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.
- Slides will be added to this master version as I get the chance to build them out.
- 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

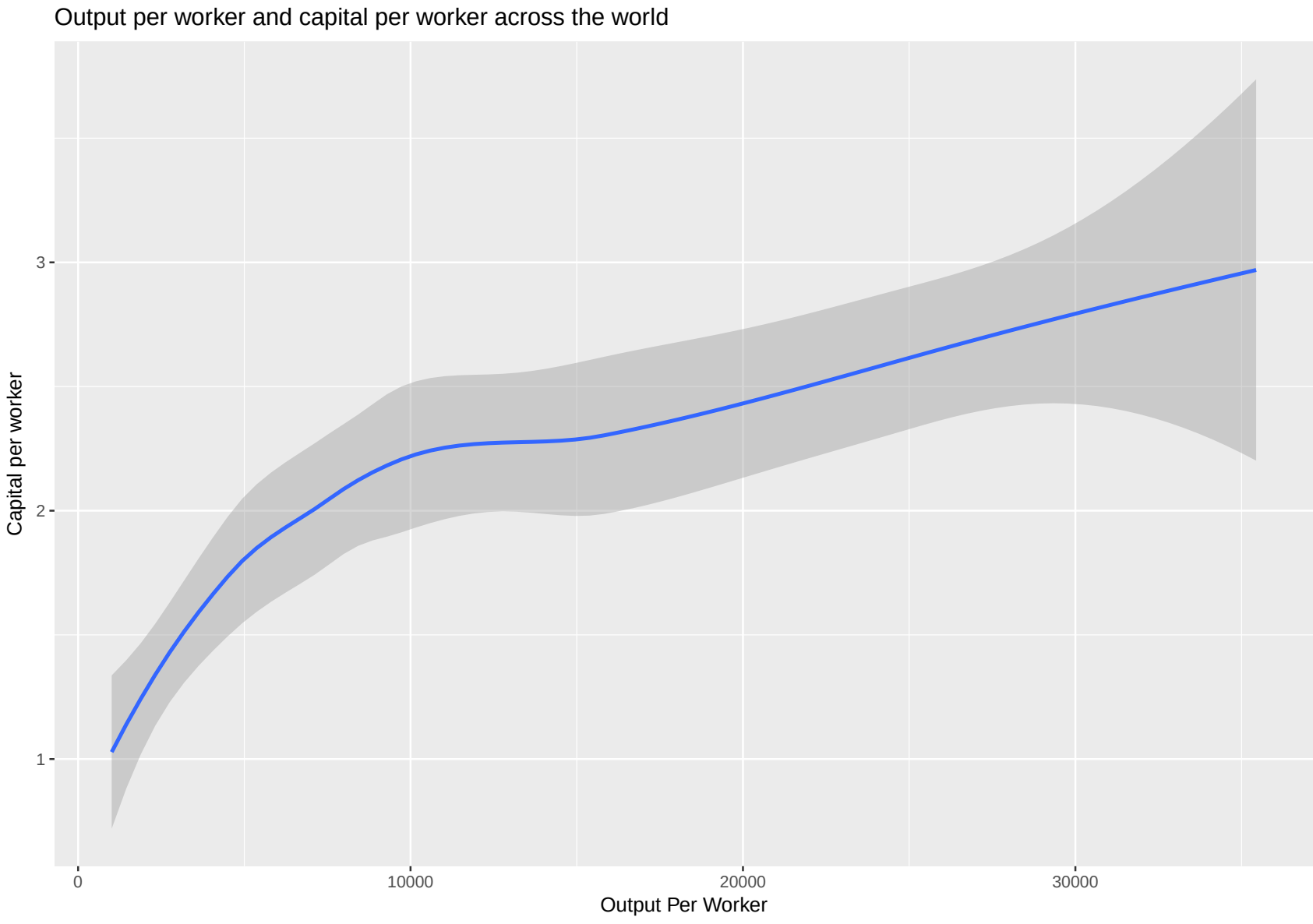
Key Software resources

- 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

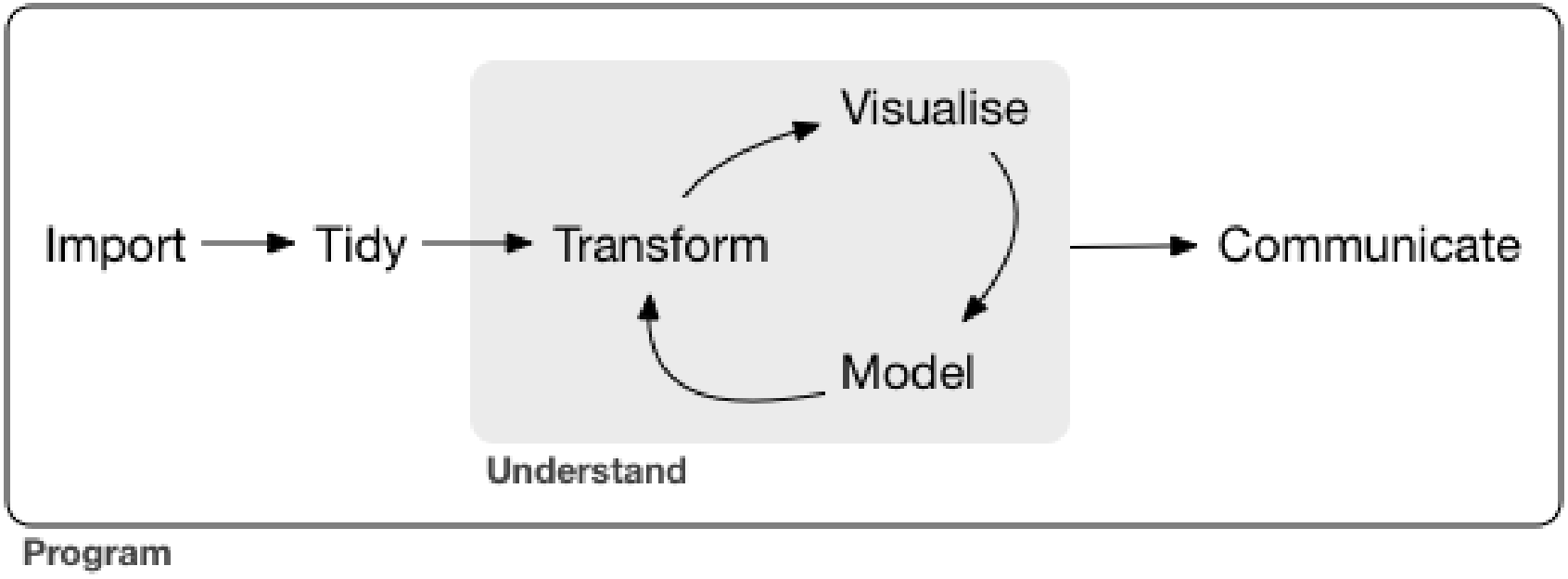


Figure 1: Source: Wickham, 2016

Lecture 1: Movitation, statistical basics and data handling

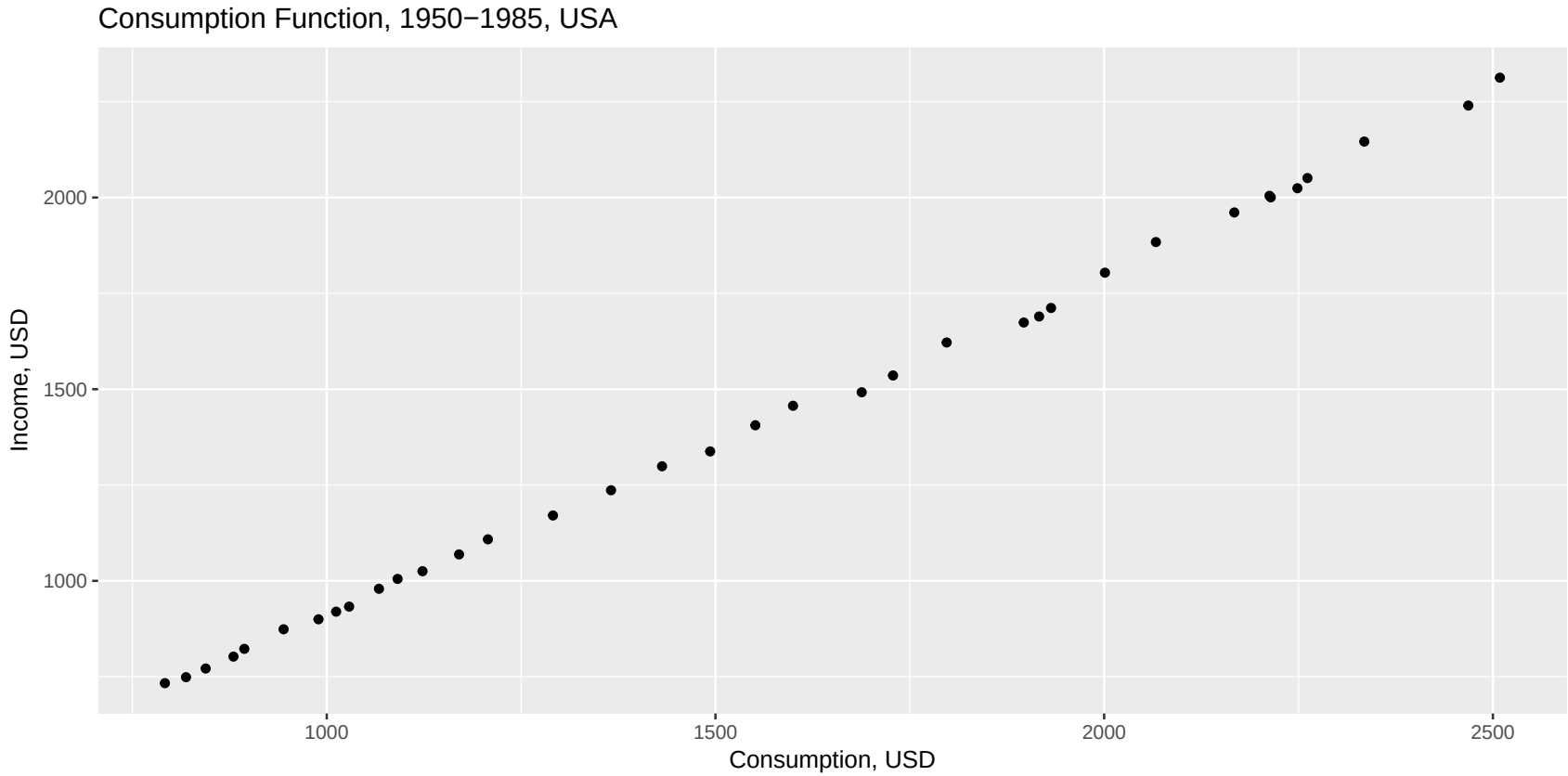
- Types of economic data: micro and macro
- Observation studies and experiments
- Statisitical inference, probabilit distributions, fitting a model.
- Graphical methods
- Descriptive statistics
- Expected Values and Variances
- Example: Hall and Jones, Growth Accounting, 1999.
- Reading: Koop, Chapter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

