# Economics 6400: Econometrics

Lecture 1: Introduction

CSU, East Bay

September 26, 2017

# Course description

- The class equips students with the basic theoretical knowledge and computer skills needed to analyze data, and to read and conduct economic research. The focus is on the method of ordinary least squares and on analyzing cross sectional data

# Syllabus overview: Assessment

- Midterm - 25%
  - Multiple choice and/or written
- Homework - 35%
  - Mostly computer/Stata exercises
  - 8 assignments (5% each), drop 1 assignment
  - Can work in groups but assignments *must be written up individually*
  - **Always** attach printouts of your Stata output and any do files
  - *See the Academic Honesty Standards on Blackboard*
- Final exam - 40%
  - Multiple choice and/or written
  - Comprehensive with a bias to last half of the material
  - *You must score above 50% to pass the class*

# Syllabus overview: Accessing Stata

- We will be using Stata/IC 14 or 15. . . a lot
  - ECON 6511 (Advanced Econometrics) and 6896 (Research Methods) will also use Stata
- Accessing Stata
  - Annual license can be purchased for $89 (**recommended**) at http://www.stata.com/order/new/edu/gradplans/student-pricing/
  - Computers in the Learning Commons area of the library: lc 67-78, lc09-lc14, and lc 25-30.
- *Small Stata* is not sufficient for this class as it limits the size of datasets that can be used
  - Do not purchase this version
- MSBA students may prefer using R (or SAS), which is fine
  - Slides and exams will using examples from Stata however

# Syllabus overview: Textbook

- *Introductory Econometrics: A Modern Approach* (6th edition) by Jeffrey Wooldridge
- Absurdly expensive but is also the textbook for ECON 6511 (Advanced Econometrics) and 6896 (Research Methods)
  - Earlier editions (used or new) are fine
- Data sets, which we will use for the homework assignments, are nicely integrated

# A few other things...

- Lecture slides, assignments, datasets, and practice exams will be available on *Blackboard*
- Some students find this class very challenging. Put a lot of time into the assignments. They are the key to passing the class
- I try to respond to email (ryan.lampe@csueastbay.edu) instantly so please don't hesitate to email me when you are stuck or have any other questions
- Questions?

# What is econometrics?

- Econometrics is the "science and art of using economic theory and statistical techniques to analyze economic data" (Stock and Watson, 2015, p. 1)
- Typical goals of econometric analysis
  - Estimating relationships between economic variables
  - Testing economic theories and hypotheses
  - Forecasting economic variables
  - Evaluating and implementing government and business policy

# Types of questions answered using econometrics

- Economics questions we might answer using econometric techniques:
  - Do smaller class sizes improve elementary school education?
  - Is there racial discrimination in the market for home loans?
  - How much do cigarettes taxes reduce smoking?
  - How much will U.S. GDP increase by next year?
- Often the goal is to obtain a numerical relationship between two variables
  - E.g. between standardized test scores and class size: a 10% drop in class size leads to a 2% increase in average test scores
  - We will also need information on how precise this number is

# Why do we need econometrics?

- Econometricians typically analyze *nonexperimental* data that is not accumulated through controlled/laboratory experiments
- Randomized control experiment (RCE): Imagine you were interested in the causal effect of fertilizer on tomato yields
  - Conduct an RCE in which you **randomly** add fertilizer to some plots and none to other plots
  - At the end of the growing season, the difference between the average yield of plots with fertilizer (the "treatment" group) and the average yield of plots without fertilizer (the "control" group) is a measure of the causal impact of fertilizer
- The random assignment of fertilizer to plots eliminates the possibility of a systematic relationship between the characteristics of plots (e.g. how sunny the plot is) and the assignment of fertilizer
  - Any systematic difference between yields is due to the fertilizer

# Why do we need econometrics?

- Consider a similar experimental methodology applied to estimating the impact of class size on education outcomes
  - Randomly assign students in a district to classrooms of two distinct sizes
  - Compute the difference between the average test scores from both class sizes
- Clearly, such experiments are seldom possible because they are often unethical, impossible to execute, and expensive
  - If you were interested in the relationship between the use of seat belts and highway fatalities, would it be ethical to randomly remove seat belts from some cars but not others?

