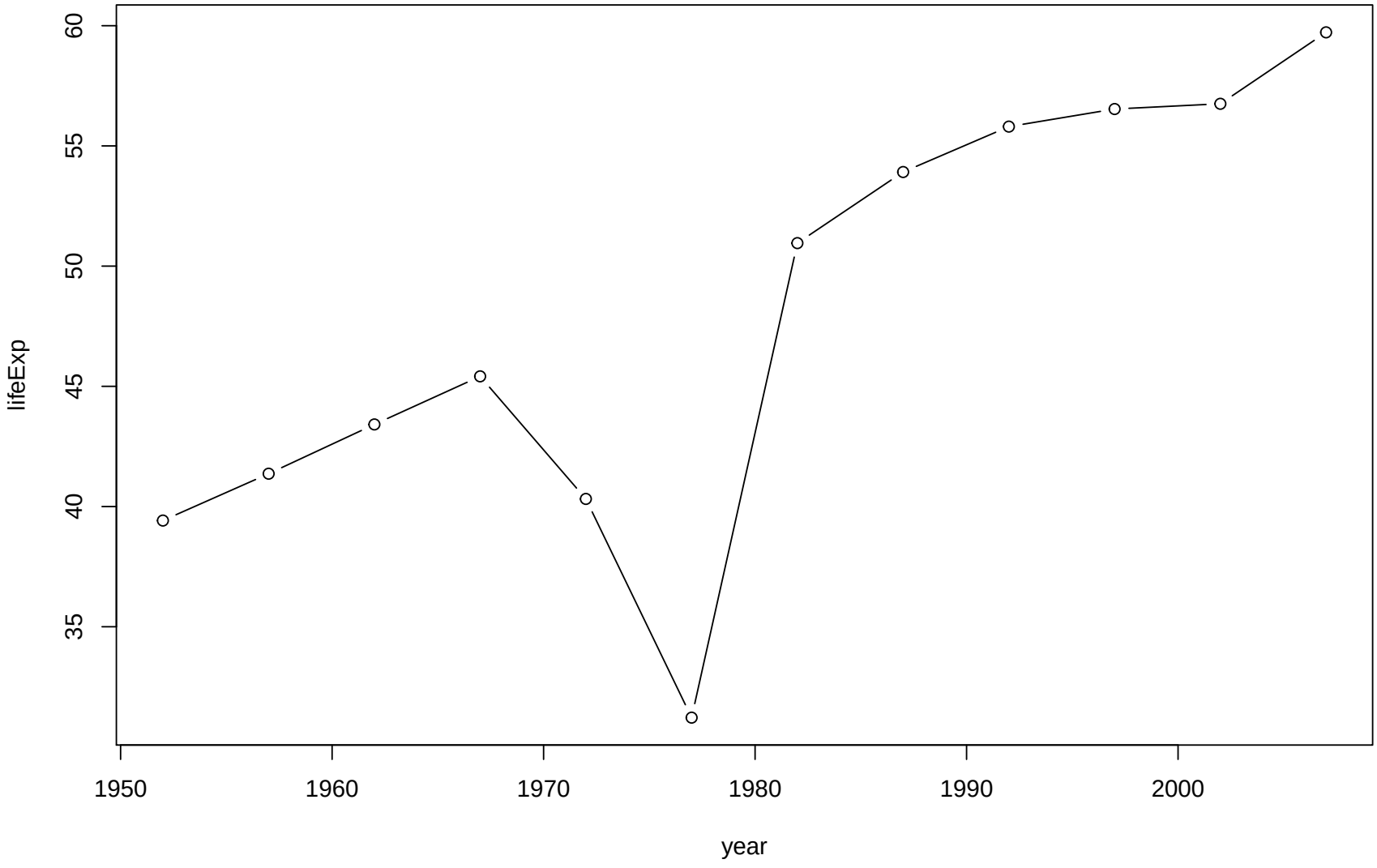
```
## # A tibble: 5 × 2
##   continent lifeExp
##      <fctr>   <dbl>
## 1    Africa 52.9265
## 2  Americas 72.8990
## 3      Asia 72.3960
## 4    Europe 78.6085
## 5   Oceania 80.7195
```

Working with the data

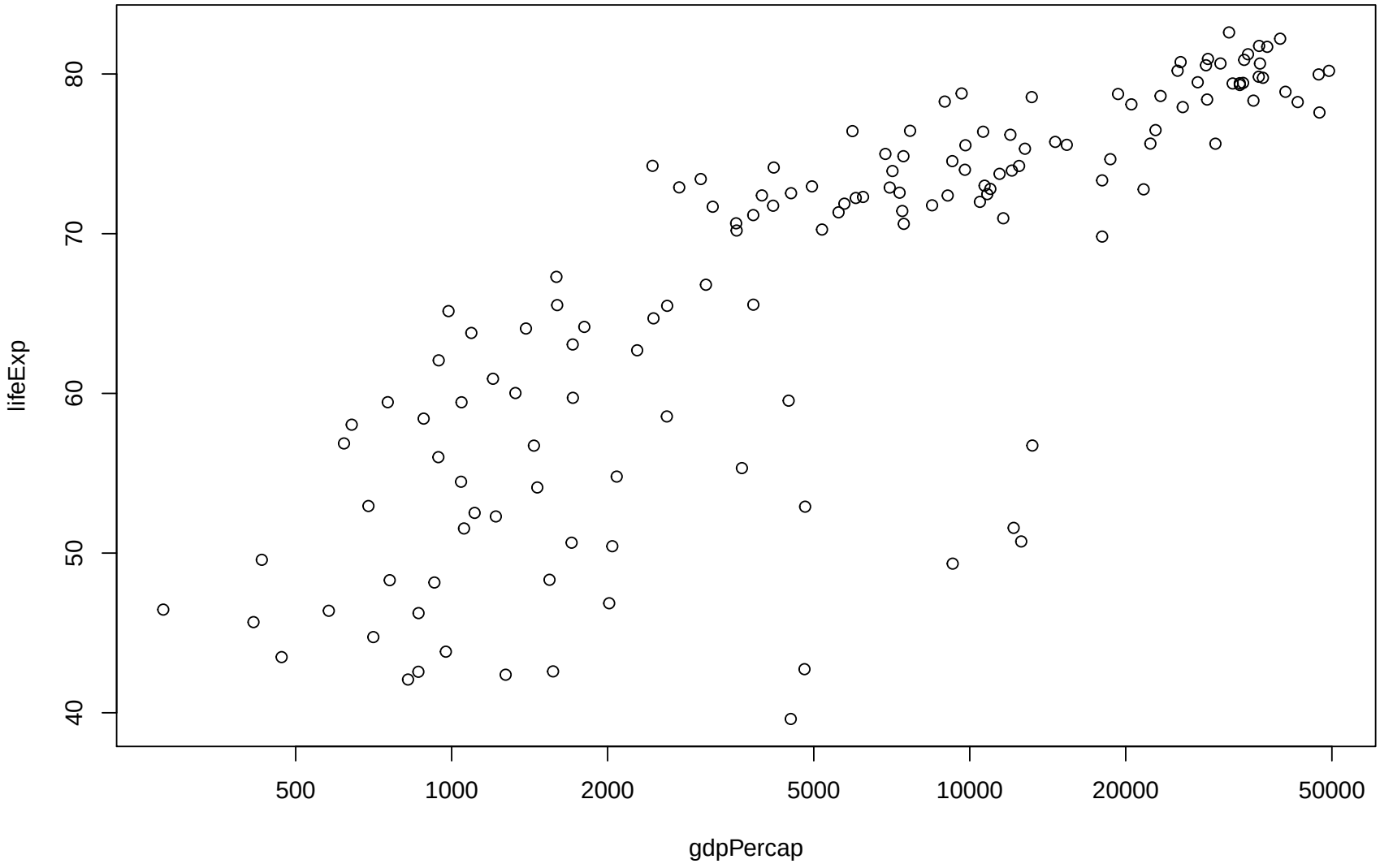
```
##
##   Africa Americas     Asia   Europe  Oceania
##      624      300      396      360       24
```

```
##   continent lifeExp
## 1    Africa 47.7920
## 2  Americas 67.0480
## 3      Asia 61.7915
## 4    Europe 72.2410
## 5   Oceania 73.6650
```