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

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

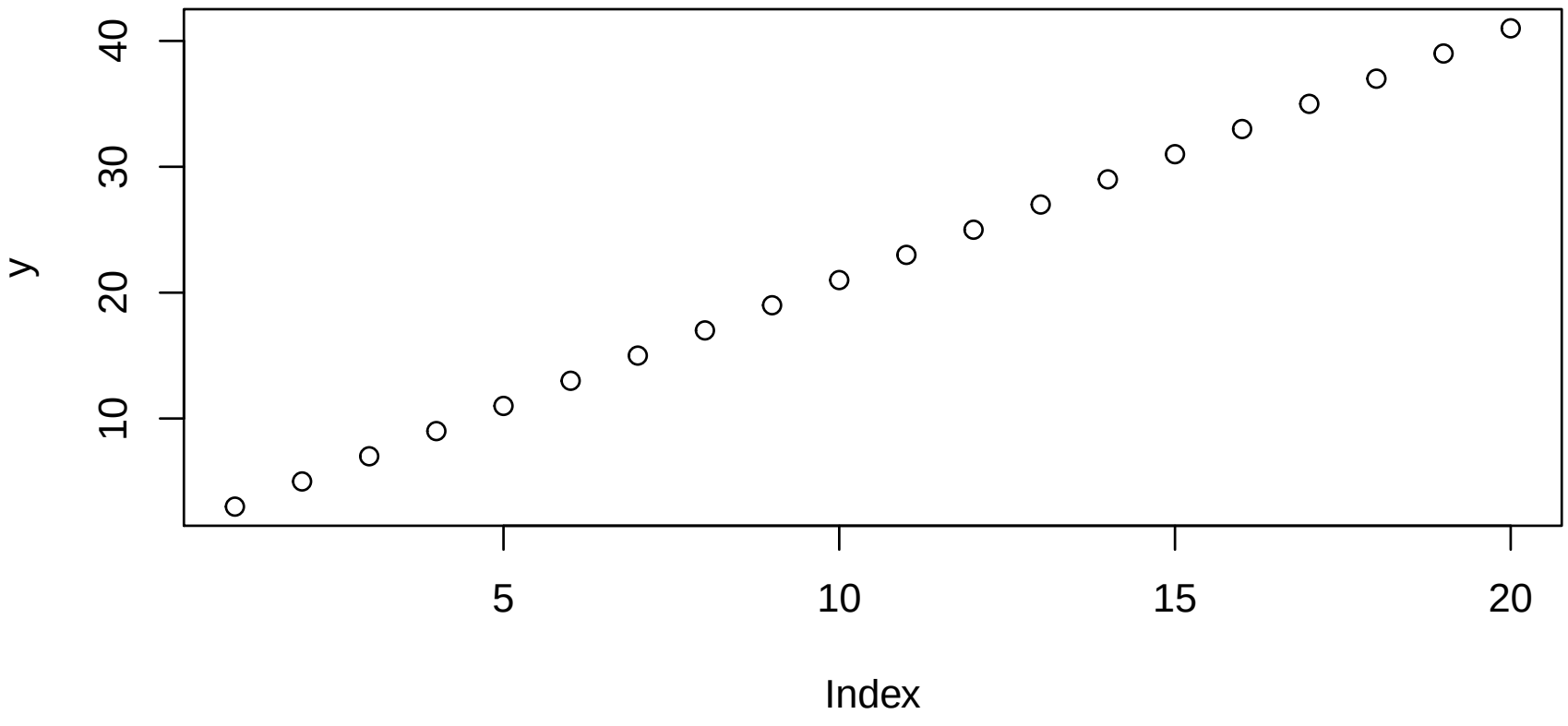
where α and β are coefficients, which determine a particular line. So, for instance, setting $\alpha = 1$ and $\beta = 2$ defines one particular line while $\alpha = 4$ and $\beta =$ -5 defines a different line.

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

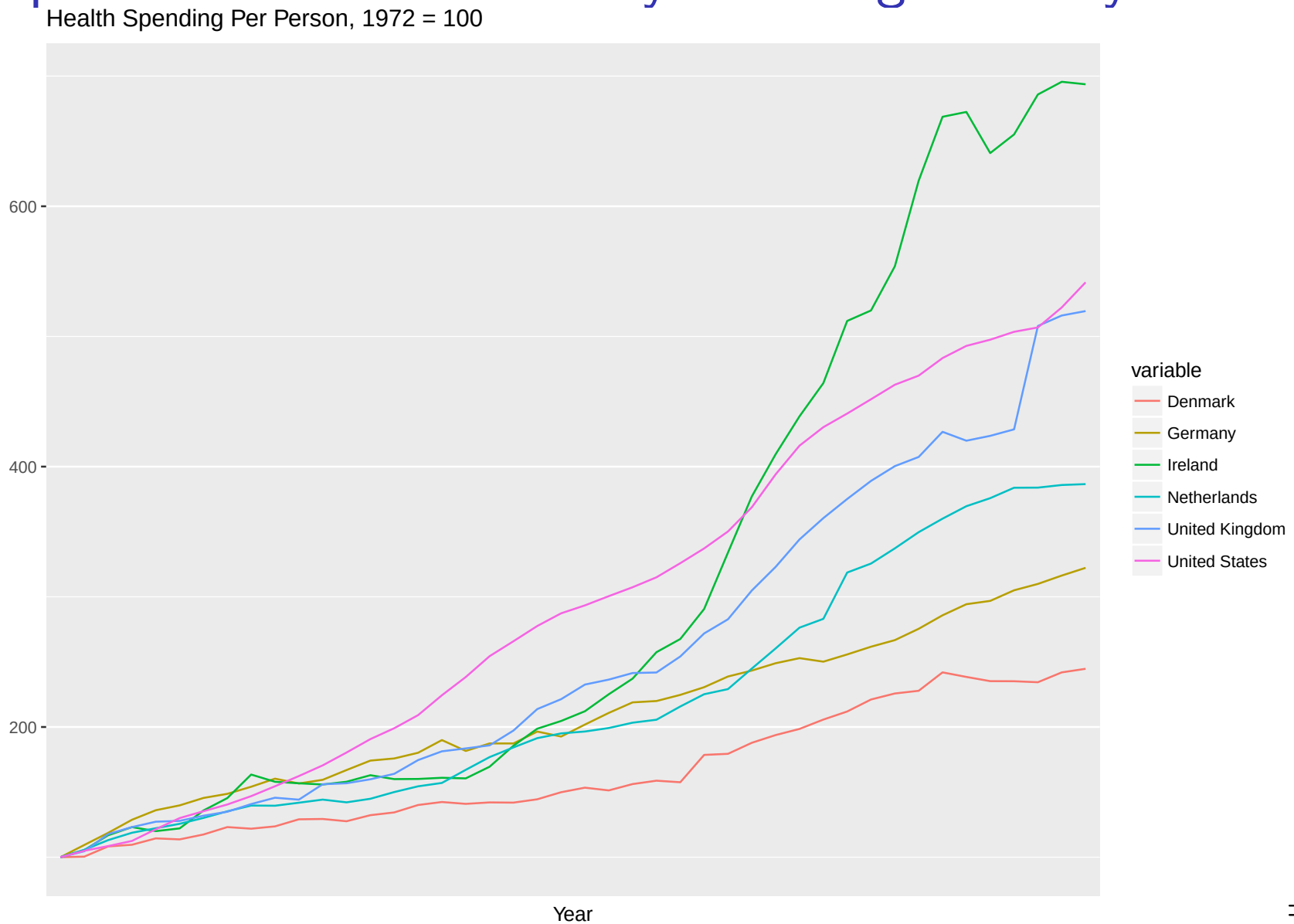
Logarithms

- 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.

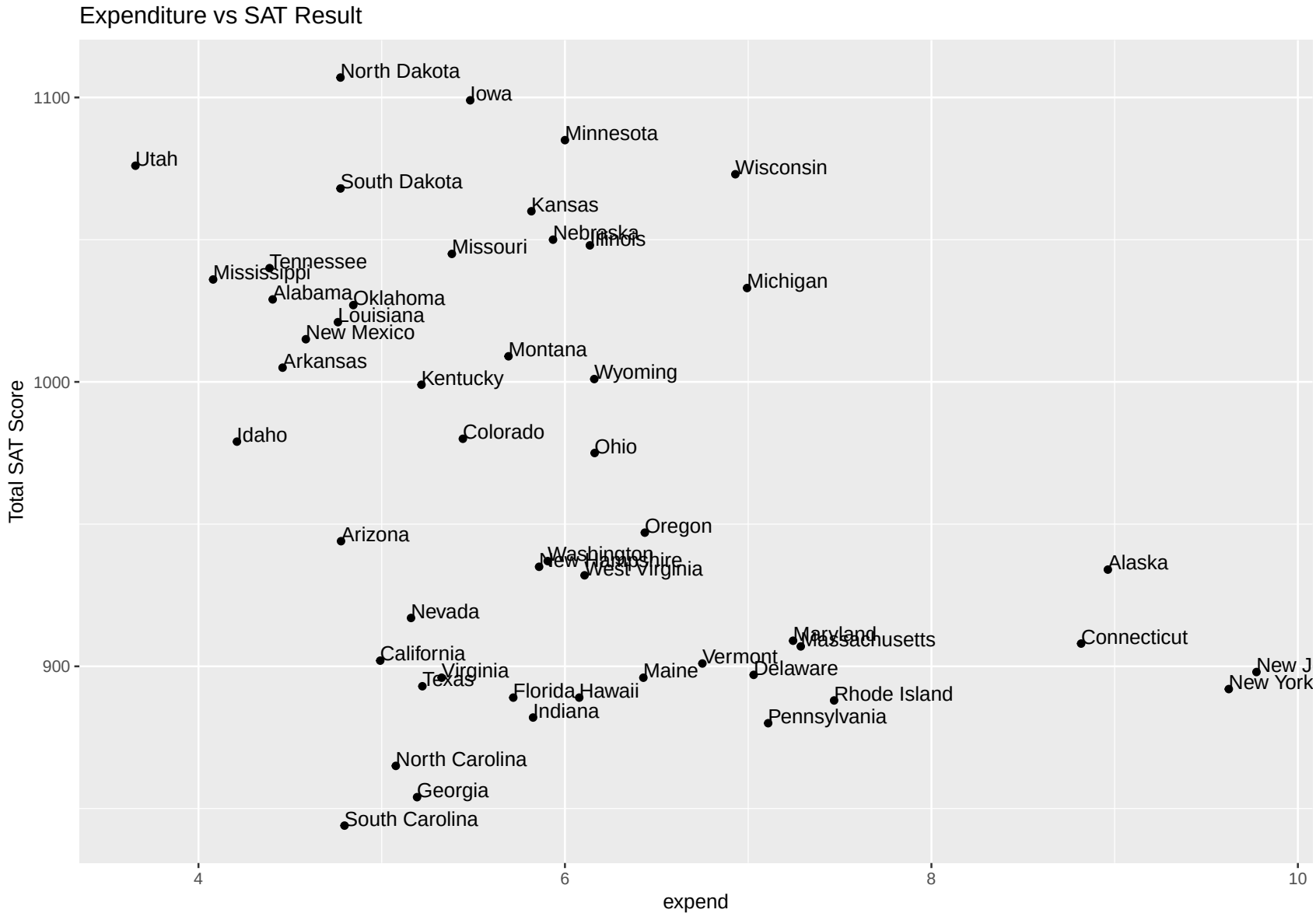
Example: does spending more on education improve outcomes? US Data

```
x<-summary(SAT)
pander(x)
```

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

Looking at the data



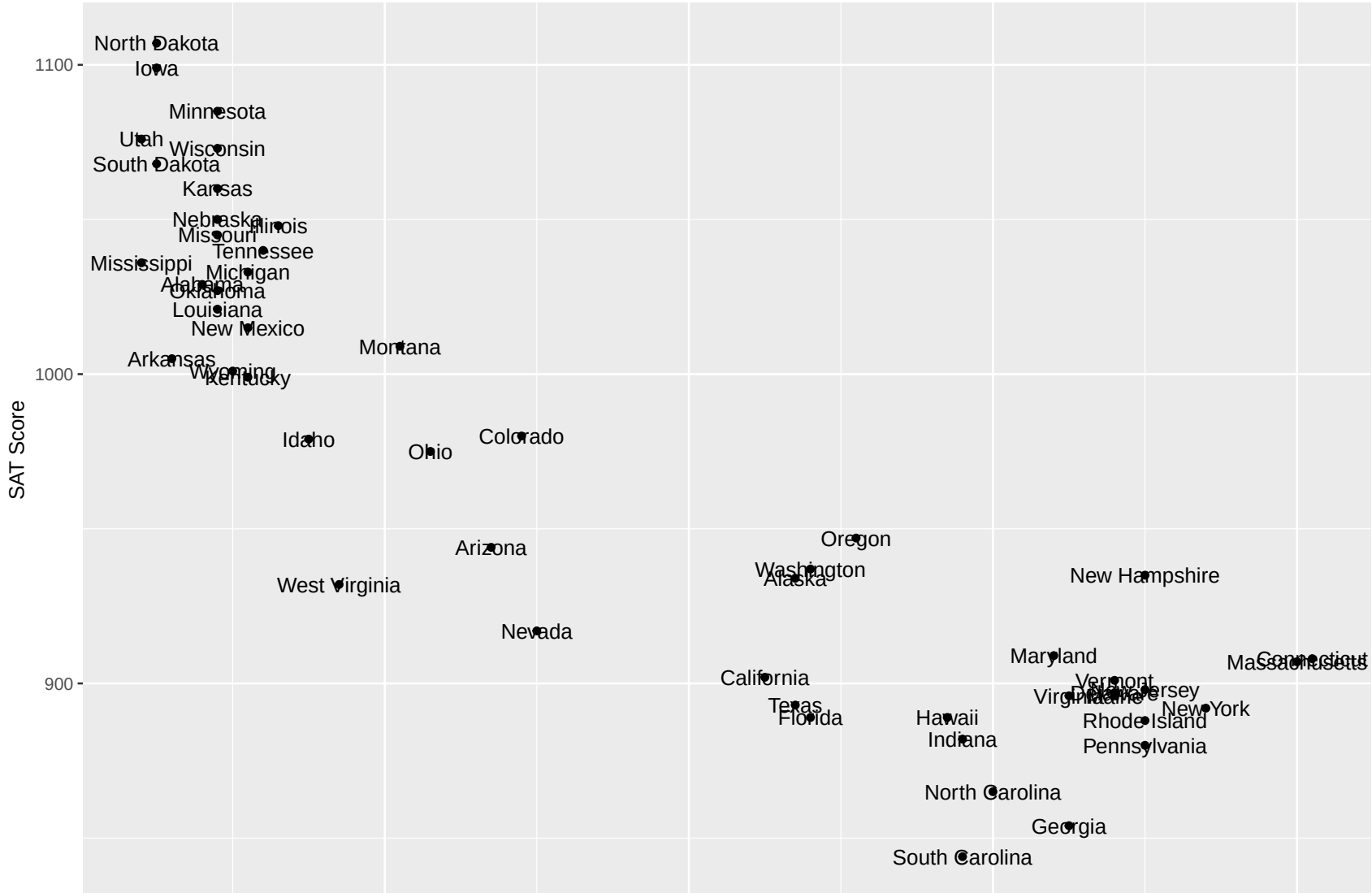
Careful interpreting this dataset!