# Why do we need econometrics?

- Most data is therefore observational data, e.g. from surveys, administrative records, historical records
- These data pose major challenges, **which econometrics is design to tackle**
  - Suppose you obtained observational data on class sizes and standardized test scores and found that test scores are systematically higher for students in smaller classes
  - This would not be strong evidence since it is possible that students in smaller classes tend to live in wealthier districts that have more opportunities for learning outside the classroom
- Econometrics provides a toolbox to **control** for the existence of these opportunities and determine a **causal** relationship between class size and educational outcomes

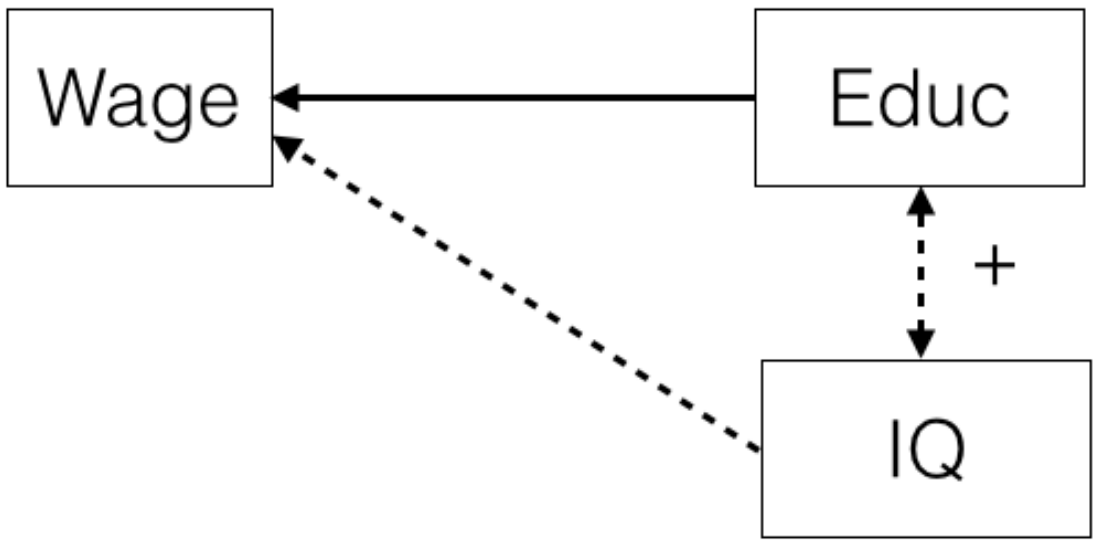
# Causality and the notion of ceteris paribus

- Definition of causal effect of $x$ on $y$:
  - "How does variable $y$ change if variable $x$ is changed *but all other relevant factors are held constant*?"
  - E.g. "How do standardized test scores change if class size is reduced but all other relevant factors (family wealth, availability of tutoring etc.) are held constant?"
- Most economic questions are ceteris paribus questions
  - Example: How much do cigarettes taxes reduce smoking?
- It is important to define which causal effect one is interested in

# Causality and the notion of ceteris paribus

- Measuring the return to education
  - "If a person is chosen from the population and given another year of education, by how much will his or her wage increase?"
  - Implicit assumption: all other factors that influence wages such as experience, family background, IQ etc. are held fixed (or controlled for)
- Infeasible RCE:
  - Choose a group of people; randomly assign different amounts of education to them (infeasible!); compare wage outcomes
- Problem without random assignment: amount of education is related to other factors that influence wages
  - E.g., intelligence (IQ), which makes the cost of obtaining education lower, and also may lead to overestimating the returns to that education

# Causality and the notion of ceteris paribus




- IQ (omitted variable) is positively correlated with training and wage

# Economic model of education and worker productivity

- What is the effect of an additional year of education on worker productivity?
- Assume wage is a mathematical function of education and IQ:

$$wage = f(education, IQ)$$

where

  - *wage* is hourly wage
  - *educ* is years of education
  - *IQ* is a measure of the worker's intelligence
- Functional form of relationship not specified
  - We will assume it is linear on the next slide