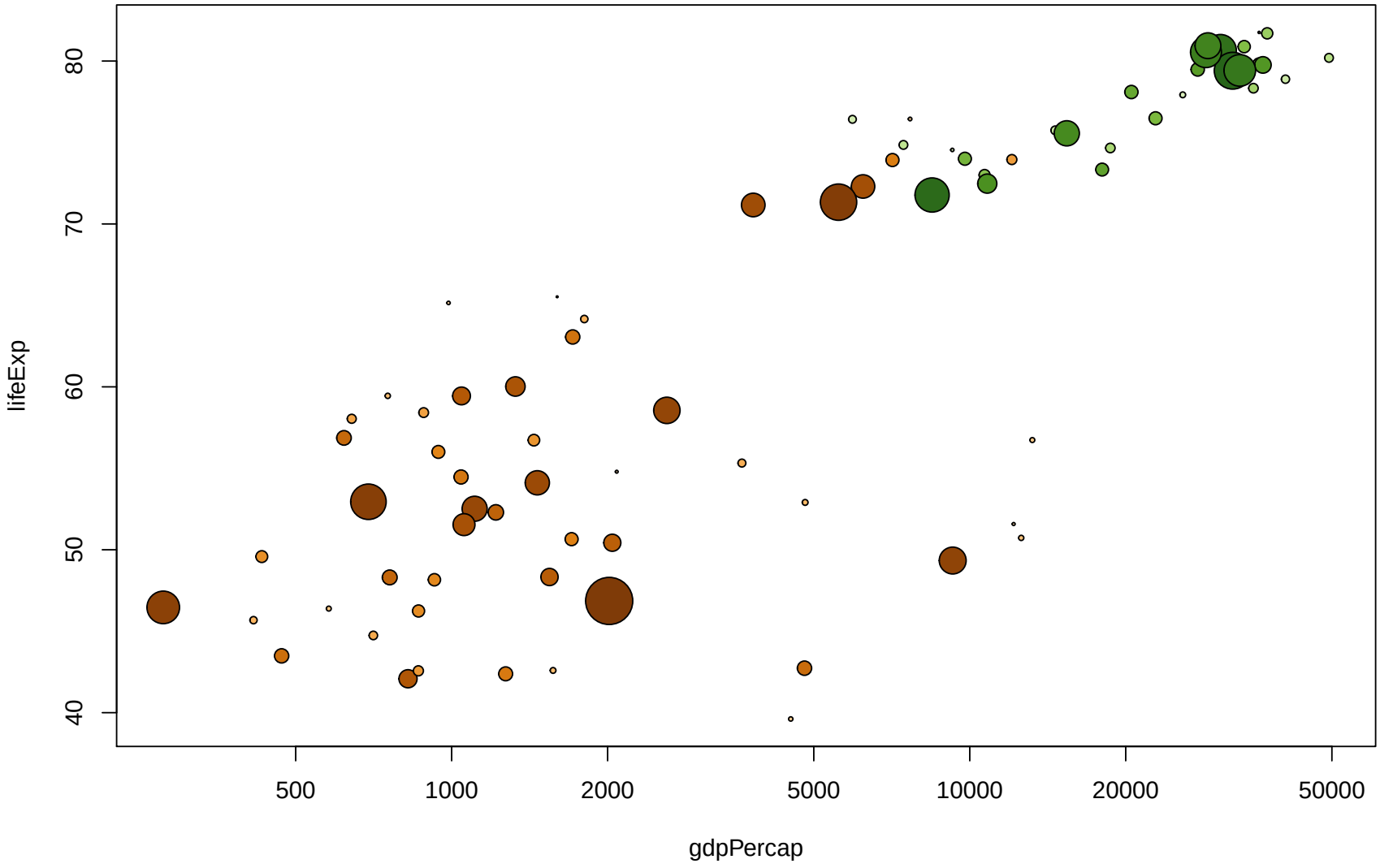
Working with the data



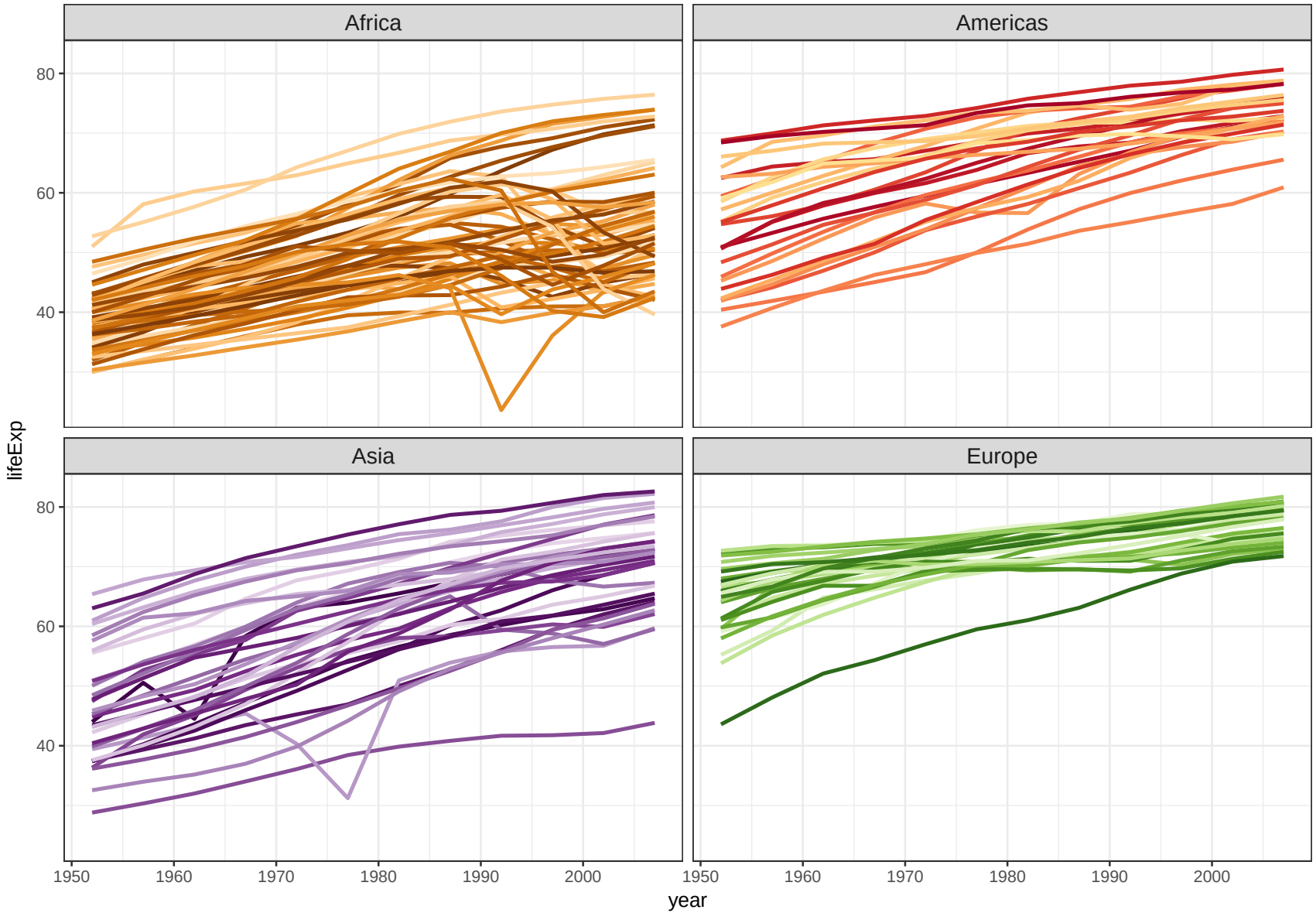
Working with the data



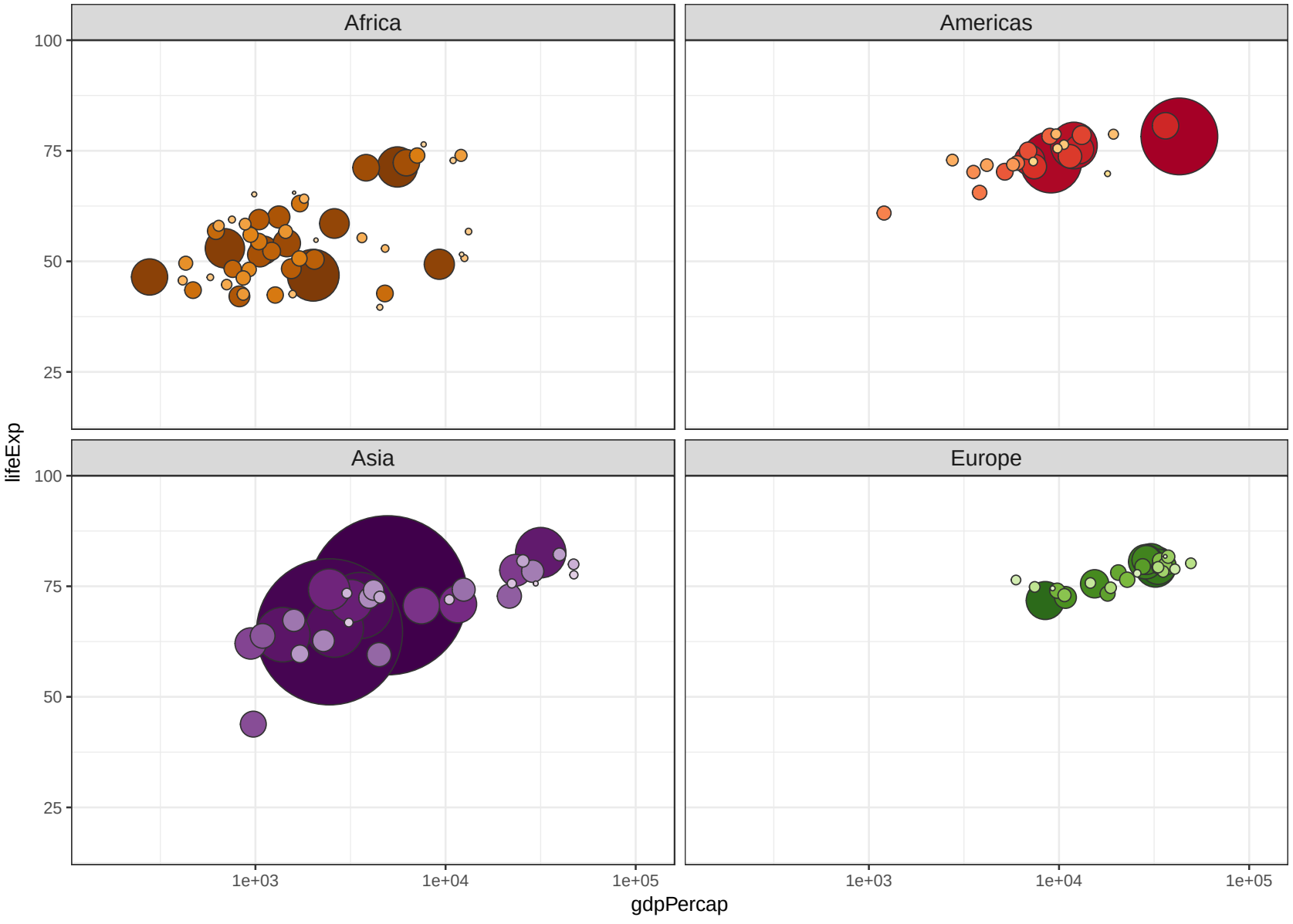
Working with the data



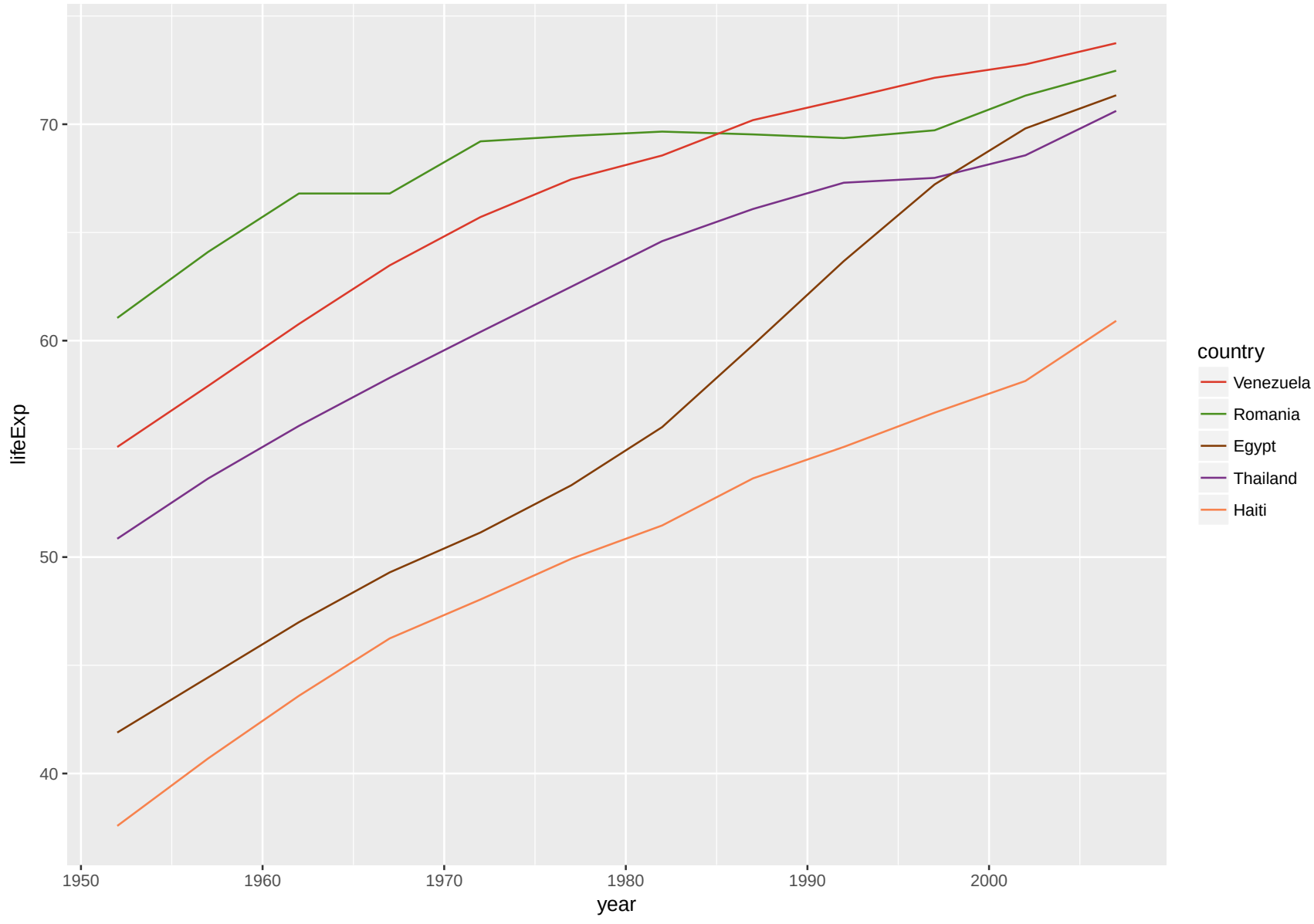
Bubble plots



Working with different graph types



Sub setting for 5 countries



Summary

- With dependent data, can still learn from standard regression methods
- If data stationary and weakly dependent, have consistency and asymptotic normality
- If errors still correlated, can use robust variance estimate
- If nonstationary, can model form of nonstationarity and get rid of it
  - Include time trend (or season dummies, etc) in regression
  - Difference or filter data
- Once stationarity and weak dependence restored, estimation, inference, interpretation

Lecture 7 Machine Learning

- Lots of hype. But ML studies the design of algorithms that can learn.
- Learning means it improves its performance as it receives more data.
- Typical machine learning tasks are concept learning, classification, function learning or "predictive modeling", clustering and finding predictive patterns
- These tasks are learned through available data that were observed through experiences or instructions
- The ultimate goal is to improve the learning in such a way that it becomes automatic, so that humans don't need to interfere any more.

Input knowledge

- data that contain features (wage, sector, etc)
- in R this is just a data.frame()
- Use dim() or str() or summary() to figure out size of the data set

Notion of 'learning'

- Supervised vs Vs UnSupervised