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

```
ggplot(data=SAT)+ aes(x = perc, y = total, label = state) +
```



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.

```
##install.packages("tidyquant")
library(tidyquant)
AMZN <- tq_get("AMZN", get = "stock.prices", from = "2007-0
FANG <- c("FB", "AMZN", "NFLX", "GOOG") %>%
    tq_get(get = "stock.prices", from = "2007-01-01", to =
## Setup dates for zoom window
end <- ymd("2017-01-01")
```

Example from Freedman

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.

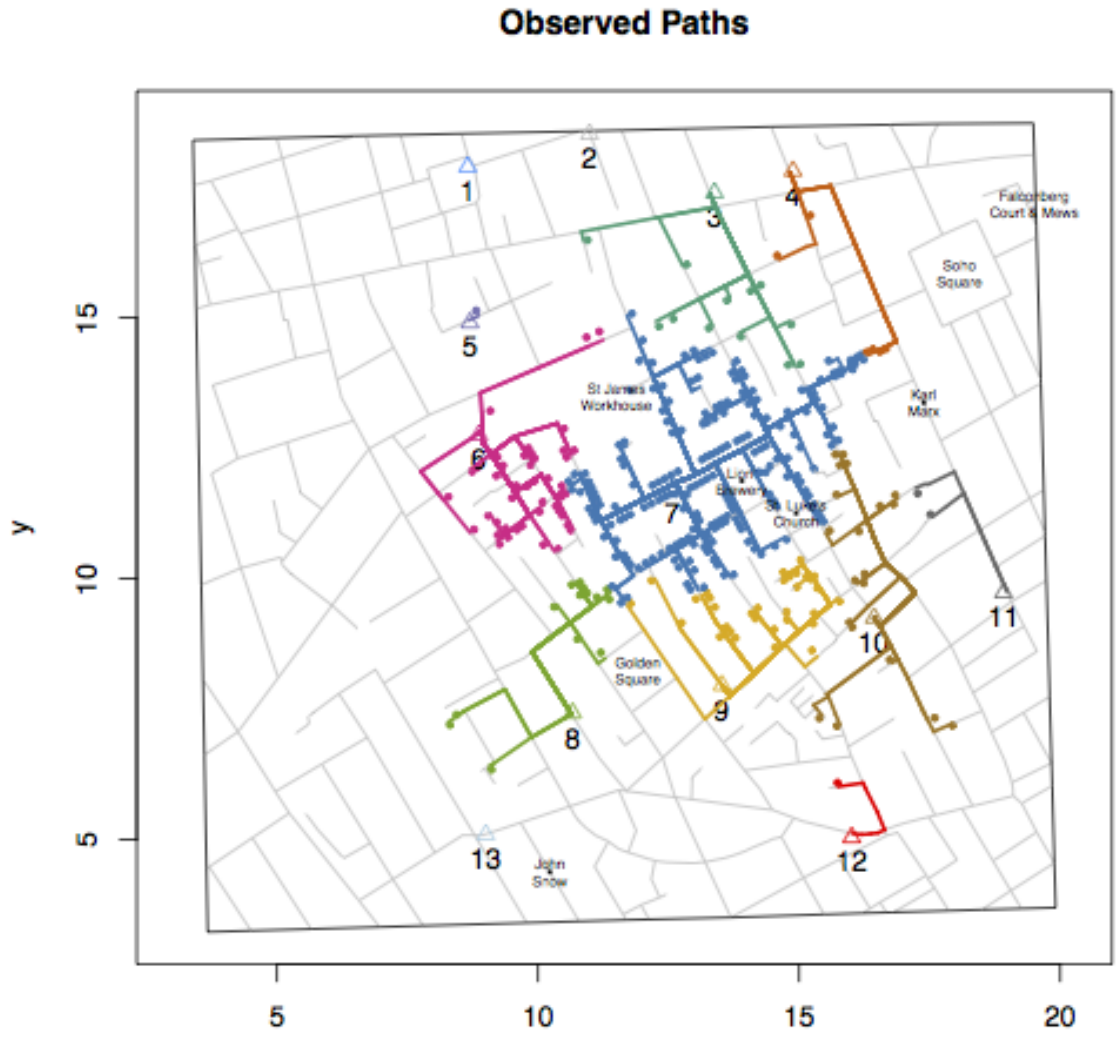
An even earlier example: John Snow and Cholera

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.

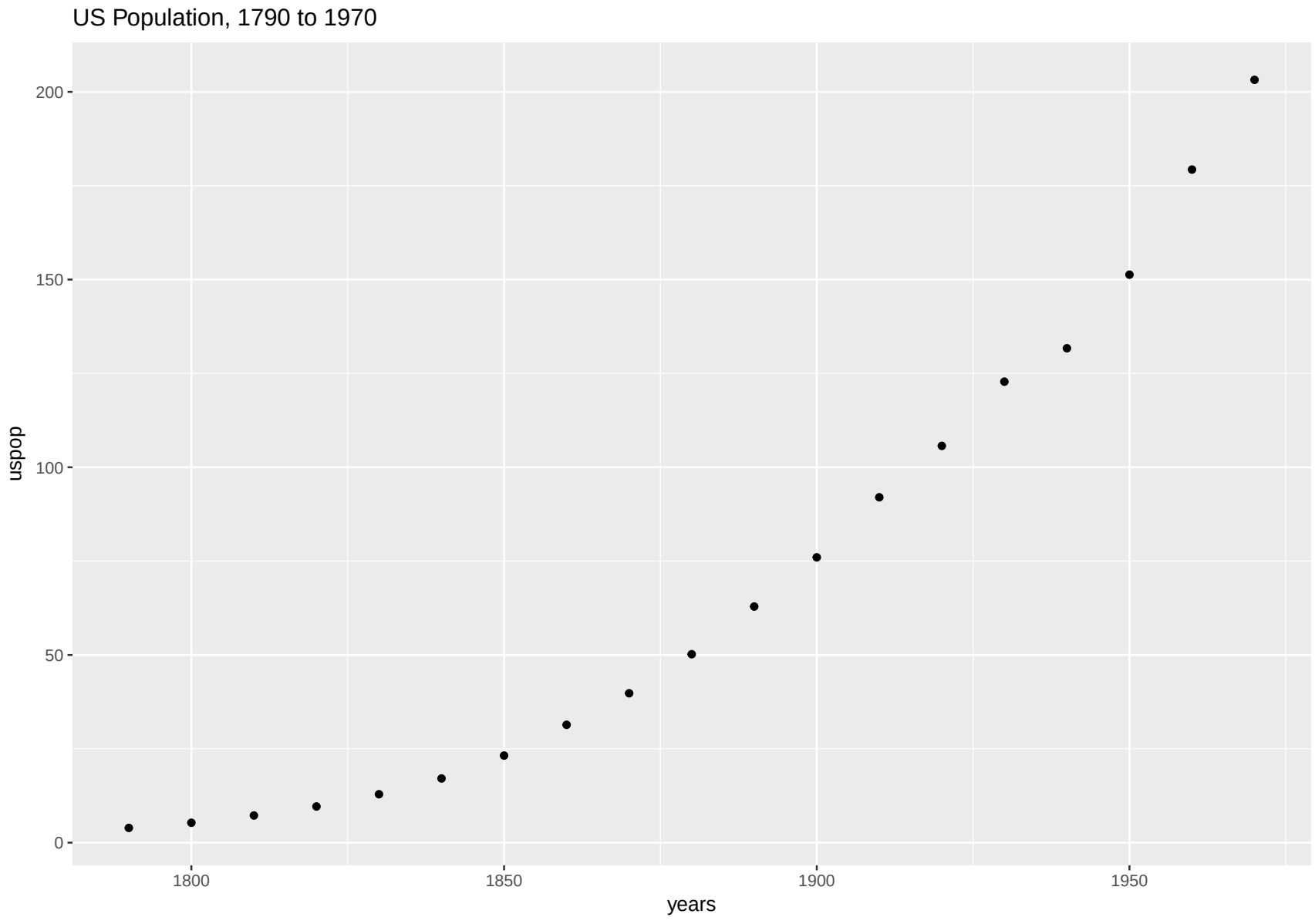
By doing this he found there was a significant clustering of the deaths around a certain pump – and removing the handle of the pump stopped the outbreak and invented epidemiology.