# Econometric model of education and worker productivity

- Simple linear equation for education and worker productivity

$$wage = \beta_0 + \beta_1 educ + \beta_2 IQ + u$$

  where
  - *wage* is hourly wage
  - *educ* is years of education
  - *IQ* is a measure of the worker's intelligence
  - $u$ is unobserved determinants of the wage, e.g. quality of education, family background, job experience etc.
- Most of econometric deals with the specification of the error $u$
- Econometric models may be used for hypothesis testing
  - E.g., the parameter $\beta_1$ represents effect of education on wage
  - How large is this effect? Is it different from zero?

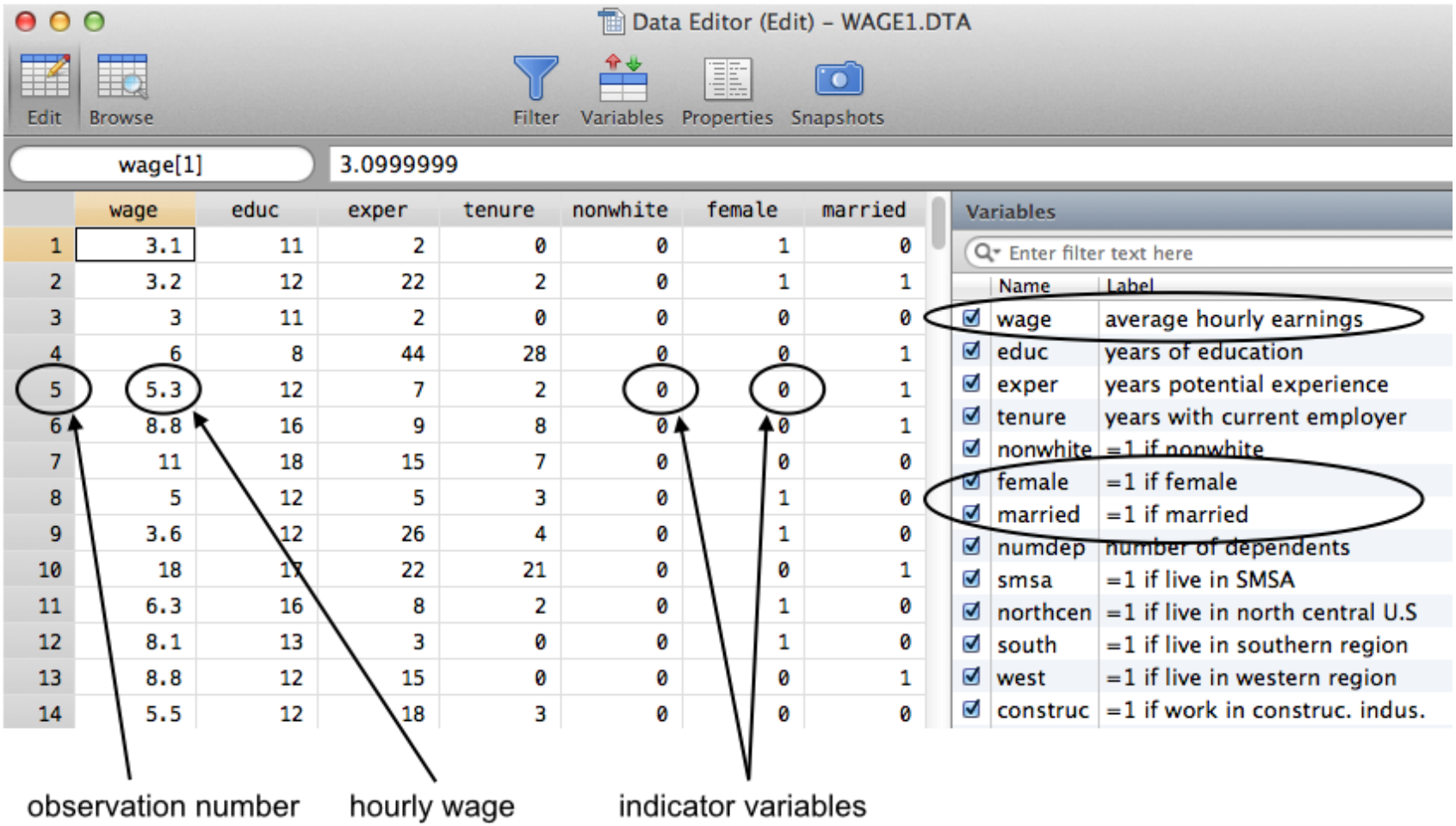
# Types of data

- Cross-sectional data
  - This course will almost exclusively deal with this type of data
- Time series data
- Pooled cross sections
- Panel/Longitudinal data

# Cross-sectional data sets

- Sample of individuals, households, firms, cities, states, countries or other unit of interest *at a given point of time/in a given period*
- Cross-sectional observations are more or less independent
  - For example, observations are sampled randomly
