# 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

```
## `geom_smooth()` using method = 'loess'
```

```
## Warning: Removed 25 rows containing non-finite values (s
```



Output per worker and capital per worker across the world
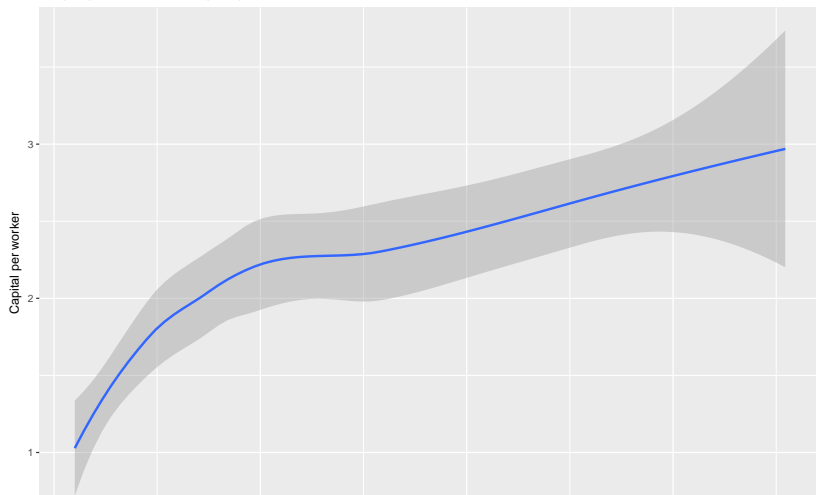
# Assessment

- ▶ 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 $\beta$
- Calculating confidence intervals for $\beta$
- 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.
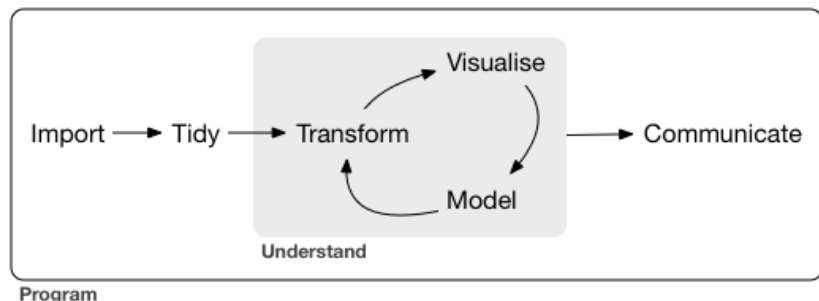
# This is the basic process of applied economic analysis



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
```

Consumption Function, 1950–1985, USA
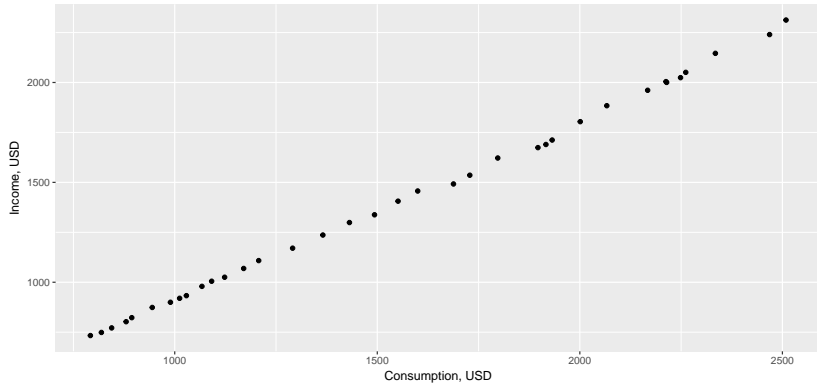


```
REG<-lm(CONS ~ Y, data=cf)
```

# 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 $\alpha$ and $\beta$ 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 ($\alpha$) and slope ($\beta$).
- 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.
```