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

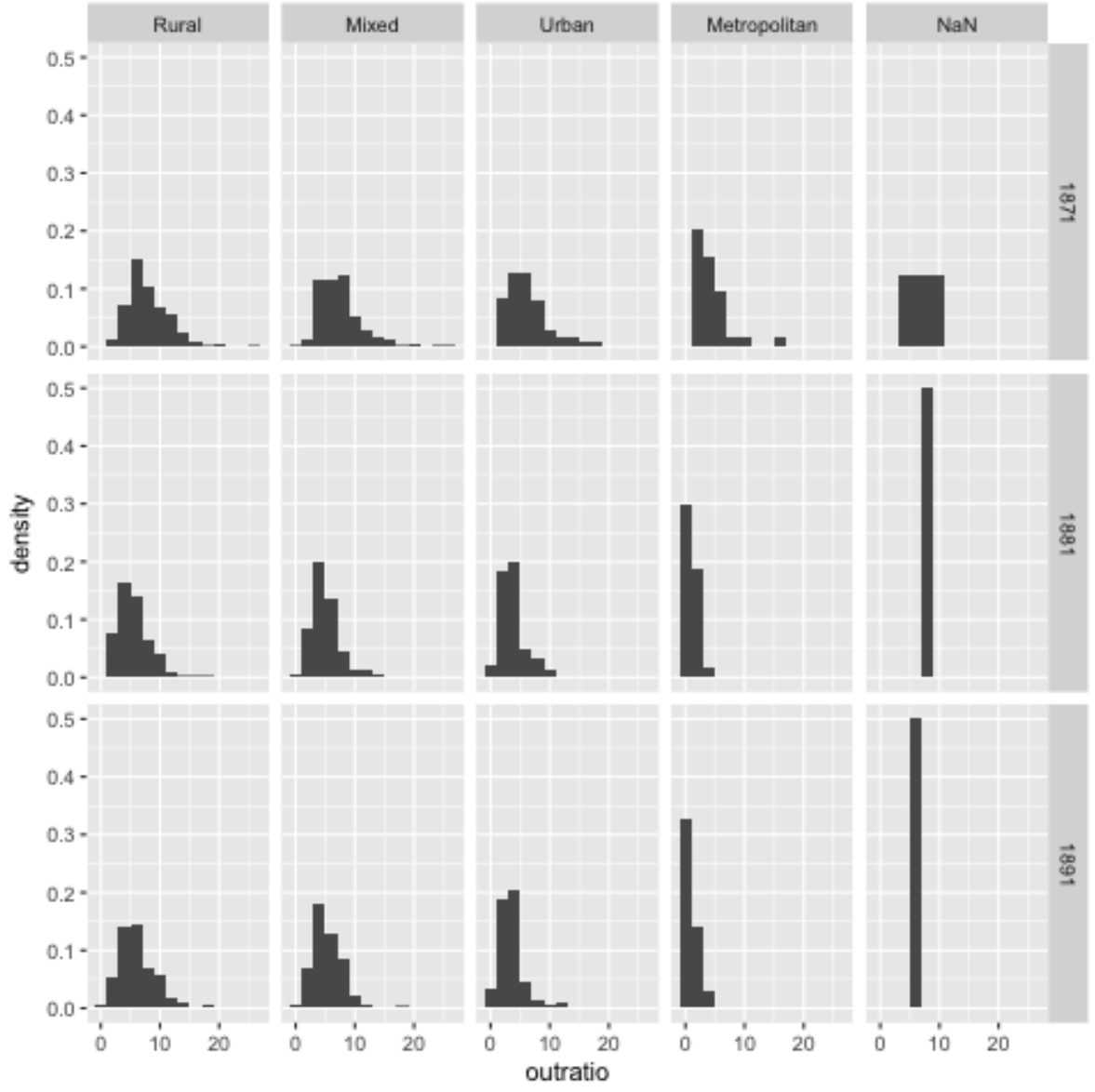


Figure 3: Created using `https://github.com/jrnold/yule`

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

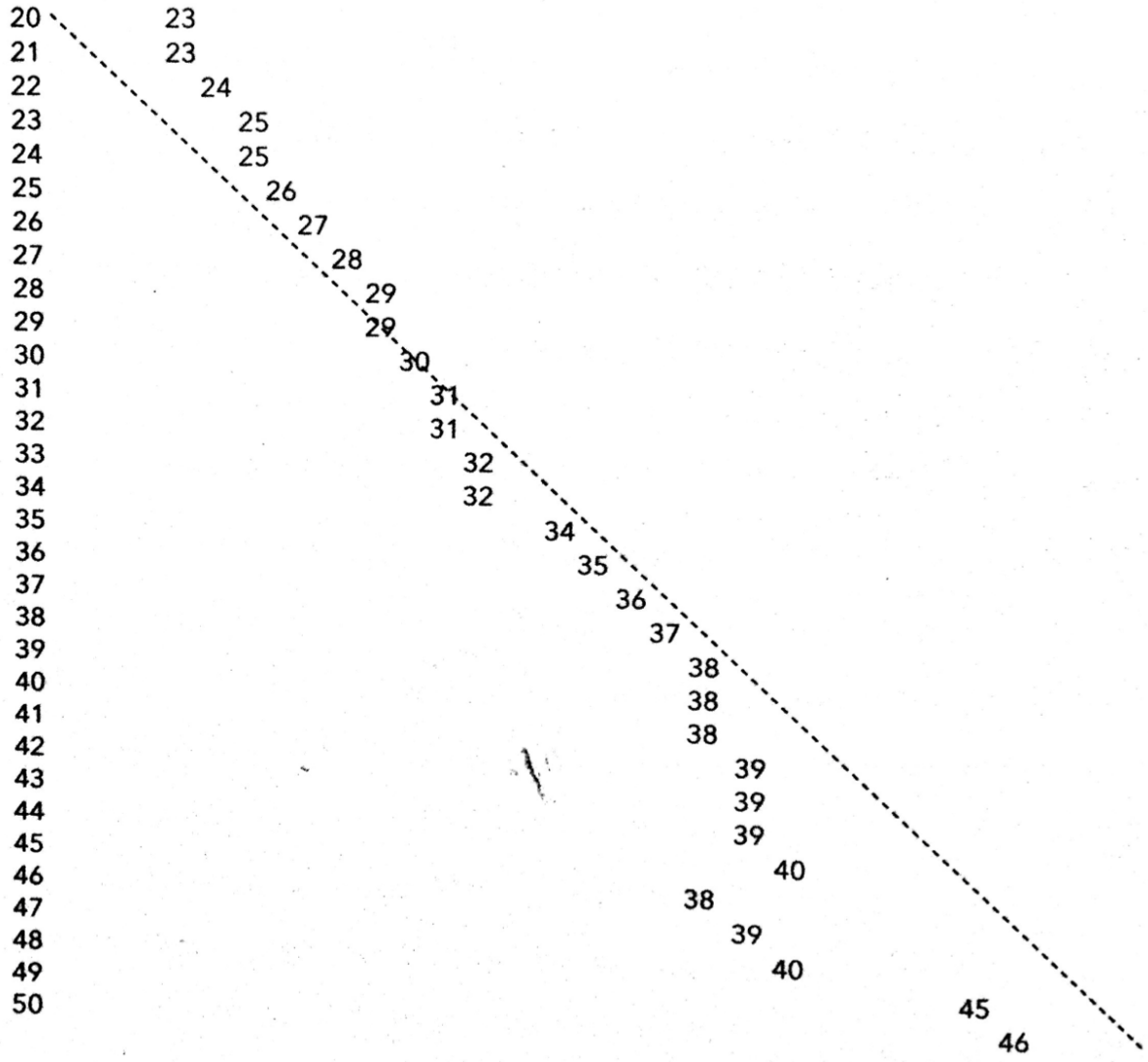
Lecture 2: Modeling (Freedman Chapter 1, Koop Chapter 4)

- 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

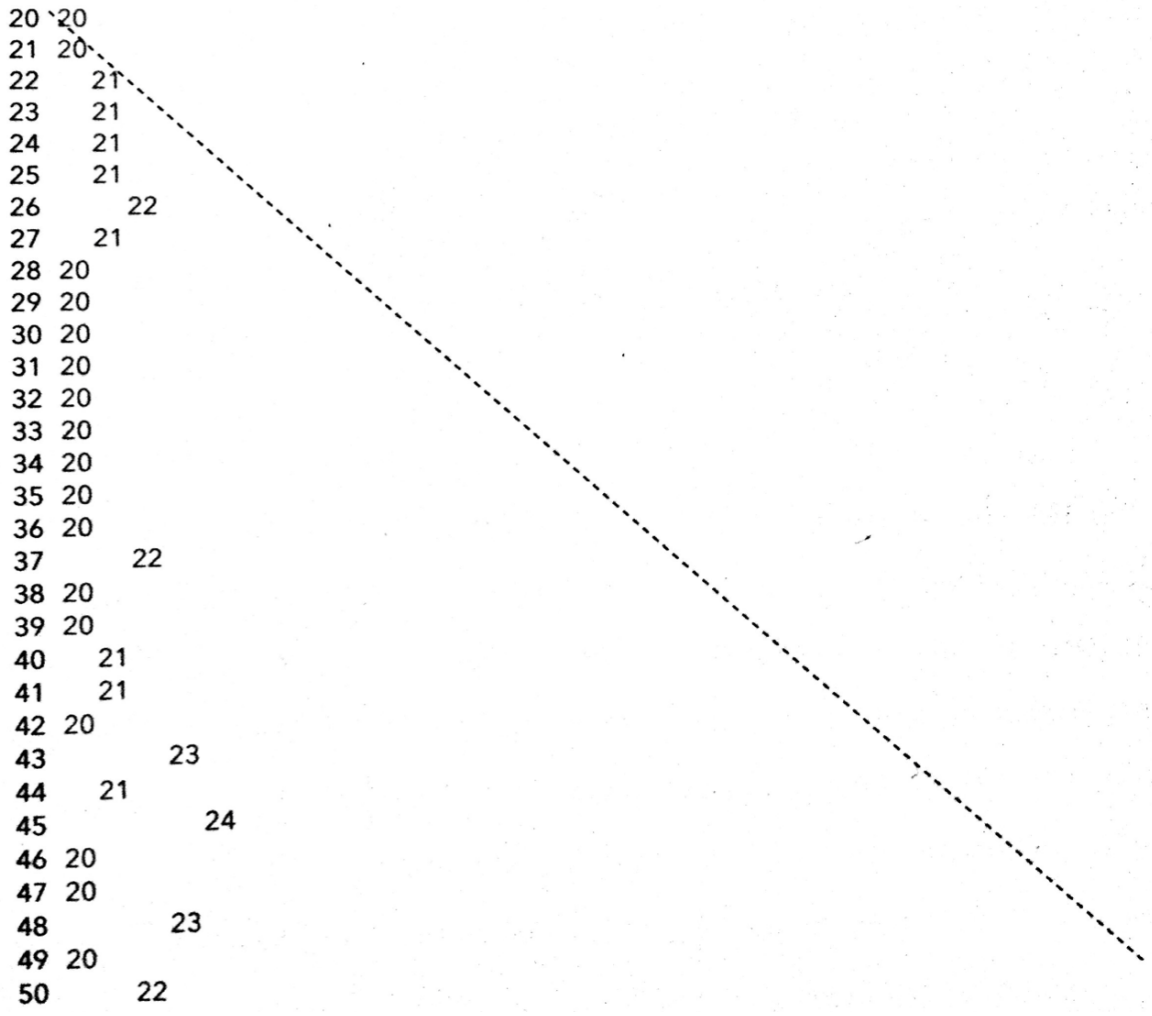
Dataclysm: a woman's age vs the age of the men who look best to her



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).