  - Sometimes pure random sampling is violated, e.g. units refuse to respond in surveys, or if sampling is characterized by clustering
- Cross-sectional data typically encountered in applied microeconomics

# Cross-sectional data set on wages and other characteristics



# Cross-sectional data set on box office statistics

| shortname | revenue | sex | violence | budget |
|---|---|---|---|---|
| Iron Monkey | 14694904 | 3 | 6 | 11 |
| Chicken Run | 106834564 | 1 | 4 | 45 |
| Fantastic Four | 154696080 | 4 | 5 | 100 |
| Frida | 25885000 | 7 | 7 | 12 |
| From Hell | 31602566 | 7 | 10 | 35 |
| the Rings: The Fellowship of the Ring | 313364114 | 1 | 7 | 93 |
| Mission: Impossible II | 215409889 | 3 | 6 | 125 |
| Resident Evil | 40119709 | 5 | 8 | 33 |
| The Shipping News | 11434216 | 7 | 6 | 38 |
| X-Men | 157299717 | 1 | 6 | 75 |
| Men in Black II | 190418803 | 2 | 5 | 140 |
| Titan A.E. | 22753426 | 2 | 5 | 75 |
| The Emperor's New Groove | 89302687 | 0 | 3 | 100 |
| ars: Episode II - Attack of the Clones | 302191252 | 2 | 5 | 115 |
| ars: Episode III - Revenge of the Sith | 380270577 | 2 | 7 | 113 |
| Return to Me | 32722798 | 2 | 3 | 24 |