Index

#
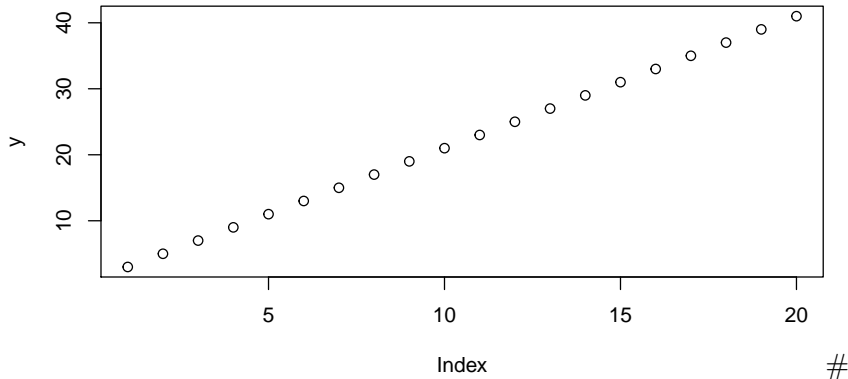
Notation

# 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.



Health Spending Per Person, 1972 = 100

variable
- Denmark
- Germany
- Ireland
- Netherlands
- United Kingdom
- United States
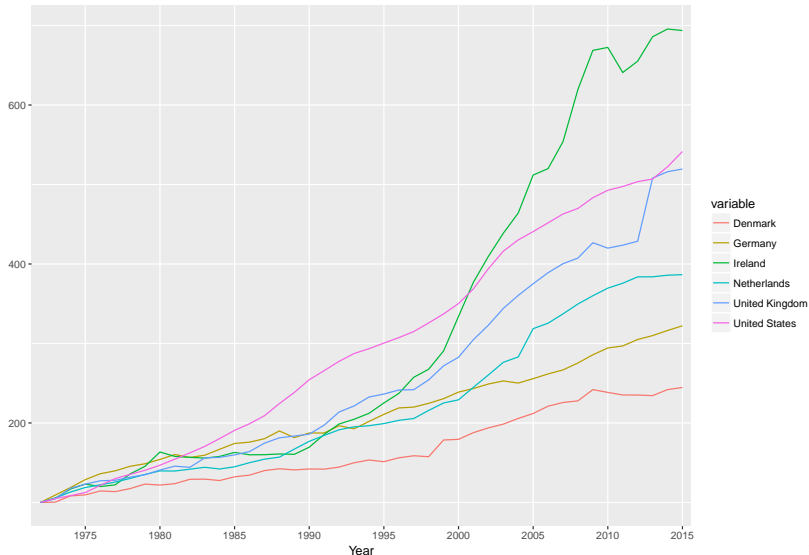
Graphing proportional change over time

#

# We're also interested in:

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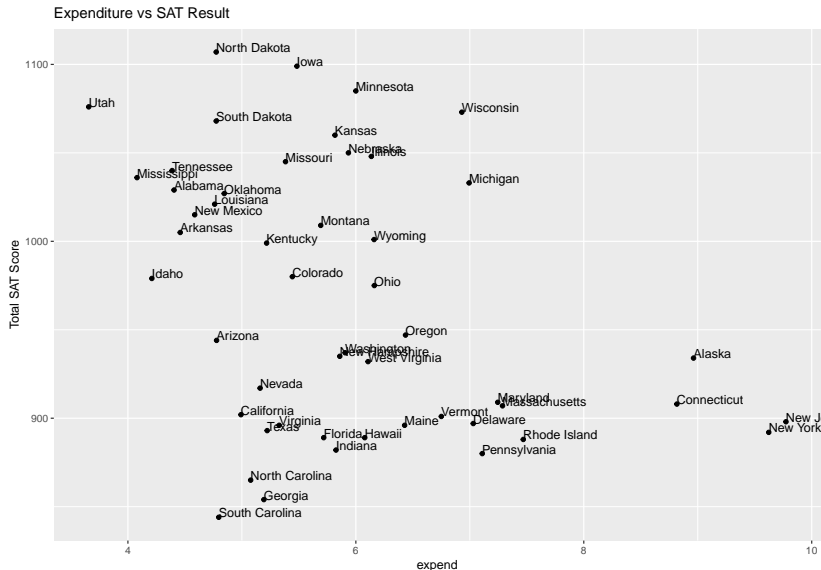4. Assessing strength of evidence.