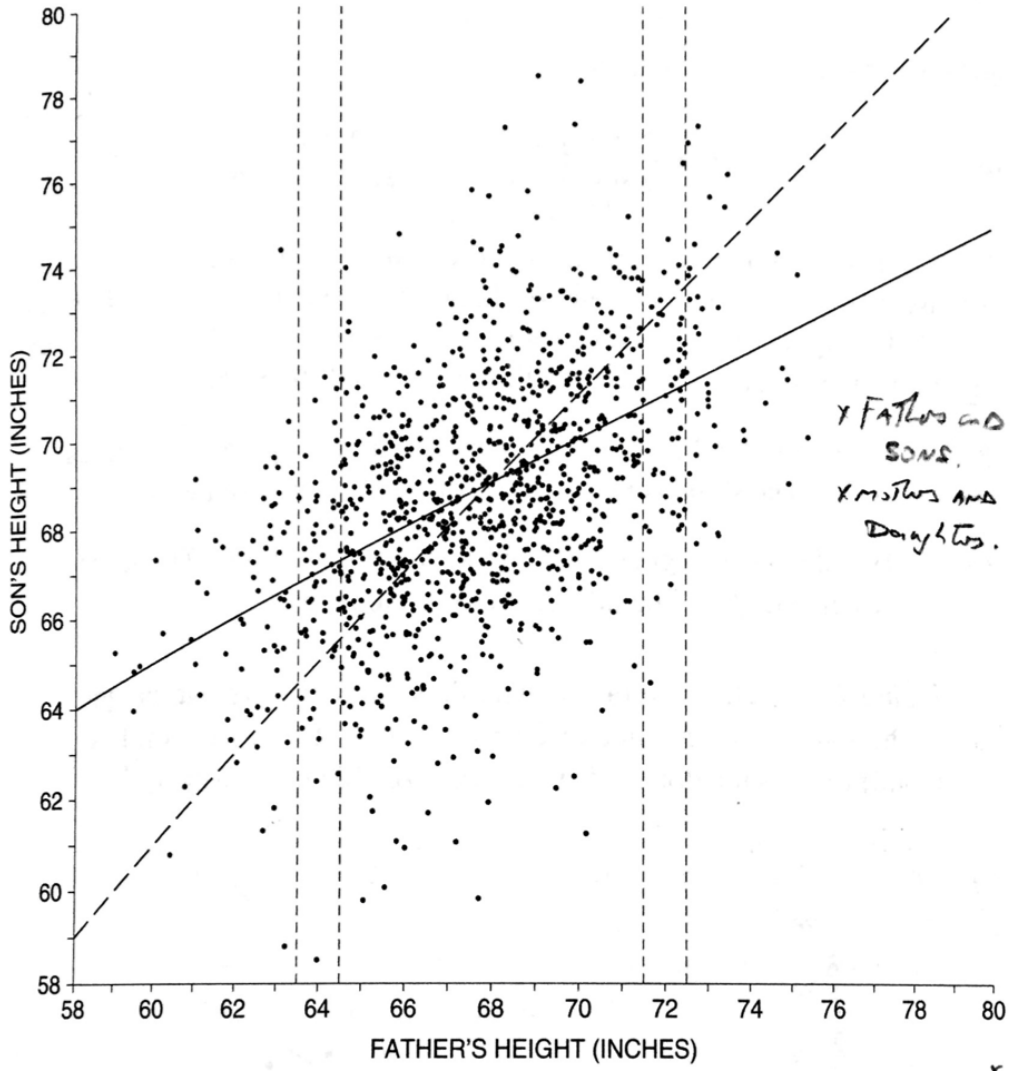


Figure 4: Source: Freedman, 2012

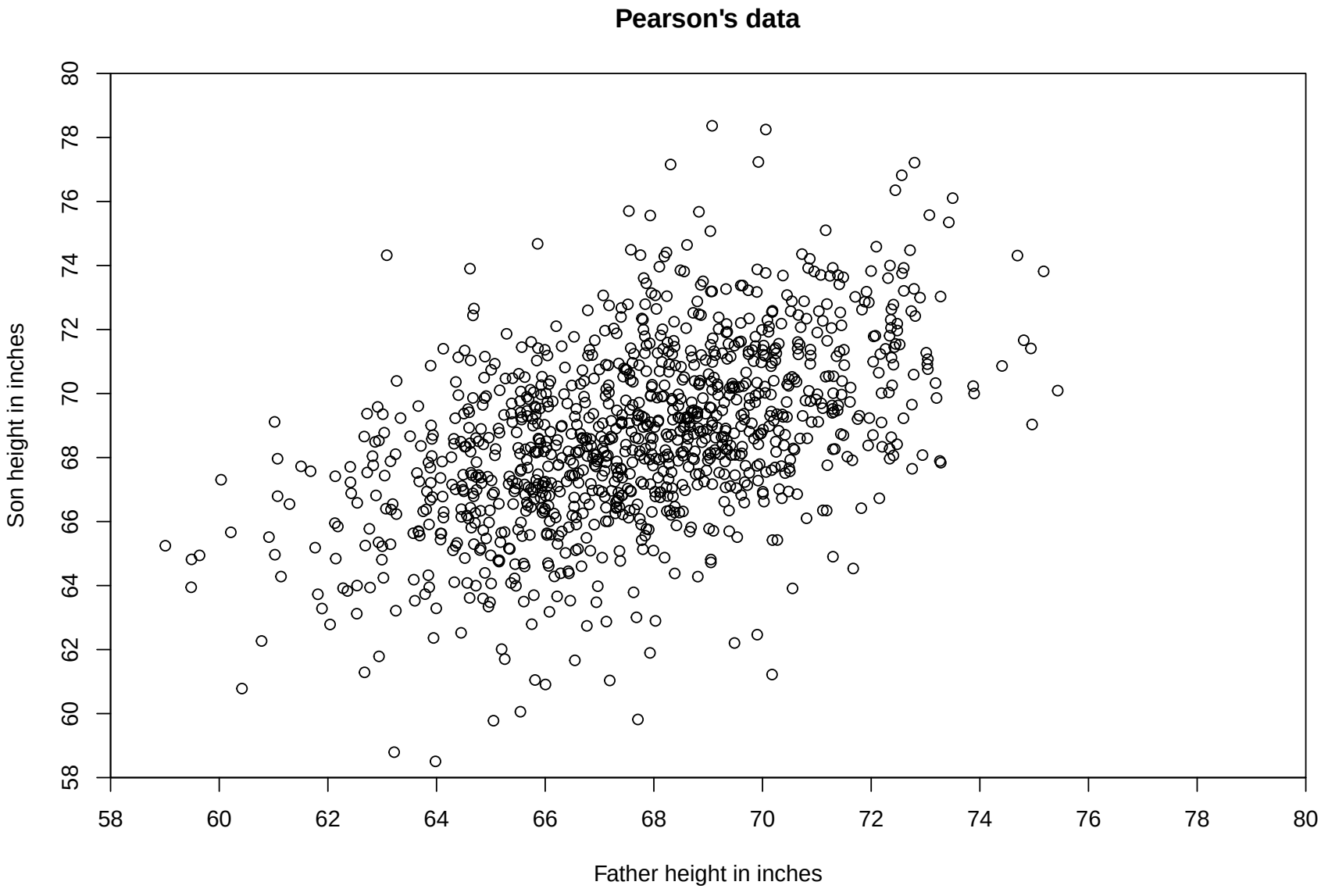
Understanding correlation

- 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
```

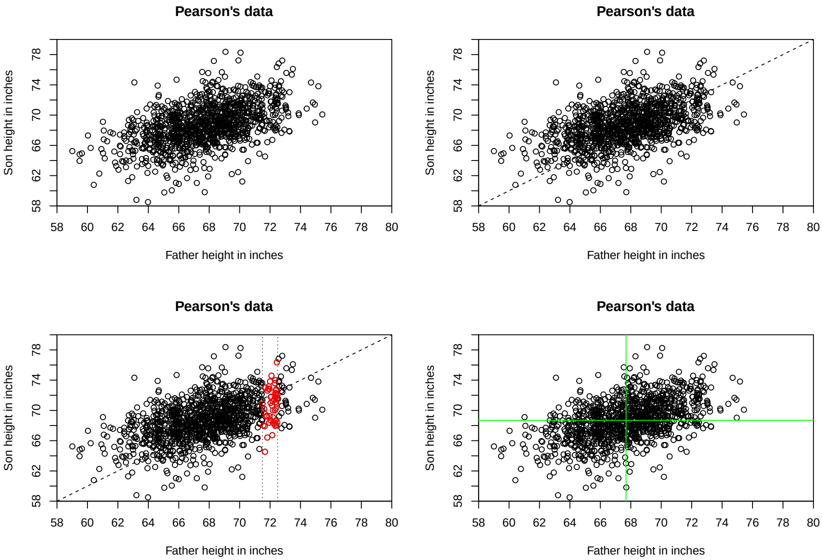
And this


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.

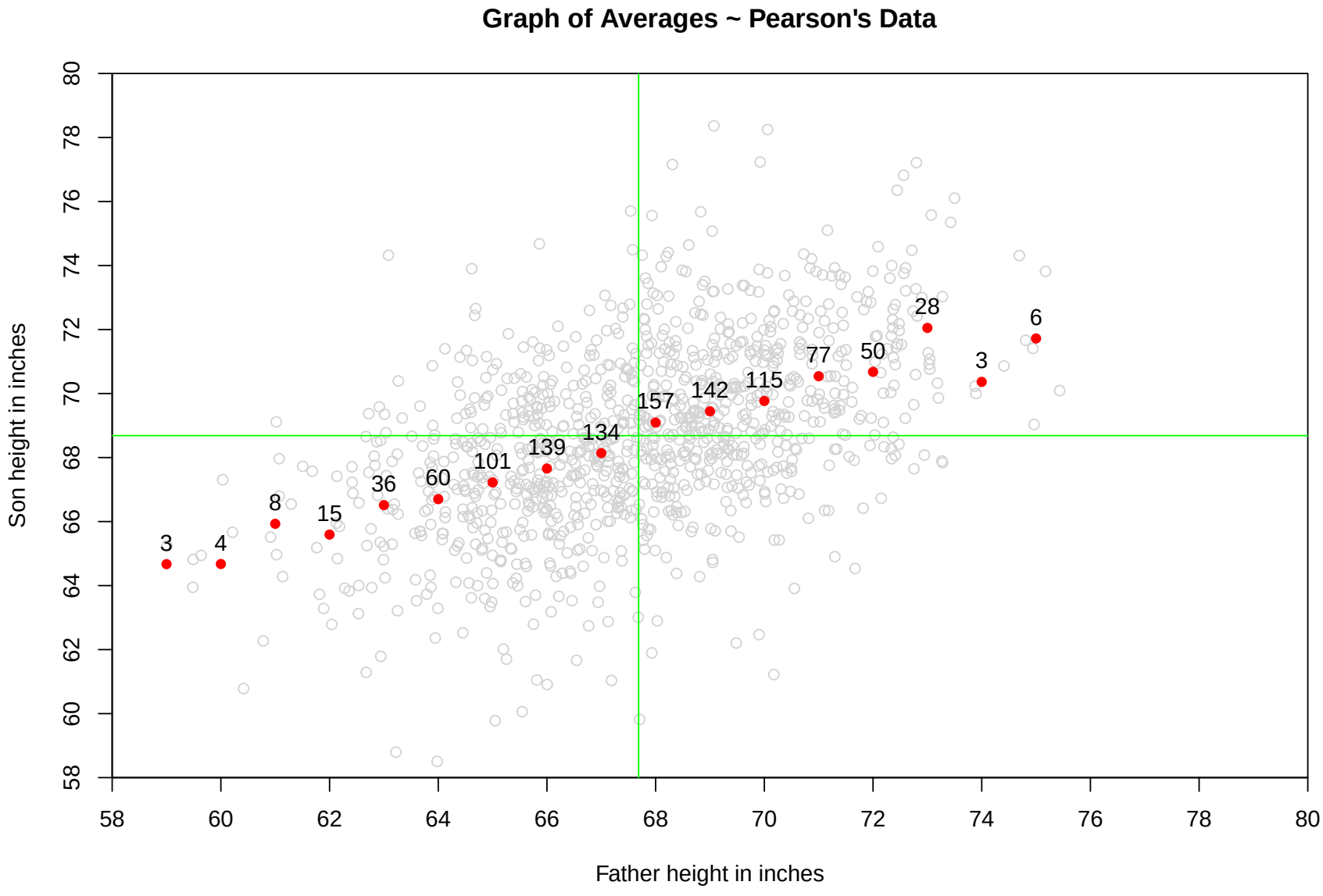
Interpreting these figures 2

- 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?

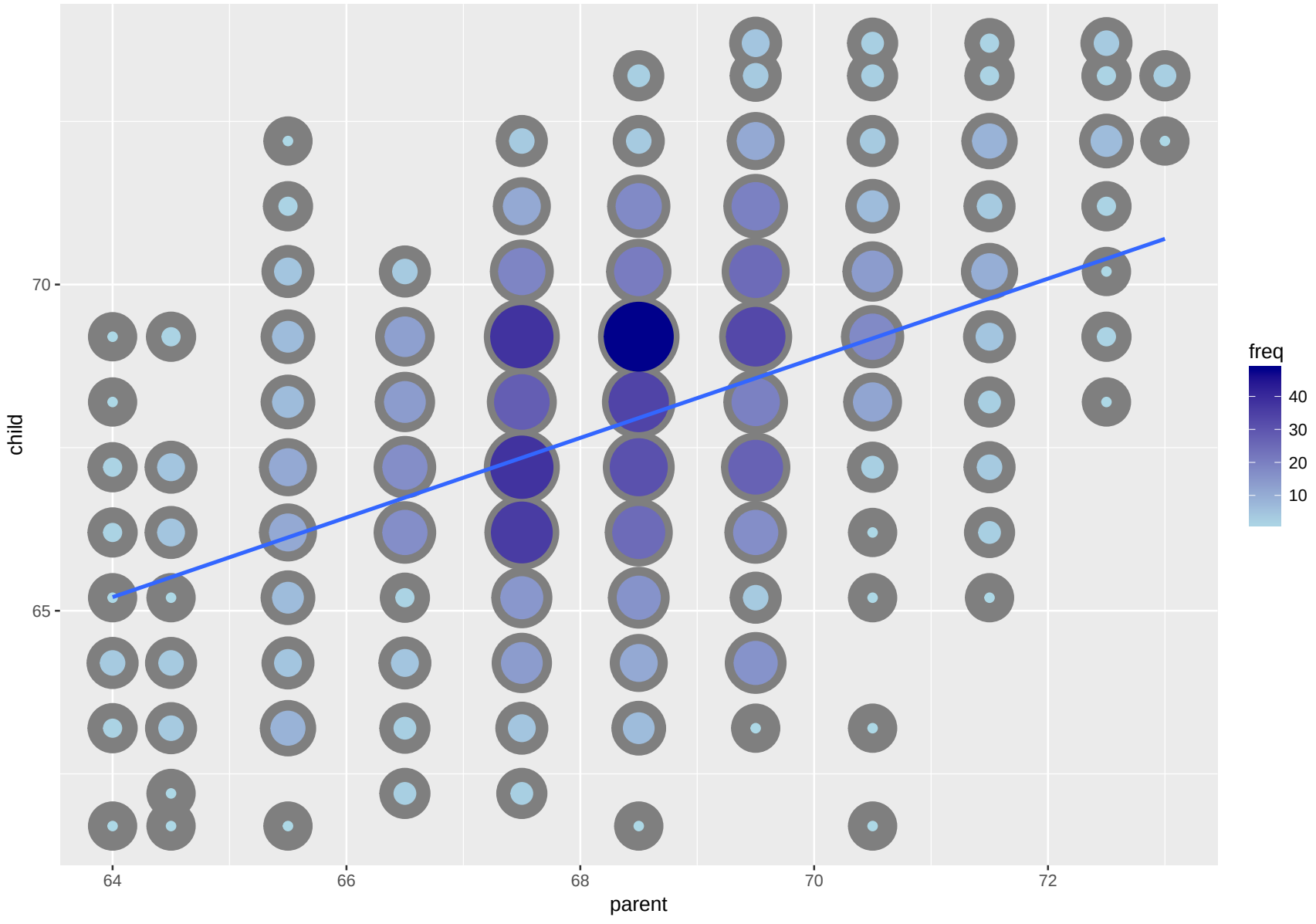
The regression line

- 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



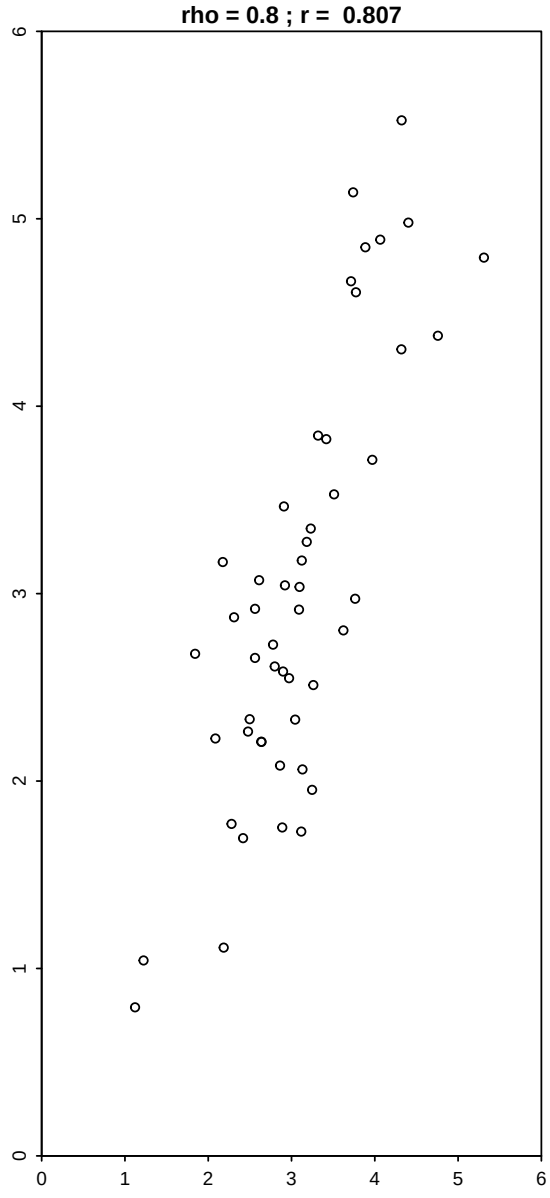
A subtle distinction