- Supervised: finding a function which can be used to assign a class or value to unseen observations given a set of labeled observations. (Regression)
- Unsupervised: When you don't need labels (Clustering). Problem: you can't measure the performance of your model directly.
- Semi-supervised: when you have only a few labels but let the algorithm produce most of the clusters itself.

Learning is applying a function to inputs to generate an output

- that might seem like it's just regression.
- But it's not. The goal is model building for prediction.

Example: Using Wage data, predict the wage for a 60 year old worker. The prediction is ML.

```
library(ISLR)
data(Wage)
# Build Linear Model: lm_wage
lm_wage <- lm(wage ~ age, data = Wage)
# Define data.frame: unseen
unseen <- data.frame(age = 60)
# Predict the wage
predict(lm_wage, unseen)
```

```
##        1
## 124.1413
```

Interpreting the result

Wage for a 60 year old worker is predicted to be 124 dollars per day.

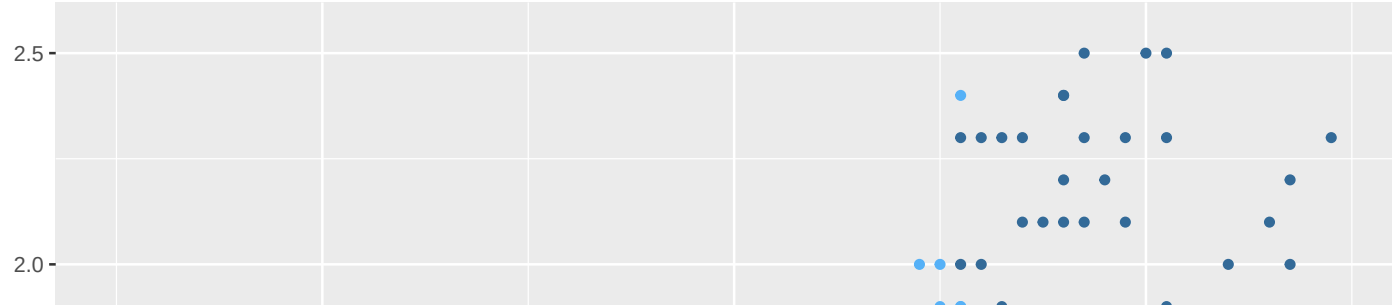
Common ML approaches to solving problems

1. Classification: predicting whether a new observation falls into a certain category. Very useful for medical diagnosis and facial recognition. The output here is qualitative and the classes are known.
2. Regression: estimate a function that maps inputs to outputs based on a function and predict from it. Output is quantitative. Eg predicting weight and height. You need to find good values for the coefficients from previous values.
3. Clustering. Group similar objects while making sure clusters are distinct. Here you don't need knowlege about labels. K-means is a good method. It clusters your data in K clusters.