# 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
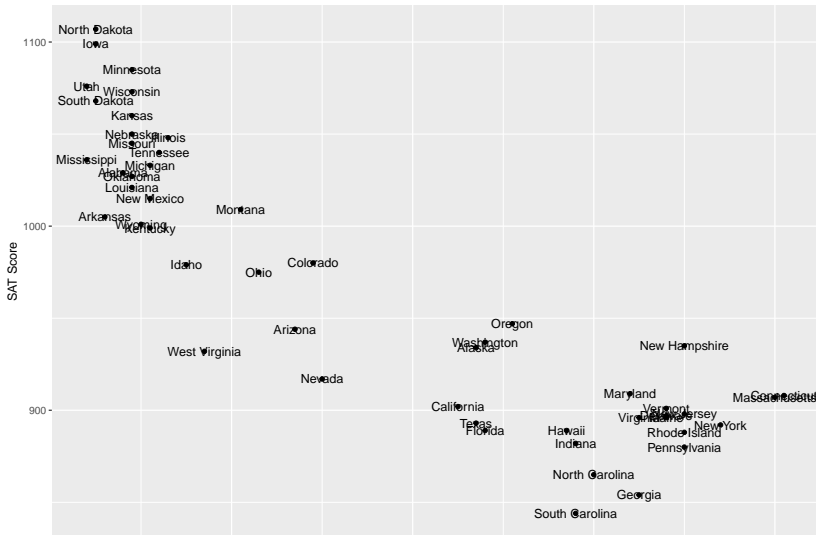


Expenditure vs SAT Result

# 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.

# 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



**Observed Paths**
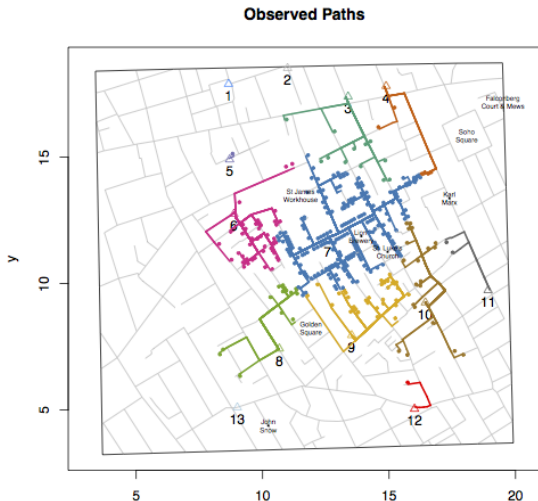
# Thinking in terms of economic analysis: Population



US Population, 1790 to 1970
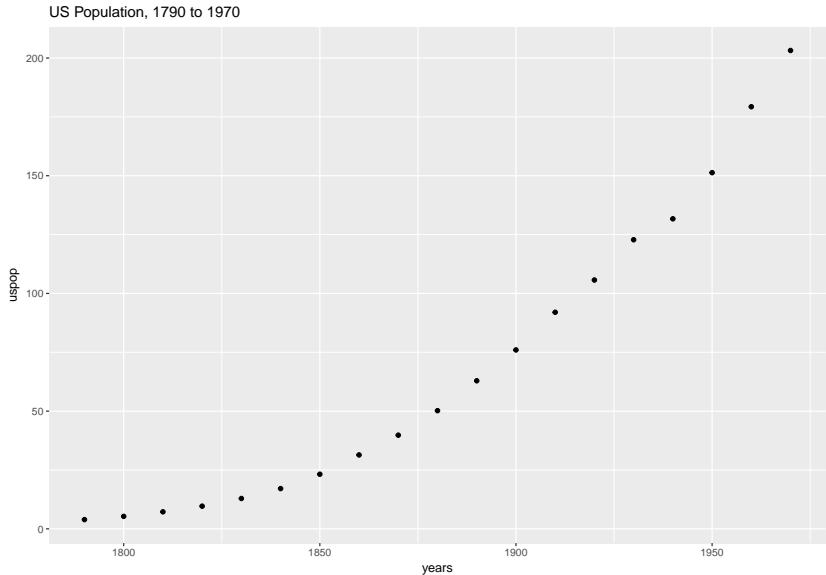
# Yule: What causes poverty?