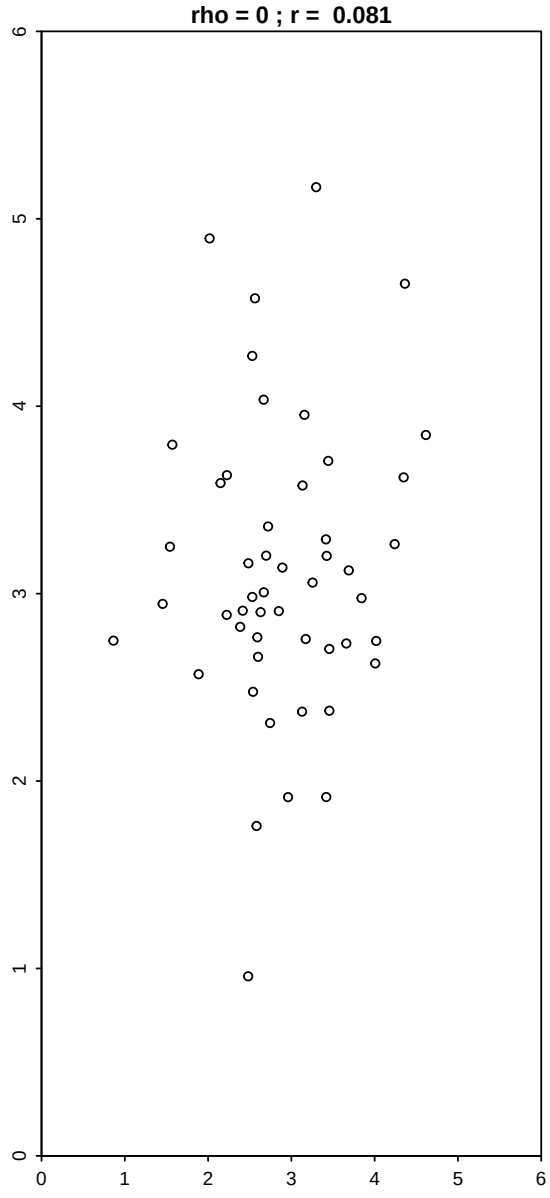
- Recall the distinction between association and causation
- Before graphing models like this was a few clicks, it made a lot of sense to summarise the count data using summary statistics.

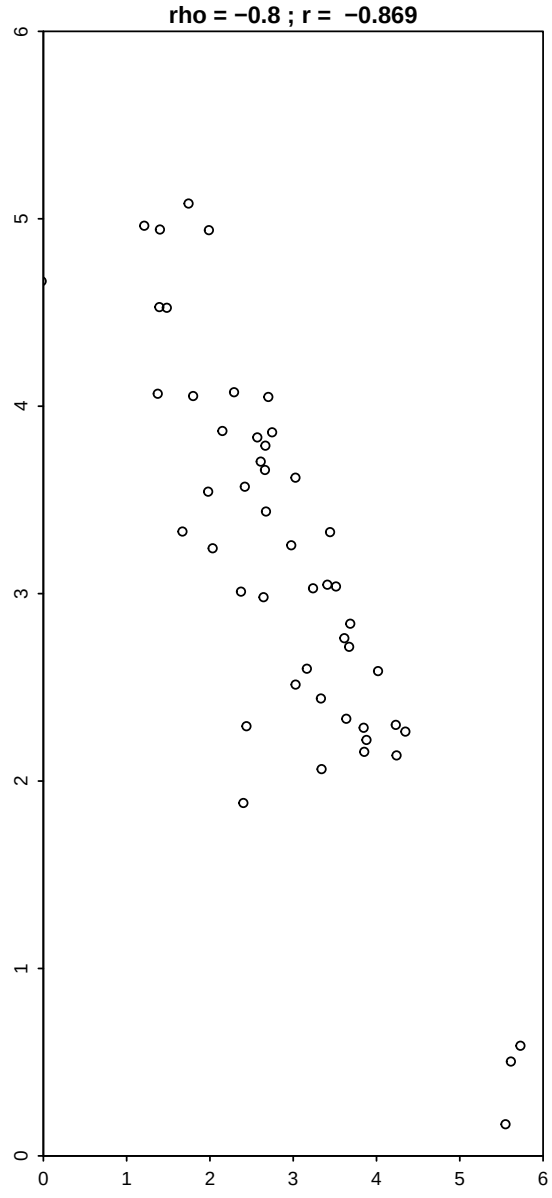
The Pearson correlation concept.

- 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 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

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

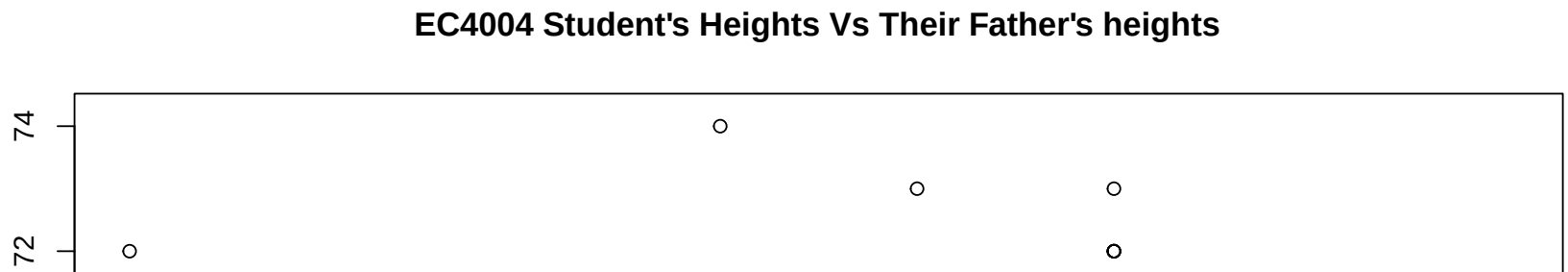
This time

- 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

EC4004 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.