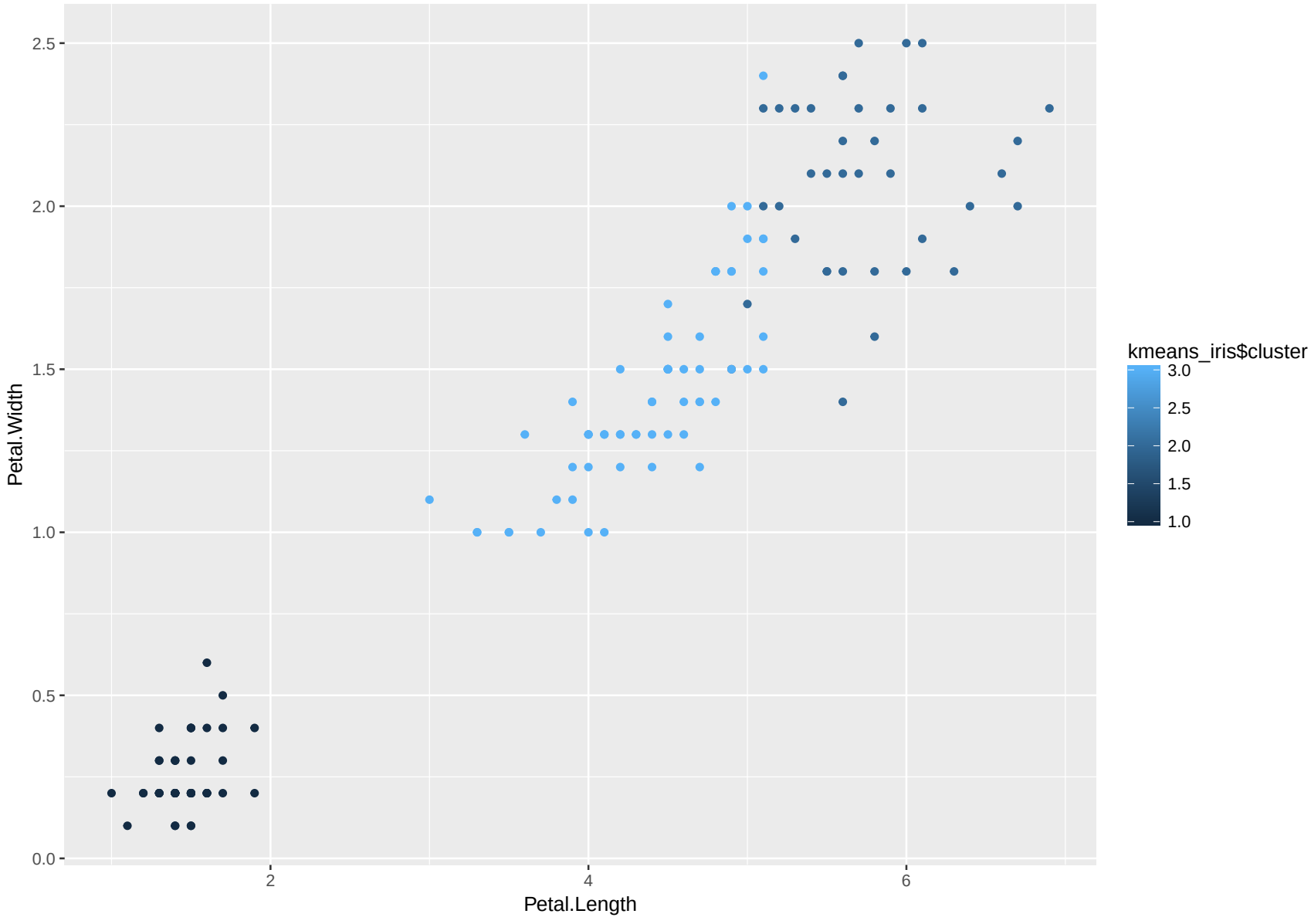
Regression and clustering are quite similar. You're trying to find a function which can be used to assign a class or value to unseen observations. T

Clustering (code)

```
# Set random seed. Don't remove this line.
set.seed(1)
#Load the data
data(iris)
# Chop up iris in my_iris and species
my_iris <- iris[-5]
species <- iris$Species
# Perform k-means clustering on my_iris: kmeans_iris
kmeans_iris <- kmeans(my_iris, 3)
# Compare the actual Species to the clustering using table
table(species, kmeans_iris$cluster)
# Plot Petal.Width against Petal.Length, colouring by clus
ggplot(data=my_iris)+aes(x = Petal.Length, y = Petal.Width,
```



Clustering Output



Why is machine learning important for public policy?

- Example. A common ML approach is to train a system by showing it a vast quantity of data on, say, students and their achievements.
- The software chews through the examples and learns which characteristics are most helpful in predicting whether a student will drop out.
- Once 'trained', it can study a different group and accurately pick those at risk.
- State systems are stuffed with administrative data and newer data types amenable to that kind of analysis.
- taxation and expenditure systems/health and social protection systems/newer sources of data (eg toll data, phone data)

Training

- Training is just regression based on a very large dataset, you've seen that before.
- The difference is the method used to train, the solution system, and the way these methods are applied.

Prediction & training & application

- In hospitals doctors try to predict heart attacks so they can act before it is too late.
- Manual systems correctly predict around 30%.
- A machine-learning algorithm created by Sriram Somanchi of Carnegie Mellon University and colleagues, and tested on historic data, predicted 80%—four hours in advance of the event, in theory giving time to intervene.

Example: Understanding racial disparities in New York City's Stop and Frisk Policy

- Annals of Applied Statistics, 2016.
- Stop and frisk: NYC Zero tolerance policy. Idea is that by stopping smaller transgressions, larger ones can be avoided.
- Problems with this – most of the time (90% +) nothin happens. The person is hassled and time is wasted. – The people most likely to be stopped are blacks and hispanics (+80%)
- Problem: how to see if it is discriminatory?

Discrimination or fair policy?

- Previous studies analysed the racially motivated stops using instruments.
- These guys looked at 3 million stops between 2008 and 2012, focusing in on 760,000 weapon stops.
- Legally cops can't stop people unless they think there's a likelihood of something going down.
- this is the likelihood based only on information available to officers prior to the stop decision that the stopped individual has a weapon.
- They find that in 43% of the approximately 300,000 CPW stops between 2011 and 2012, there was at most a 1% chance of finding a weapon on the suspect.

Discrimination or fair policy?

- They find that blacks and Hispanics were disproportionately involved in low hit rate stops
- They trace this disparity to two factors: (1) the highly localized nature of the policy, and (2) discriminatory enforcement.
- High crime areas, particularly NYC public housing, have lower stop thresholds, presumably reflecting more aggressive efforts to reduce crime in those locations. Since these areas are home to large numbers of blacks and Hispanics, members of these groups were disproportionately impacted by stop standards that differed by location.

One of these isn't machine learning. Which?

1. Given a viewer's shopping habits, recommend a product to purchase the next time she visits your website.
2. Given the symptoms of a patient, identify her illness.
3. Predict the USD/EUR exchange rate for February 2016.
4. Compute the mean wage of 10 employees for your company.

Using R For k-Nearest Neighbors (KNN)

- Instance-based learning. new data are classified based on stored, labeled instances.
- the distance between the stored data and the new instance is calculated by means of some kind of a similarity measure.
- This similarity measure is typically expressed by a distance measure such as the Euclidean distance, cosine similarity or the Manhattan distance. In other words, the similarity to the data that was already in the system is calculated for any new data point that you input into the system
- Then, you use this similarity value to perform predictive modeling. Predictive modeling is either classification, assigning a label or a class to the new instance, or regression, assigning a value to the new instance.
- Whether you classify or assign a value to the new instance depends of course on your how you compose your model with KNN.

More on KNN

- distance of the new point to all stored data points has been calculated, the distance values are sorted and the k-nearest neighbors are determined
- The labels of these neighbors are gathered and a majority vote or weighted vote is used for classification or regression purposes
- In the case of regression, the value that will be assigned to the new data point is the mean of its k nearest neighbors.

Example (code)

```
library(class) # This has the KNN algorithm
attach(iris)
summary(iris[c("Sepal.Width", "Petal.Width")])
normalize <- function(x) {
  num <- x - min(x)
  denom <- max(x) - min(x)
  return (num/denom)
}
iris_norm <- as.data.frame(lapply(iris[1:4], normalize)) #
summary(iris_norm)
```

Example (output)

```
library(class) # This has the KNN algorithm
attach(iris)
normalize <- function(x) {
  num <- x - min(x)
  denom <- max(x) - min(x)
  return (num/denom)
}
summary(iris[c("Sepal.Width", "Petal.Width")])
iris_norm <- as.data.frame(lapply(iris[1:4], normalize)) #
summary(iris_norm)
```

Example: Training

```
set.seed(1234)
ind <- sample(2, nrow(iris), replace = TRUE, prob = c(0.67,
iris.training <- iris[ind == 1, 1:4]
iris.test     <- iris[ind == 2, 1:4]
iris.trainLabels <- iris[ind == 1, 5]
iris.testLabels  <- iris[ind == 2, 5]
```

Example: The model itself

```
iris_pred <- knn(train = iris.training, test = iris.test, c
iris_pred # Get some results back. The result of this comm
```

```
##  [1] setosa     setosa     setosa     setosa     setosa
##  [7] setosa     setosa     setosa     setosa     setosa
## [13] versicolor versicolor versicolor versicolor versico
## [19] versicolor versicolor versicolor versicolor versico
## [25] virginica  virginica  virginica  virginica  versico
## [31] virginica  virginica  virginica  virginica  virgini
## [37] virginica  virginica  virginica  virginica
## Levels: setosa versicolor virginica
```

Evaluating the model

```
df <- data.frame(iris_pred, iris.testLabels)
names(df) <- c("Predicted Species", "Observed Species")
df
```

```
##    Predicted Species Observed Species
## 1             setosa            setosa
## 2             setosa            setosa
## 3             setosa            setosa
## 4             setosa            setosa
## 5             setosa            setosa
## 6             setosa            setosa
## 7             setosa            setosa
## 8             setosa            setosa
## 9             setosa            setosa
## 10            setosa            setosa
## 11            setosa            setosa
## 12            setosa            setosa
## 13        versicolor        versicolor
```

Generating a cross table.