film title

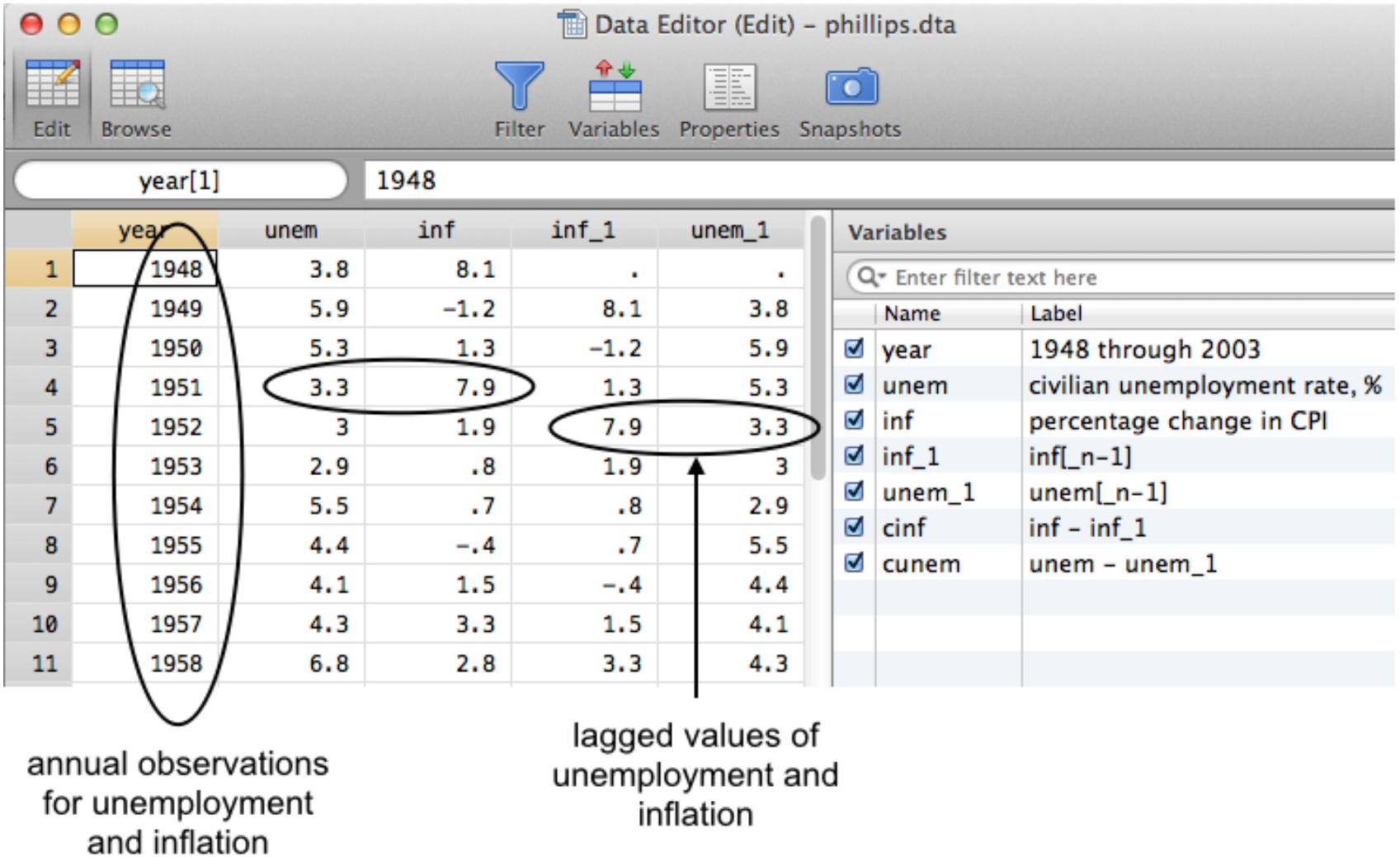
box office revenue

measures of sex & violence

# Time series data

- Observations of a variable or several variables over time
- For example, stock prices, money supply, consumer price index (CPI), gross domestic product (GDP), annual homicide rates, automobile sales,. . .
- Time series observations are typically serially correlated
- Ordering of observations conveys important information
- Data frequency: daily, weekly, monthly, quarterly, annually,. . .