```
library(dplyr)
df<-read_csv("classpearson.csv") # Import the data
height<-sapply(strsplit(as.character(df$Height),"'|\""),
        function(x){12*as.numeric(x[1]) + as.numeric(x[2])}
fheight<-sapply(strsplit(as.character(df$Fheight),"'|\""),
        function(x){12*as.numeric(x[1]) + as.numeric(x[2])}
mheight<-sapply(strsplit(as.character(df$Mheight),"'|\""),
        function(x){12*as.numeric(x[1]) + as.numeric(x[2])}
df<-data.frame(df$Gender, height, fheight, mheight)
plot(fheight, height, main="EC4004 Student's Heights Vs The
```



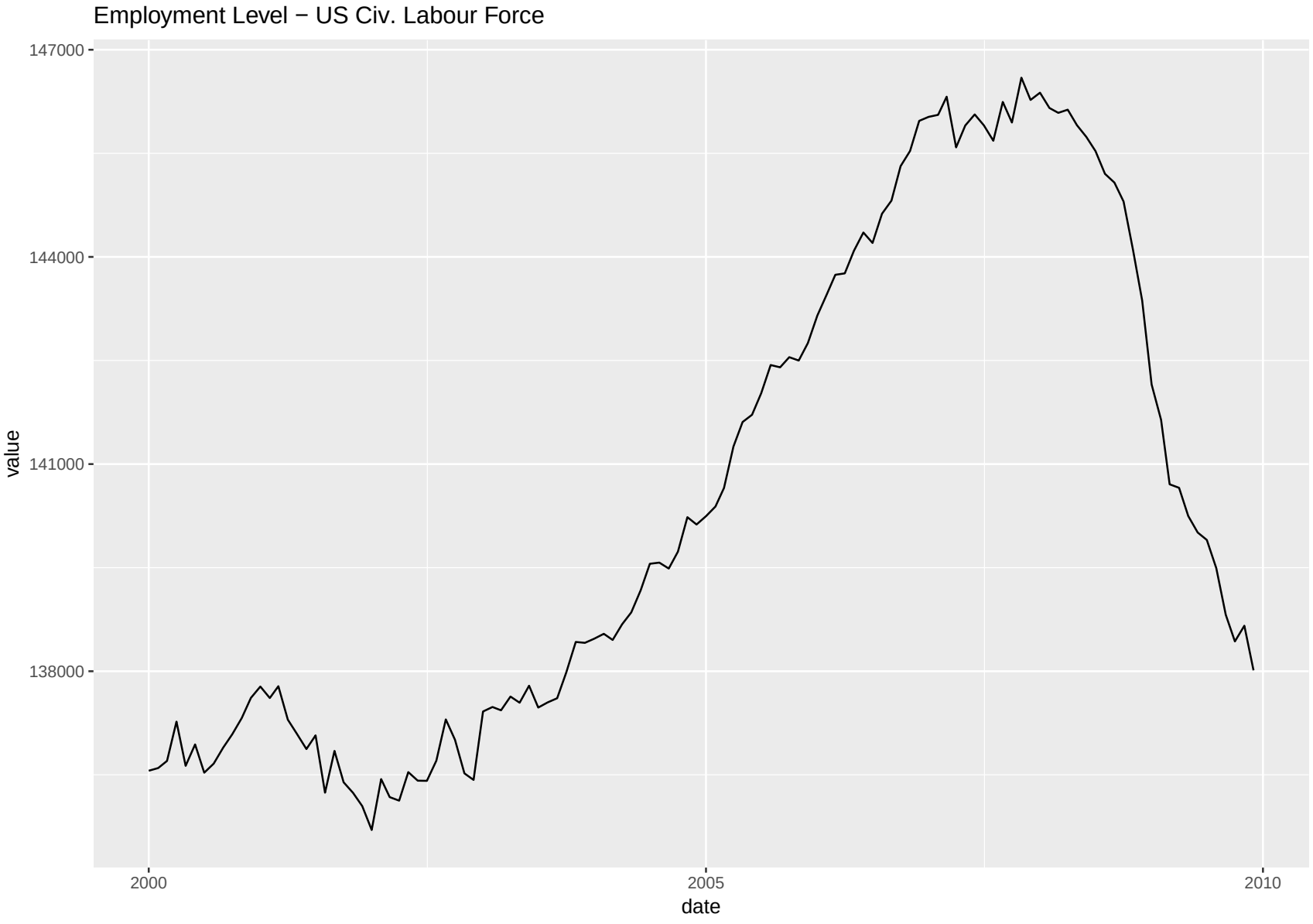
Concept of regression

- 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.

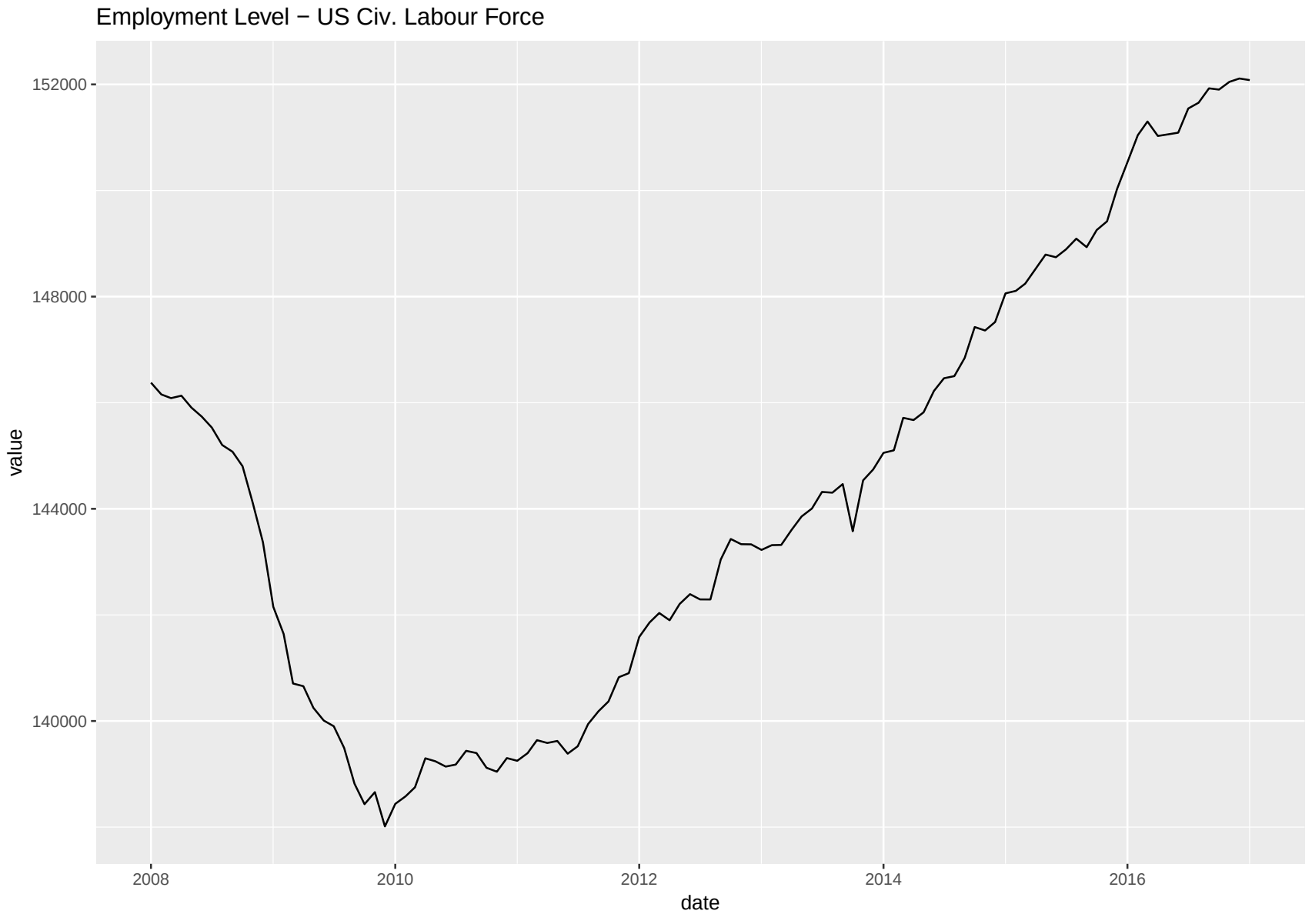
The Econometric Approach

- 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?

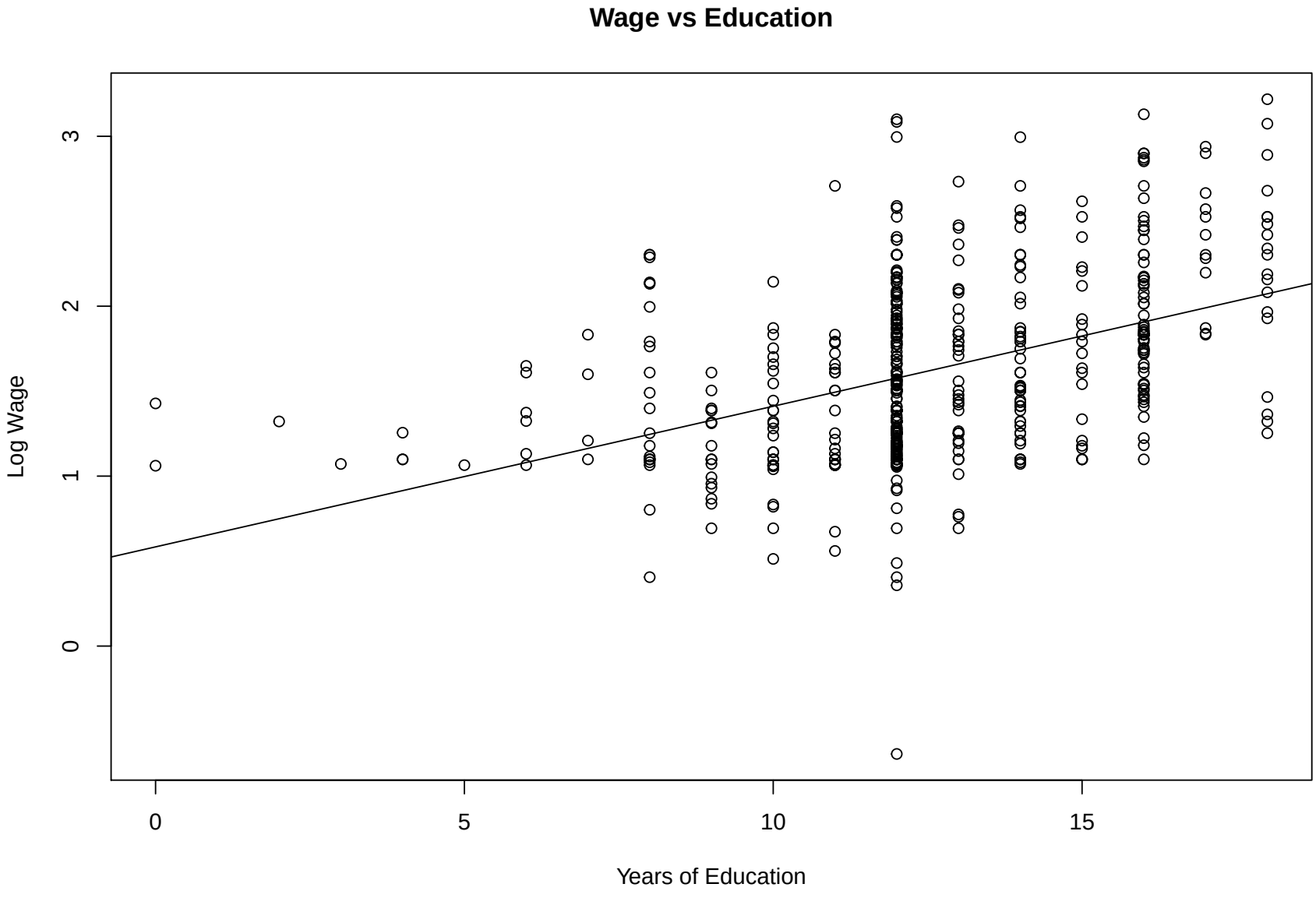
Example: US Employment Data. How would you model this?



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.

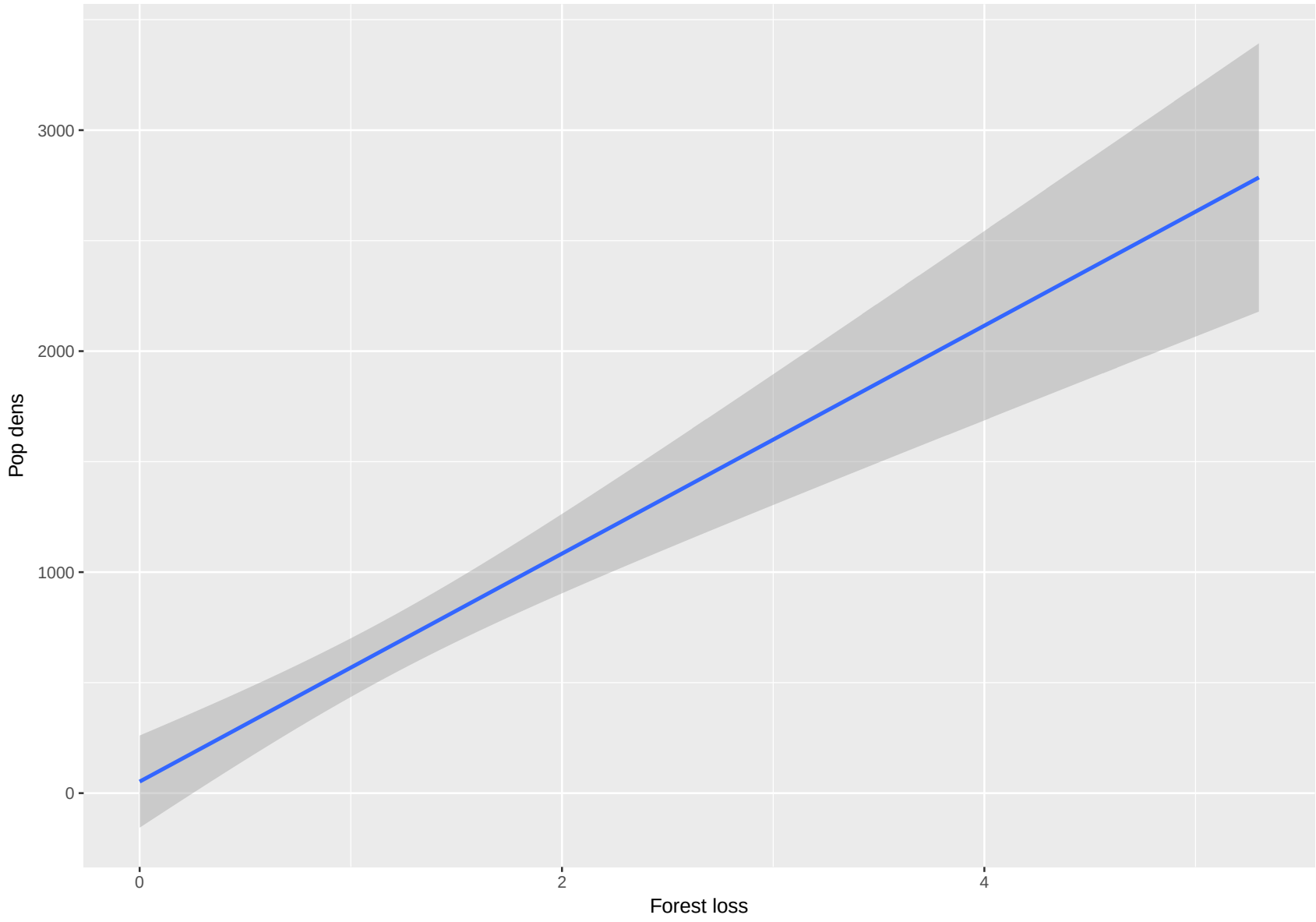


Regression results–what do they mean?