Result: Pretty good prediction

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
##
## Total Observations in Table:  40
##
##
##
##                  | iris_pred
## iris.testLabels |     setosa | versicolor |  virginica
```

Why is this important?

- KNN is just one, very simple example of a range of techniques being developed to understand patterns in large data *and* make predictions from it.
- Iris is a very small dataset, and so is tractable. We can use massive datasets and the KNN algorithm will do a pretty good job.

What about causal inference?

- In policy world we often want to ask what causal effects will be of, say, a price change or a tax increase (or whatever)
- This really ammounts to changing the underlying training data
- Currently Machine Learning doesn't do well with this, but lots of people are working on making it better.

Lecture 8. Modeling for public policy

Motivation

We're typically interested in statistical models. Statistical models are used for

1. Identifying patterns in data
2. Classifying events
3. Untangling multiple causal influences
4. Assessing strength of evidence.

Statistical models give you more power to analyse the real world than simple tests than (say) the t test.

Literature in Computer science very different to economics

- focus in this literature is typically on methods that "work," more than on deriving asymptotic (large sample) results of the type that are common in the econometrics and mathematical statistics literature.
- There is also less emphasis on confidence intervals and standard errors.
- Heavy emphasis on out-of-sample comparisons, in particular cross- validation.
- Also less emphasis on causal effects as opposed to prediction.

See Breiman, L. (2001a), "Statistical Modeling: The Two Cultures," Statistical Science, Vol. 16(3): 199-215.

Types of Machine Learning

- Unsupervised

– Clustering – Principal components analysis

- Supervised

– Regression – K nearest neighbour – Trees – Nearest neighour – Support vector machines – Random Forests – Ensemble Methods

- Semi-Supervised

A model is a representation for a purpose.

- representation: it stands for something in reality we'd like to know more about.
- Purpose: we care about attributing something like causality, etc.

These models are useful for one or more purposes, but have severe and important limitations. We build models because they are much more understandable than the real world.

Models can be made of paper, clay, plastic, cardboard, and mathematics. They needn't be equations at all. Think of blueprints for a house or a model plane, or a flight simulator to train pilots.

Statistical models are versions of mathematical models, informed by and based on data.

We typically use these things to ask defined questions we call hypotheses. We then accept or reject the hypotheses based on what the output of the model is.

We can also move from experimental results to a prediction.

Example 1

Say you've a model of student test scores by gender.

- Private schools with 10% more boys than girls score 10% below the national average.
- Private schools with 10% more girls than boys score 10% above the national average.
- What happens if we double the girls to boys ratio? That is, we increase from 10% more boys than girls to 20% more boys than girls?

Use this very common linear strategy for making your prediction:

If the change in output going from 0 to 1 is X, then the change in output going from 1 to 2 will also be X.

Find the change in output for each of the inputs in turn. Add up the individual changes to get the anticipated change when more than one input is changed.

Can you see how this is related to the *marginal* concept we studied last semester?

Objects for statistical modeling

You need 3 things to get a model working. Again, you'll typically want to use a model to predict, or account for, some variable.

- Formulas. These relate variables to one another. They are causal statements. EG: WAGE $\sim$ EXPERIENCE $+$ GENDER. This says your wage is explained (we think) by the number of years of experience you have, and your gender. The squiggle yoke is called 'tilde'.
- Data frames–a collection of variables. Each variable gets a column, this column gets a name. The rows are cases (sometimes called elements).
- Functions. These are the building blocks of models and produce the outputs of the models. You need formulas and data frames to make functions work effectively.

Example 2

```
library(mosaic) # This attaches the library. It has interes

mean(wage ~ sector, data = CPS85) # This gives us the avera
```

```
##  clerical     const     manag     manuf     other         p
##  7.422577  9.502000 12.704000  8.036029  8.500588 11.947
##   service
##  6.537470
```

```
my_model <- lm(wage ~ exper + sector, data = CPS85) # The u
summary(my_model) # Let's spend some time looking at the in
```

```
##
## Call:
## lm(formula = wage ~ exper + sector, data = CPS85)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

Thinking (again) about interpreting the formula

The formula is a bit like a sentence. EG WAGE $\sim$ SECTOR is equivalent to

1. WAGE as a fucntion of SECTOR
2. WAGE accounted for by SECTOR
3. WAGE modeled by SECTOR
4. WAGE explained by SECTOR
5. WAGE given by SECTOR
6. WAGE broken down by SECTOR

A digression on the NULL hypothesis for a moment.

What we're actually trying to do the whoe time is explain variation.

This means we must try to distinguish random variation from purposeful, or causal, variation.

There are millions of reasons wages vary, and some of them can't be measured. But a lot of the time the explanatory variables we choose do explain a fair amount of the variation.

The rest, the unexplained bit, we think of as being essentially random.

So what you're doing all the time in this work is comparing your results to what you would get if the null hypothesis were true. If your results are very different, you're rejecting the null hypothesis, that it's all random. If your results are almost the same, then you fail to reject the null hypothesis.

You also need to assign a penalty to the null data, but we'll get to that later.

Example

```
library(statisticalModeling) #This package has the data
library(mosaic) # This package alters the 'mean' and sd() c
names(AARP) # Find the variable names in AARP
```

```
## [1] "Age"      "Sex"      "Coverage" "Cost"
```

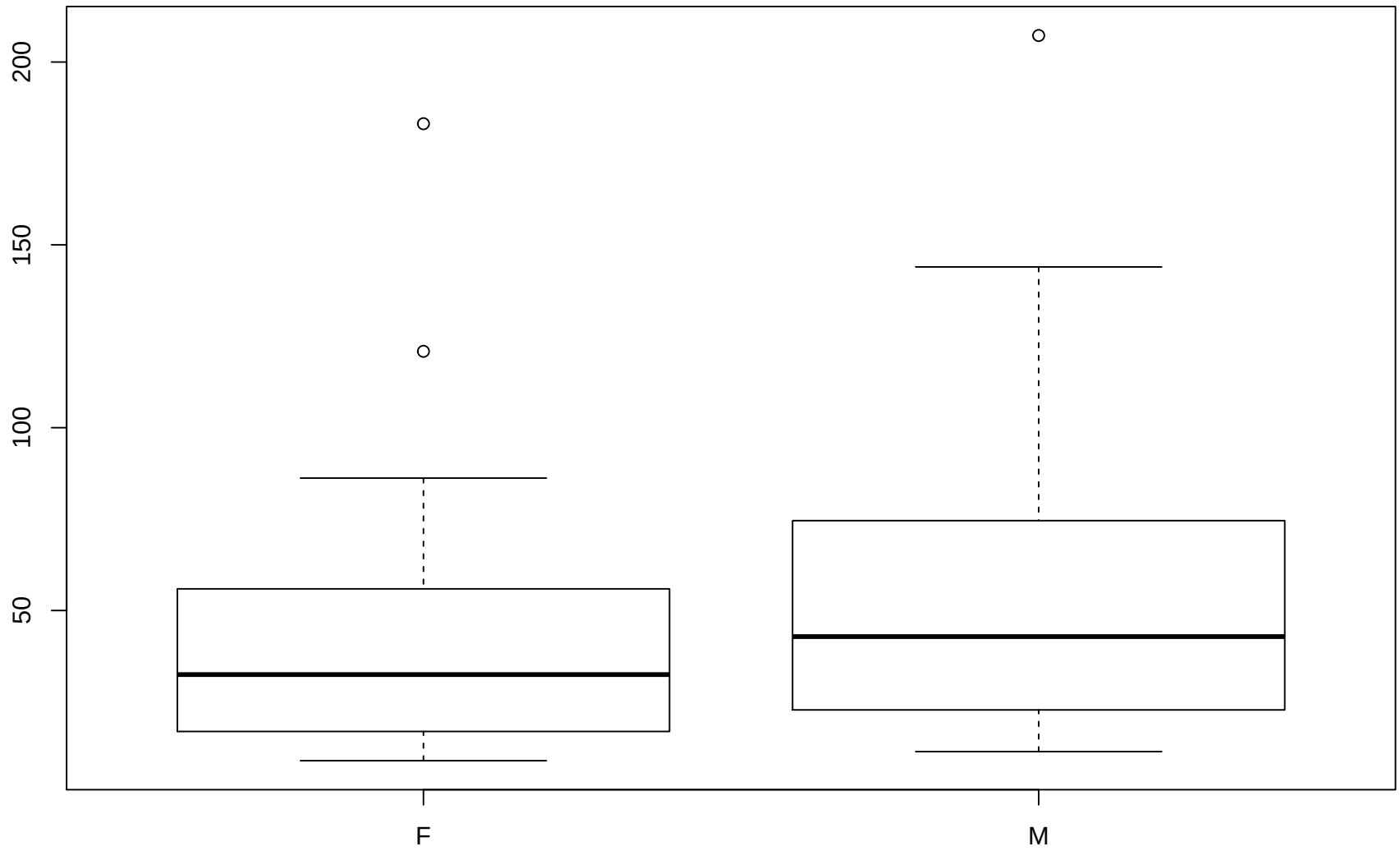
```
mosaic::mean(Cost ~ Sex, data = AARP) # Find the mean cost
```

```
##        F        M
## 47.29778 57.53056
```

Graphics with formulas

You're going to love the graphics, here are two examples.

```
boxplot(Cost ~ Sex, data=AARP) #boxplot of Cost and Sex us
```



Modeling is *always* a process

It is messy. Typically you have ideas, and then you'll test those ideas by designing a model, which you train using your data, which you've cleaned and graphed and understood using summary statistics. You then evaluate the model, test its performance, and then see how your model challenges your ideas.