- 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)

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

- $\Delta$ 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 $\Delta \text{Out}$ increased by 1 percentage point, meaning the administrative district supports more people outside the poorhouse–then $\Delta \text{Paup}$ 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 $\Delta \text{Out}$ 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
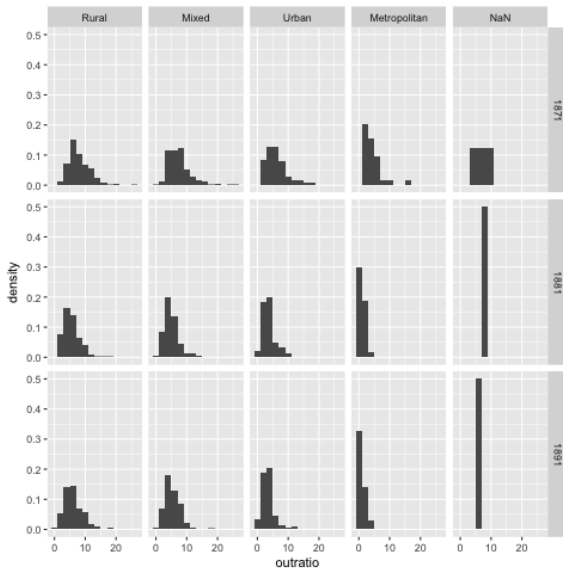
# Why causality matters

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

# Aaaand from Men

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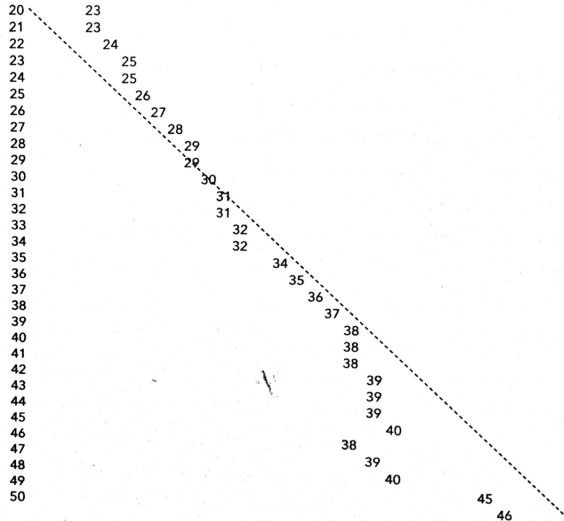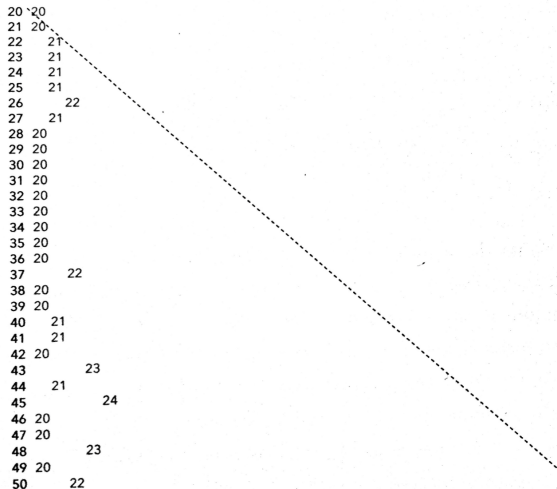


Source: Christian Rudder, *Dataclysm*

# 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 $y$, $r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x} * \frac{y_i-\bar{y}}{s)_y}\right)$
- We're tacitly assuming $s_x$ and $s_y$ aren't zero.

# An example



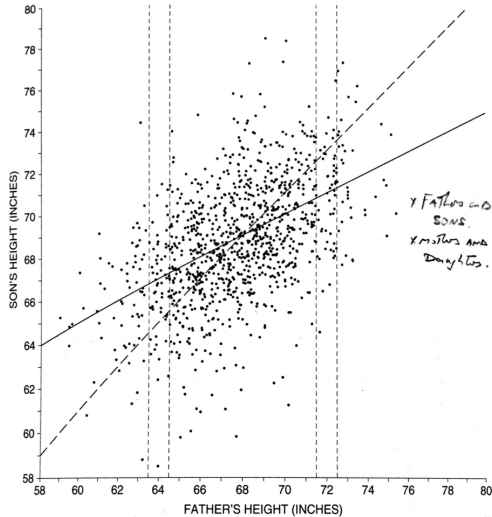Figure 1. Heights of fathers and sons. Pearson and Lee (1903).