```
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

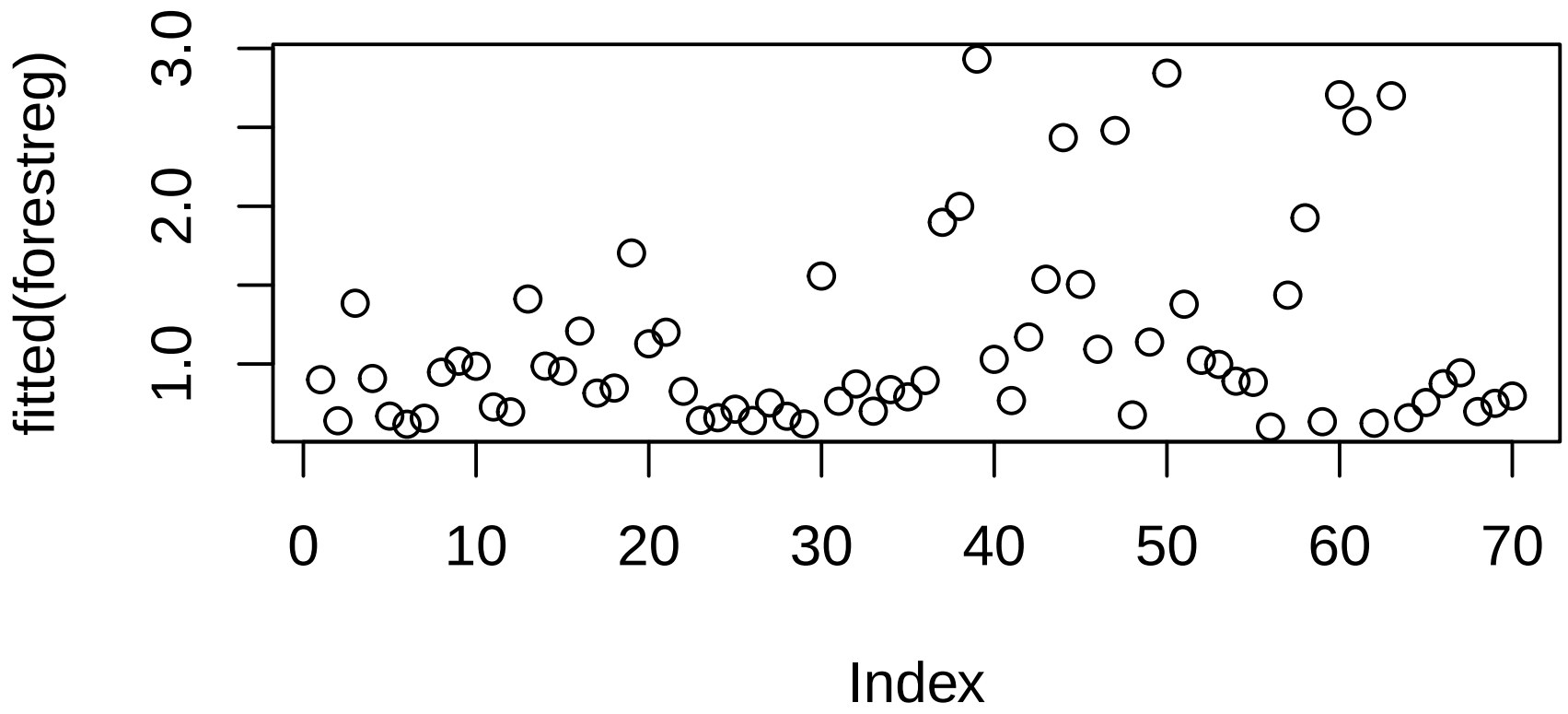
Observations	Residual Std. Error	R^2	Adjusted R^2
70	0.7031	0.4345	0.4262

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

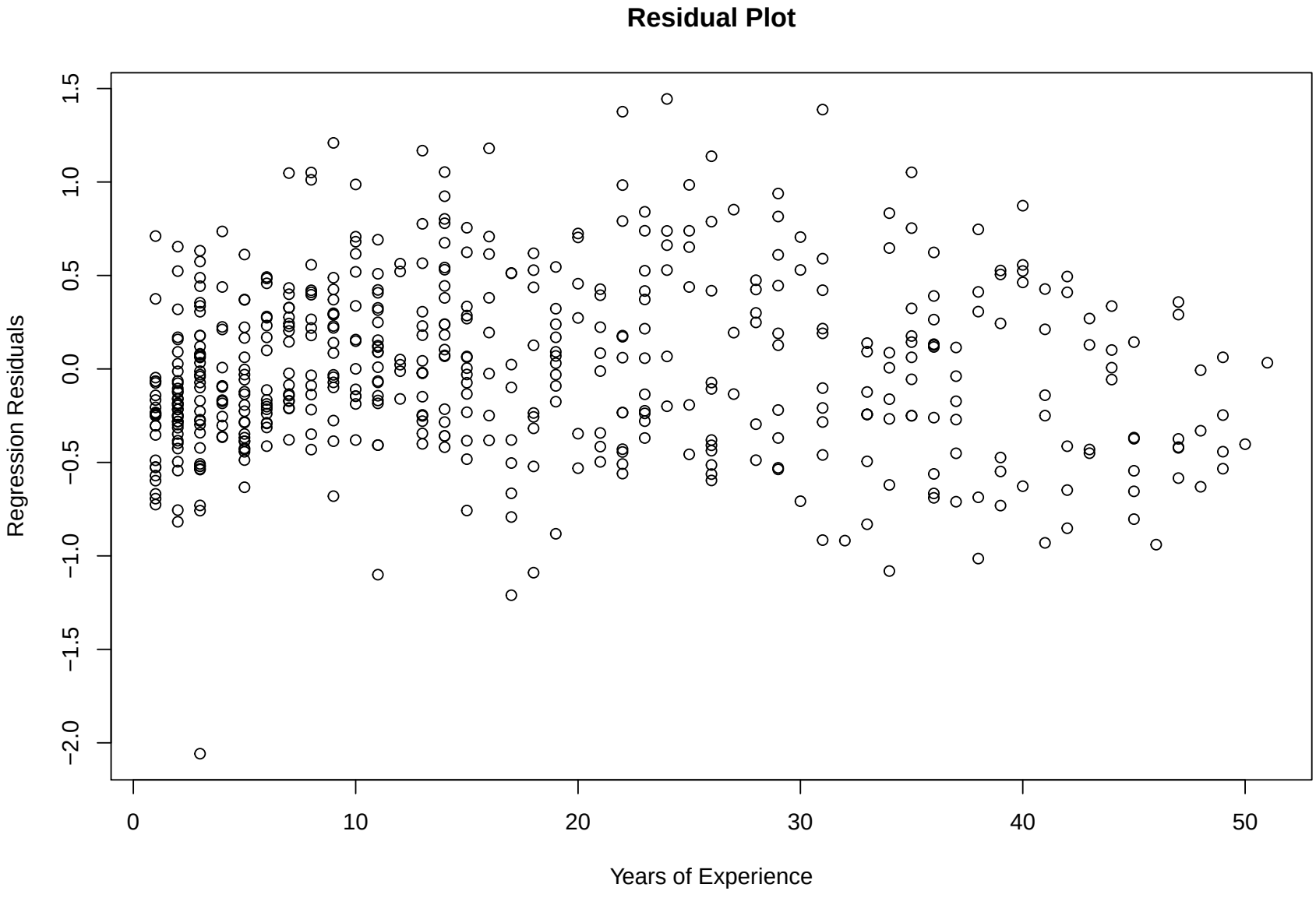
- OLS estimator is linear in β
- But we can model nonlinear functions by allowing $\mathbf{x}$ (or y) to include nonlinear transformations of the data
- For this reason, linearity assumption not as strong as it looks
- Saw this already: use of log wage instead of wage in dollars
- Multiple regression 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))
```

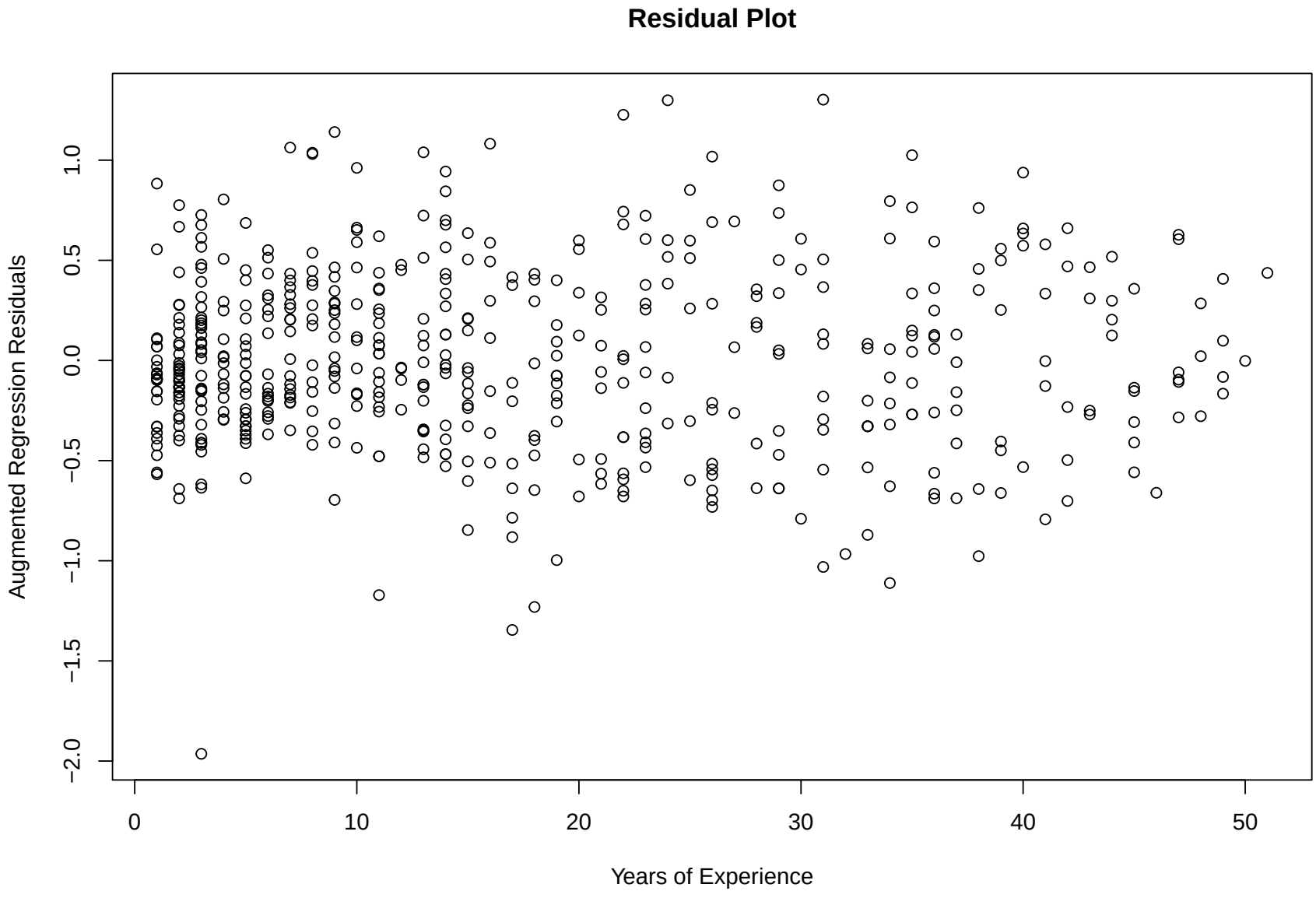
Table 10: Fitting linear model: lwage ~ educ + exper + exper2

	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
- Similarly F test of $\beta_j = 0$ for all $j = 1 \ldots k$ is produced

```
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