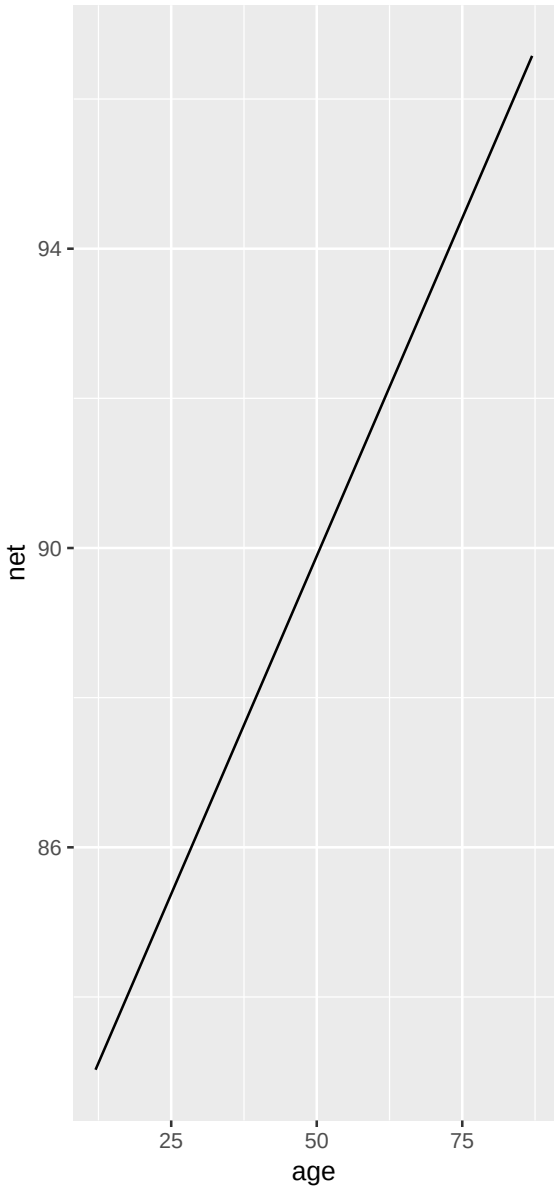
1. The first step is always, always, GET THE DATA.
2. Decide which variables are response and explanatory variables.
3. Select a model architecture. We'll typically use linear regression (lm()). We can also use recursive partitioning.
4. The computer 'trains' or 'fits' the data to the model, and vice versa.

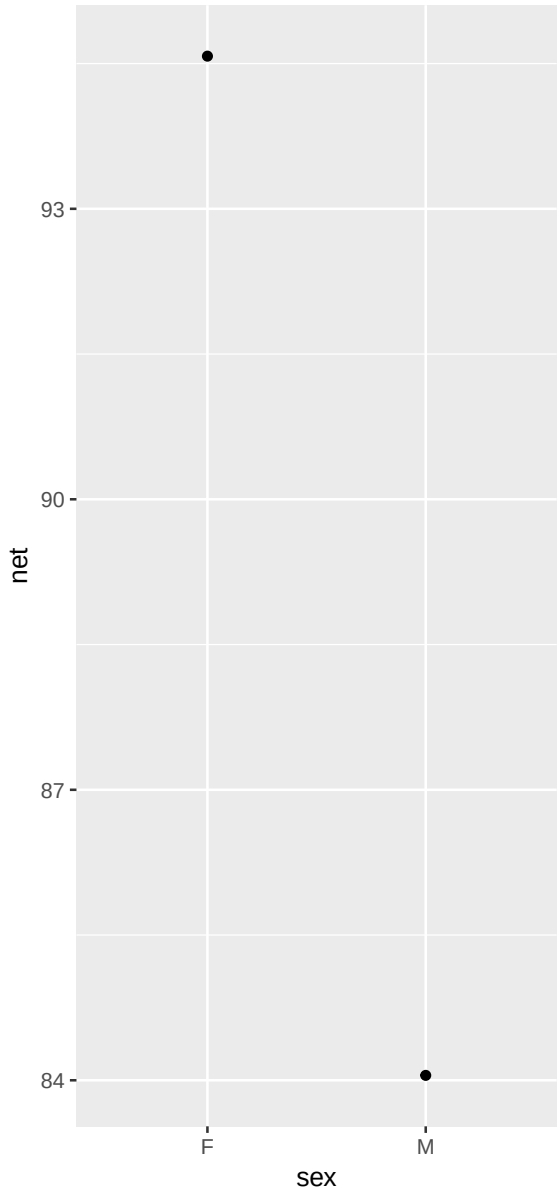
Questions to ask

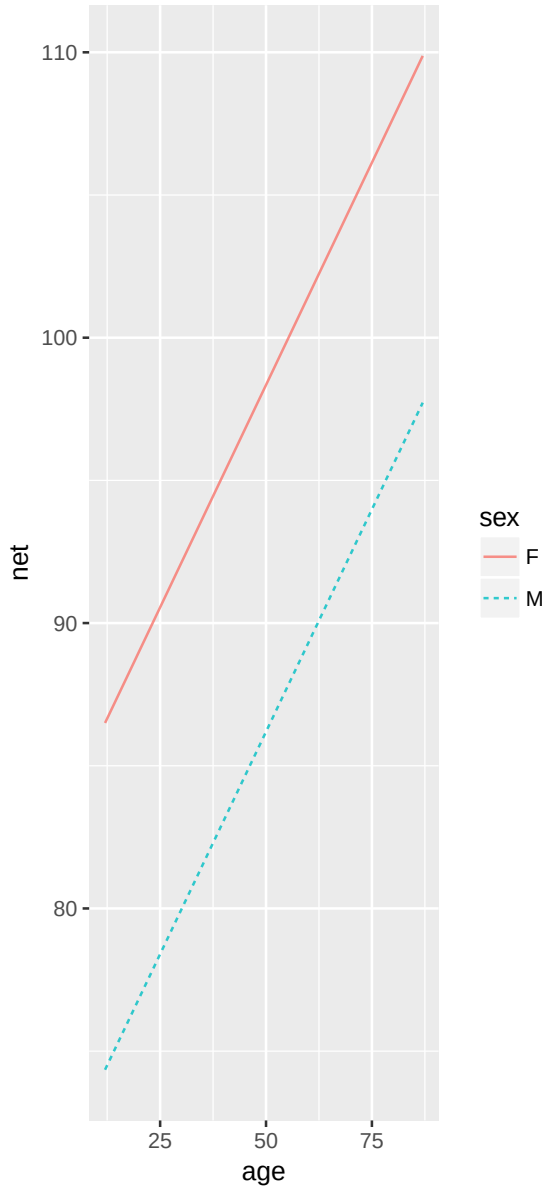
- Why do we want to choose wage as the response variable?
- Choose Educ and Exper as explanatory variables.

Let's use three models to see how different they are.

Example


net
94
90
86
25
50
75
age


net
93
90
87
84
F
M
sex


net
110
100
90
80
25
50
75
age
sex
F
M

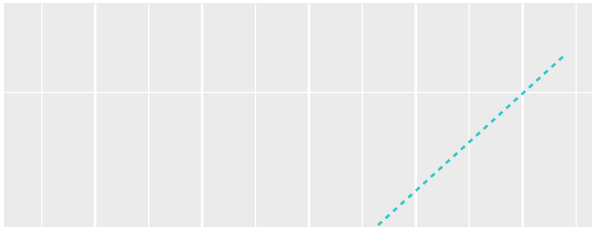
Recursive partitioning

We just used the linear modeling architecture to construct a model of a runner's time as a function of age and sex. There are many different model architectures available. Now we'll build models using the recursive partitioning architecture. The model-building function to use is rpart(), which is analogous to lm() for linear models.

The recursive partitioning architecture has a parameter, cp, that allows you to dial up or down the complexity of the model being built. Without worrying about the details just yet, you can set this parameter as a named argument to rpart().

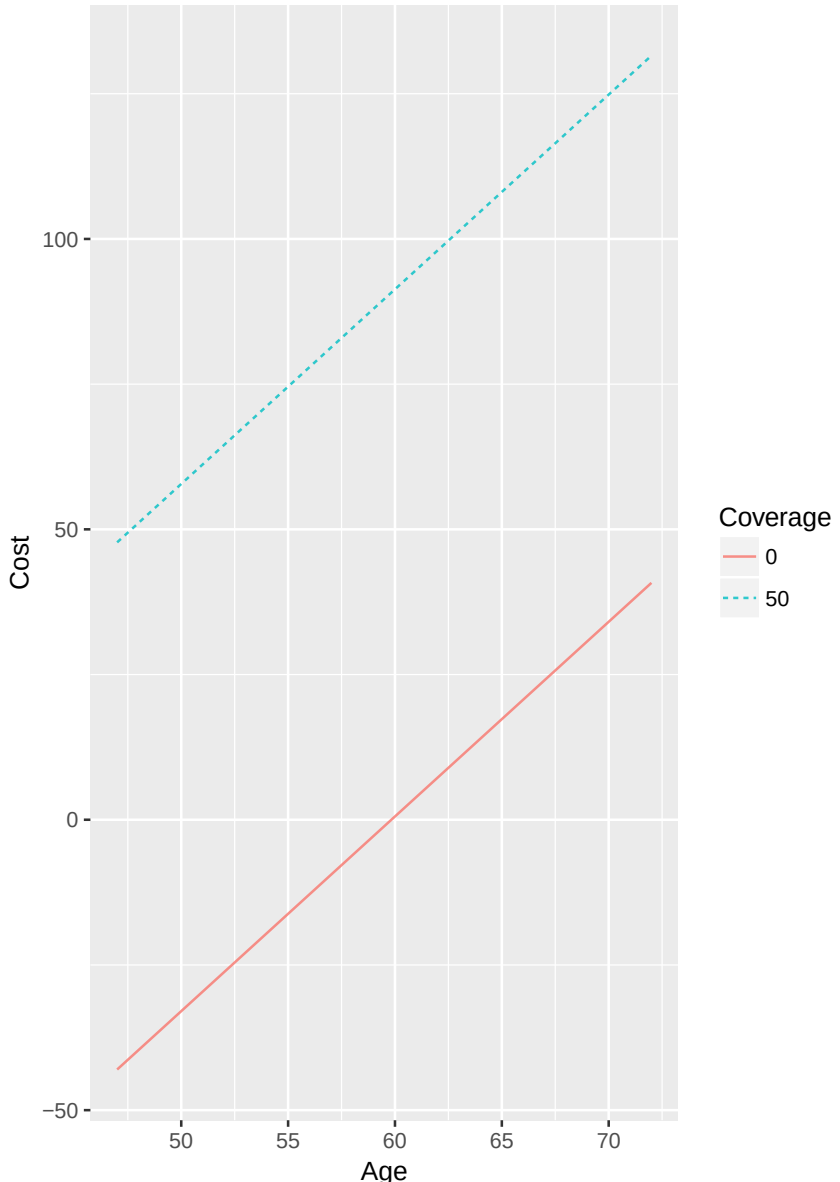
Example (code)