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
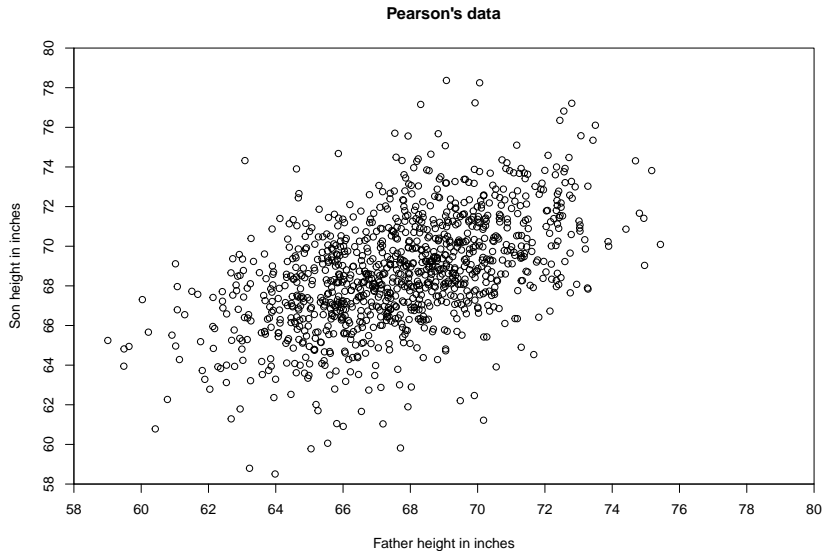
# Interpreting this figure

- 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?)
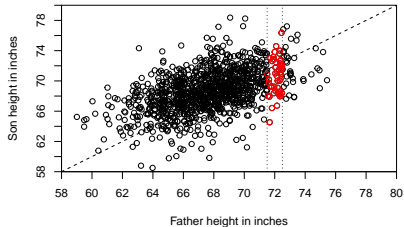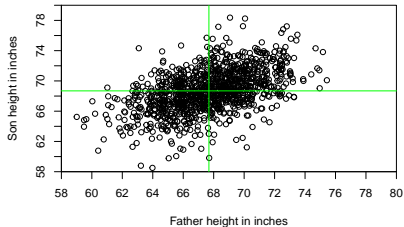
# Pearson with a 45 degree line
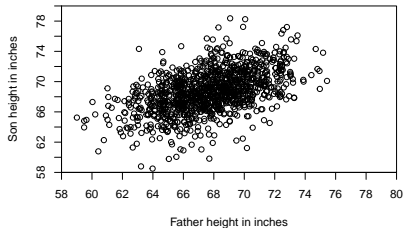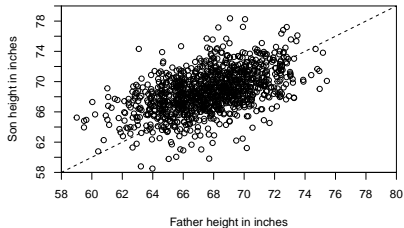
# Intepreting these figures

- 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)