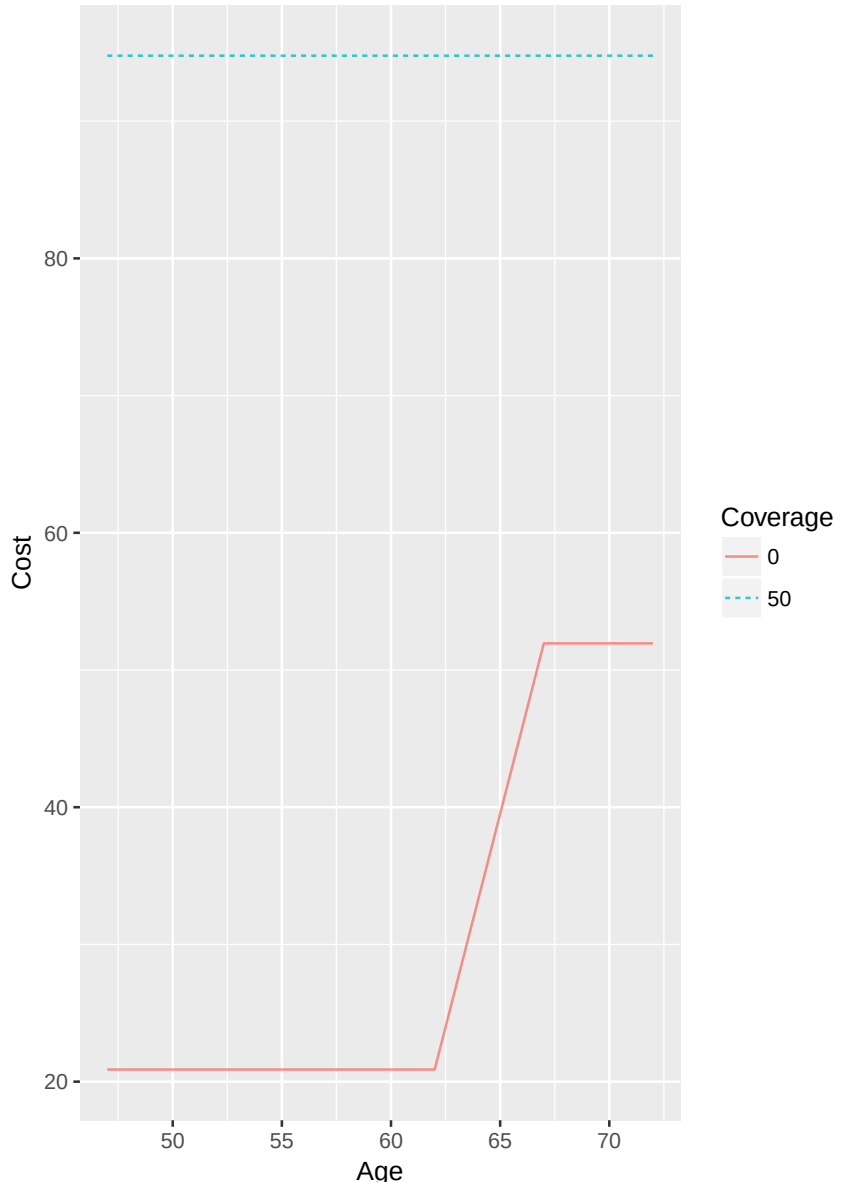
```
library(rpart)
library(mosaic)
library(statisticalModeling)
library(gridExtra)
#Let's use two different models.
model_1 <- lm(Cost ~ Age + Sex + Coverage, data=AARP) #Firs
# Build rpart model: model_2
model_2 <- rpart(Cost ~ Age + Sex + Coverage, data=AARP, cp
#Examine graph of model_1
p1<-fmodel(model_1, ~ Age +  Coverage)
# Examine graph of model_2
p2<-fmodel(model_2, ~ Age + Coverage)
# Display them side by side
grid.arrange(p1, p2, ncol=2)
```



Example (Output)

Note the different stories the different models tell. The decisions the modeler makes have an impact on the results.





Evaluating a model

- You evaluate by providing a model with inputs
- Then you calculate the resulting output. Here you use the predict() function. * Predict is a kind of fortune-telling device. We can use our actual data to figure out how good the model is, and how bad it is. That gives us a measure of 'prediction error'.

Example

```
library(mosaic)
library(statisticalModeling)
# Build a model: insurance_cost_model
insurance_cost_model <- lm(Cost ~ Age + Sex + Coverage, dat
#Construct a data frame: example_vals
example_vals <- data.frame(Age = 60, Sex = "F", Coverage =
# Predict insurance cost using predict()
predict(insurance_cost_model, newdata=example_vals)
```

```
##       1
## 363.637
```

```
# Calculate model output using evaluate_model()
evaluate_model(insurance_cost_model, data=example_vals)
```

```
##   Age Sex Coverage model_output
## 1  60   F      200      363.637
```

Extrapolation

One purpose for evaluating a model is extrapolation: finding the model output for inputs that are outside the range of the data used to train the model.

Extrapolation makes sense only for quantitative explanatory variables. For example, given a variable x that ranges from 50 to 100, any value greater than 100 or smaller than 50 is an extrapolation.

In this exercise, you'll extrapolate the AARP insurance cost model to examine what the model suggests about insurance costs for 30-year-olds and 90-year-olds. Keep in mind that the model outputs might not make sense. Models trained on data can be a bit wild when evaluated outside the range of the data.

Example

```
library(mosaic)
library(statisticalModeling)
# Build a model: insurance_cost_model
insurance_cost_model <- lm(Cost ~ Age + Sex + Coverage, dat

# Create a data frame: new_inputs_1
new_inputs_1 <- data.frame(Age = c(30, 90), Sex = c("F", "M

# Use expand.grid(): new_inputs_2
new_inputs_2 <- expand.grid(Age = c(30, 90), Sex = c("F", "

?expand.grid

# Use predict() for new_inputs_1 and new_inputs_2
predict(insurance_cost_model, newdata = new_inputs_1)
```

```
##         1         2
## -99.98726 292.88435
```

Choosing explanatory variables

What's the purpose of your model?

1. Make preductions about outcomes
2. Running experiments to study relationships
3. Exploring data to 'hunt' out relationships between variables.

Basic choices in model architecture

1. If the response variable is categorical (yes/no, etc) use rpart()
2. If response variable is numerical (eg inflation rate) use lm()

Comparing prediction results

Compare predictive performance of two models, without including a variable, and with. When you train and test a model, you use data with values for the explanatory variables as well as the response variable. Training effectively creates a function that will take as inputs values for the explanatory variables and produce as output values corresponding to the response variable.

If the model is good, when provided with the inputs from the testing data, the outputs from the function will give results "close" to the response variable in the testing data. How to measure "close"? The first step is to subtract the function output from the actual response values in the testing data. The result is called the prediction error and there will be one such error for every case in the testing data. You then summarize that set of prediction errors.

Prediction error for categorical response variables

- When the variables are numerical, it is always possible to compare the results of a model against prdedicted values.
- Imagine 2 models looking at whether marital status can help (or not) understanding data. You need to count how many mistakes the model makes.
- You're still trying to subtract the predicted response rate (or value) from the actual response rate (or value).
- With categorical variables, you're looking at the categorical error rate because you can't really think about subtracting categories.
- What about thinking about probabilities of different levels? We could do that easily using R. Now we can think about the various outputs of the models as the probability of each item happening.
- You combine the probabilities into a 'likelihood' measure, and for a few reasons, you take the log of this number.

This example compares a null model, basically a random guess, to other variable

```
 # Build the null model with rpart()
Runners$all_the_same <- 1 # null "explanatory" variable
null_model <- rpart(start_position ~ all_the_same, data = R

# Evaluate the null model on training data
null_model_output <- evaluate_model(null_model, data = Runn

# Calculate the error rate
with(data = null_model_output, mean(start_position != model
```

```
## [1] 0.5853618
```

```
# Generate a random guess...
null_model_output$random_guess <- mosaic::shuffle(Runners$s

# ...and find the error rate
with(data = null_model_output, mean(start_position != rando
```