Graph of Averages ~ Pearson's Data
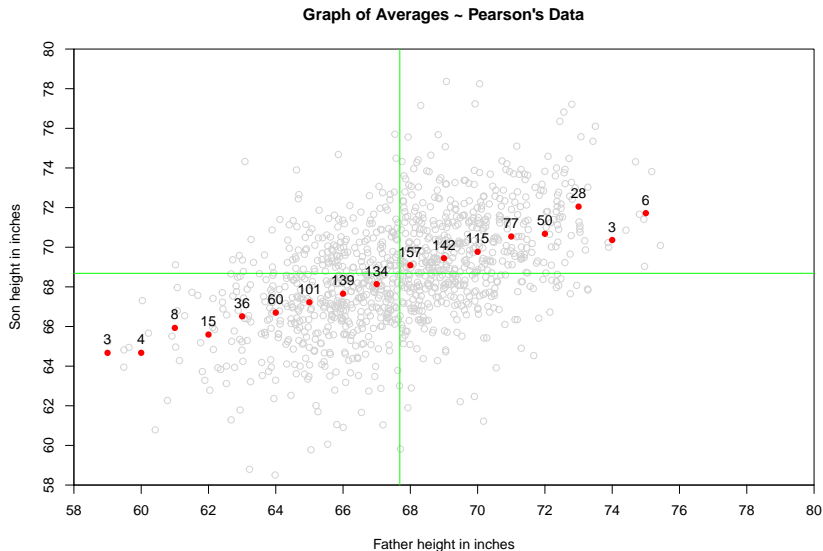
# Another way to look at the data
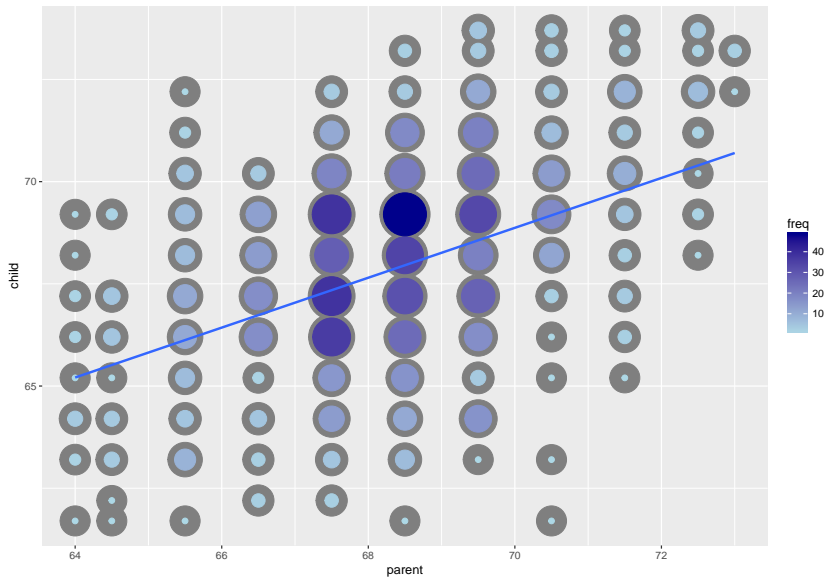
# A subtle distinction

- 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 ($\rho$) 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?
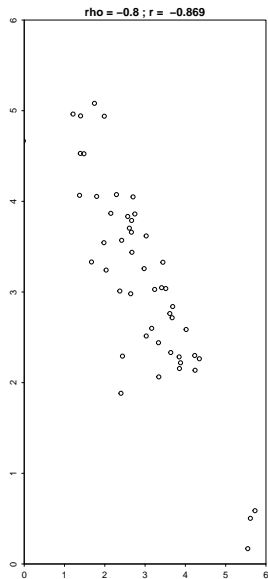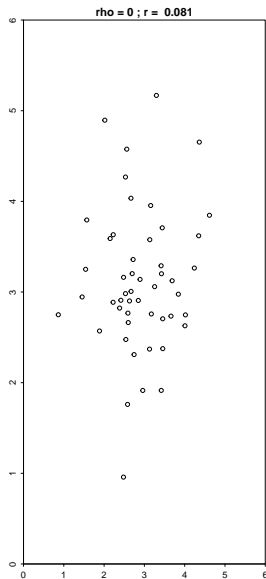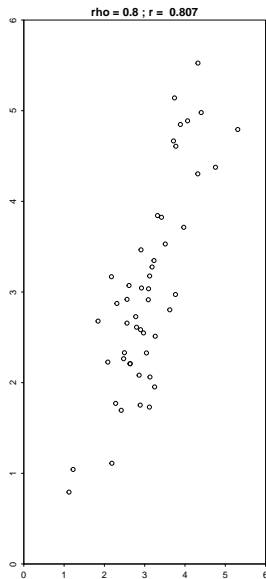
# Root Mean Square Error

- 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 $1, \dots,$ 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 $\epsilon_i$ are independent, identically distributed, mean 0, varriance $\sigma^2$.
- 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 $\epsilon_i$.

## 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 ~ 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 ~ 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.