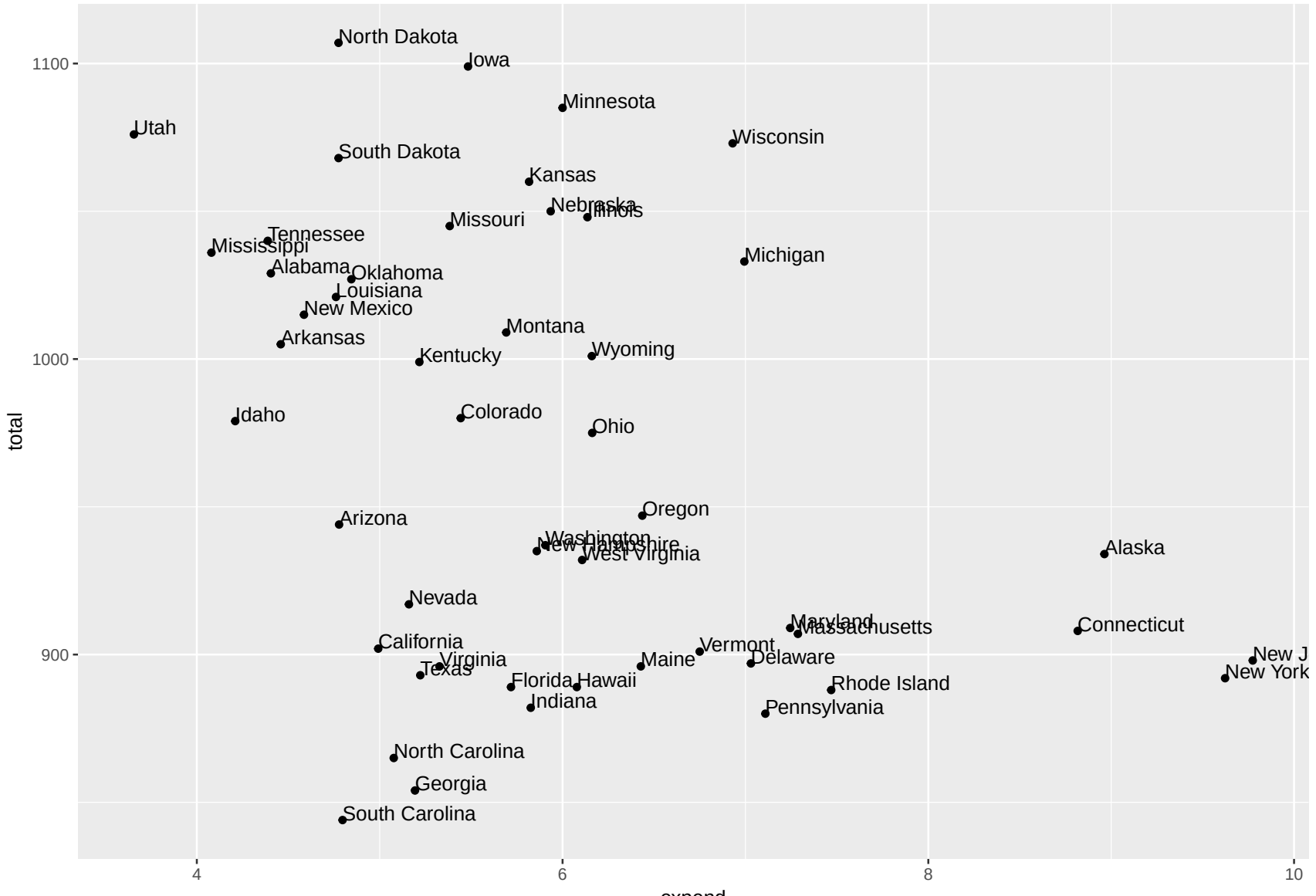
Covariates

Uses for a model

- Making predictions with available data–weather forecasting, google estimating your time of arrival when you use Google Maps.
- Exploring large data sets–the current Big Data fad.
- Analysing the outcome of an intervention in the system–very important for policy

Our well-used example: does spending more on education improve outcomes?

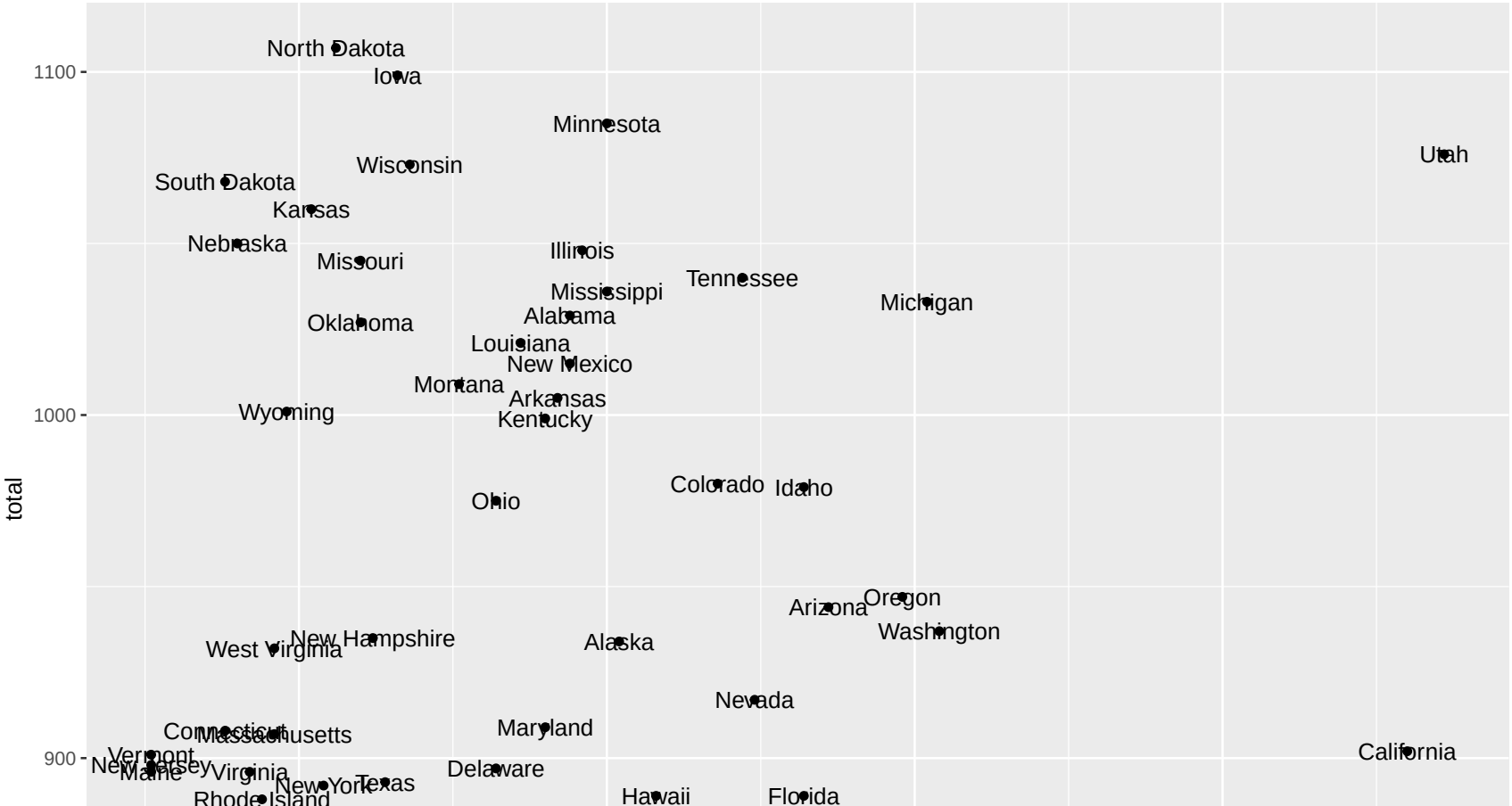
Staring at the SAT Data.



Model

The data are telling us spending more reduces your SAT score. Something is clearly wrong. What?

- Teacher salary?
- Religious ethos?
- fraction of people taking the test?



Covariates

- frac is an explanatory variable. It's not the main thing, but it is something we should possibly care about. * What you'd like to do is hold 'frac' constant and then see how SAT scores fluctuate.
- Loads of important covariates, for example expend $\sim$ state, expend $\sim$ frac $+$ state, and so on.
- By including a covariate, you are effectively determining which two groups to compare.

Let's look at building up a model using covariates to try to understand wher gender pay equity (or inequity) can be understood using this approach.

Example

```
library(mosaicData)
data("Trucking_jobs")
# Train the five models
model_1 <- lm(earnings ~ sex, data = Trucking_jobs)
model_2 <- lm(earnings ~ sex + age, data = Trucking_jobs)
model_3 <- lm(earnings ~ sex + hiredyears, data = Trucking_
model_4 <- lm(earnings ~ sex + title, data = Trucking_jobs)
model_5 <- lm(earnings ~ sex + age + hiredyears + title, da

# Evaluate each model...
evaluate_model(model_1)
evaluate_model(model_2, age = 30)
evaluate_model(model_3, hiredyears = 10)
evaluate_model(model_4, title = 'PROGRAMMER')
evaluate_model(model_5, age = 30, hiredyears = 10,
               title = 'PROGRAMMER')

# ...and calculate the gender difference in earnings using
```

Effect size

- how does changing an input model change an output value?
- here the inputs *cause* the output
- Doesn't mean the real world works that way.
- Natural units for effect sizes
- Quantiative variables: effect size is a rate.
- Eg change in wage by hour per year.
- Categorical variable: effect size is a difference.

Effect size is a property of the model, not the data

- You must include the variable you want to measure.
- Then you can compare the output of different values of the explanatory variable.

Example

```
library(statisticalModeling) #load the package with data i
#Train a model to look at life insurance
model<-lm(Cost ~ Age + Sex + Coverage, data=AARP)
# Calculate the effect size for each input, which is calcu
effect_size(model, ~Age) # This command  takes two argumen
```

```
##       slope  Age   to:Age Sex Coverage
## 1 3.351705 59.5 68.16025   F       20
```

```
effect_size(model, ~Sex)
```

```
##     change Sex to:Sex  Age Coverage
## 1 10.23278   F      M 59.5       20
```

```
effect_size(model, ~Coverage) # EG A change in coverage fr
```

```
##       slope Coverage to:Coverage  Age Sex
## 1 1.815365       20    37.23783 59.5   F
```

Summary

- A big (no pun intended) and emerging field
- Application of machine-learning willy nilly to public policy problems probably won't be successful unless we change the methods somewhat.

Next time: interview-based economic analysis.