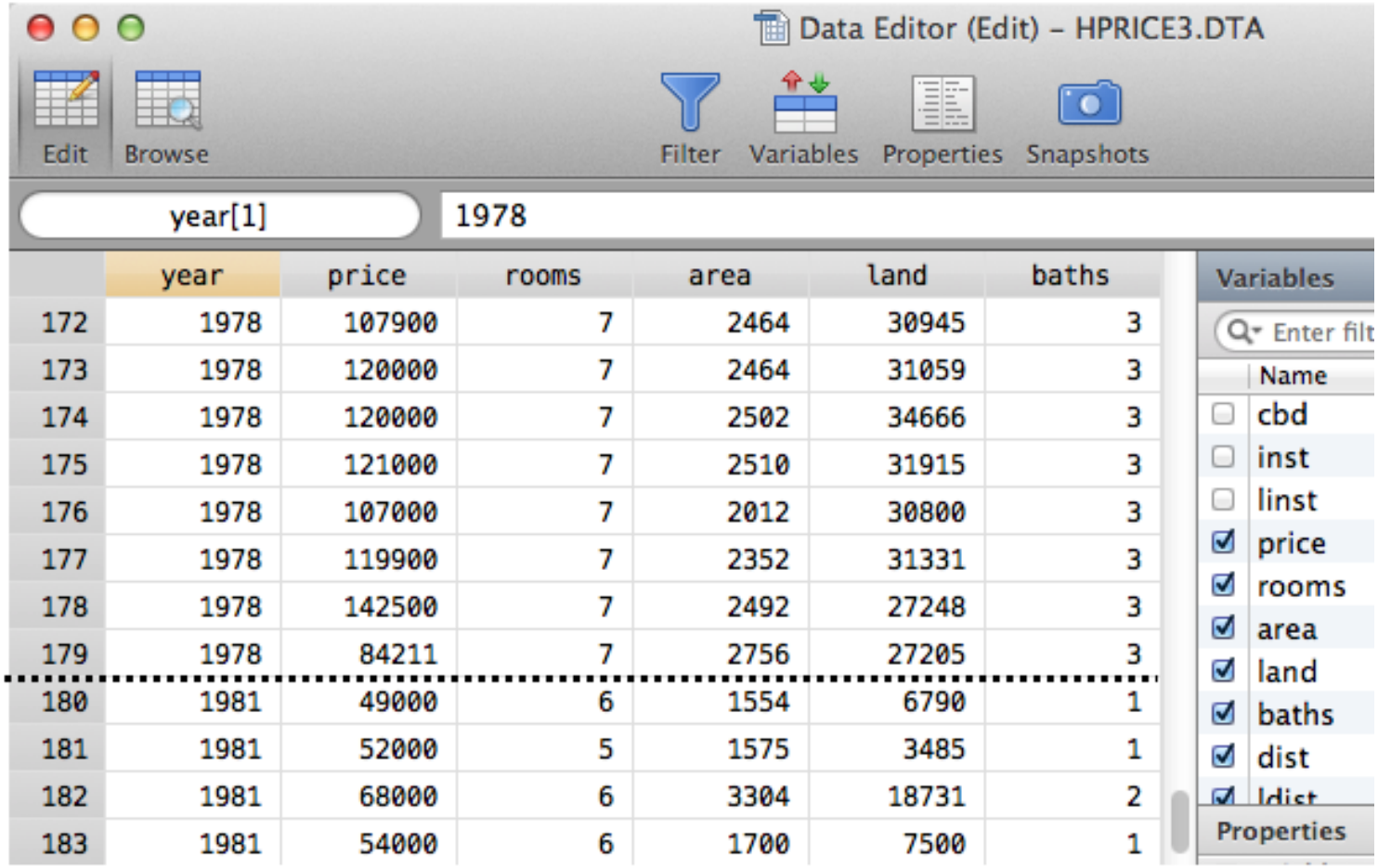
# Time series data: Unemployment and inflation



# Pooled cross sections

- Two or more cross sections are combined in one data set
- Cross sections are drawn independently of each other
- Pooled cross sections often used to evaluate policy changes
- Example: Evaluate effect of changes in property taxes on house prices
  - Random sample of house prices for the year 1978
  - A **new** random sample of house prices for the year 1981
  - Compare before/after (1978: before reform, 1981: after reform)

# Pooled cross sections: House prices in 1978 and 1981




| | year | price | rooms | area | land | baths |
|---|---|---|---|---|---|---|
| 172 | 1978 | 107900 | 7 | 2464 | 30945 | 3 |
| 173 | 1978 | 120000 | 7 | 2464 | 31059 | 3 |
| 174 | 1978 | 120000 | 7 | 2502 | 34666 | 3 |
| 175 | 1978 | 121000 | 7 | 2510 | 31915 | 3 |
| 176 | 1978 | 107000 | 7 | 2012 | 30800 | 3 |
| 177 | 1978 | 119900 | 7 | 2352 | 31331 | 3 |
| 178 | 1978 | 142500 | 7 | 2492 | 27248 | 3 |
| 179 | 1978 | 84211 | 7 | 2756 | 27205 | 3 |
| 180 | 1981 | 49000 | 6 | 1554 | 6790 | 1 |
| 181 | 1981 | 52000 | 5 | 1575 | 3485 | 1 |
| 182 | 1981 | 68000 | 6 | 3304 | 18731 | 2 |
| 183 | 1981 | 54000 | 6 | 1700 | 7500 | 1 |

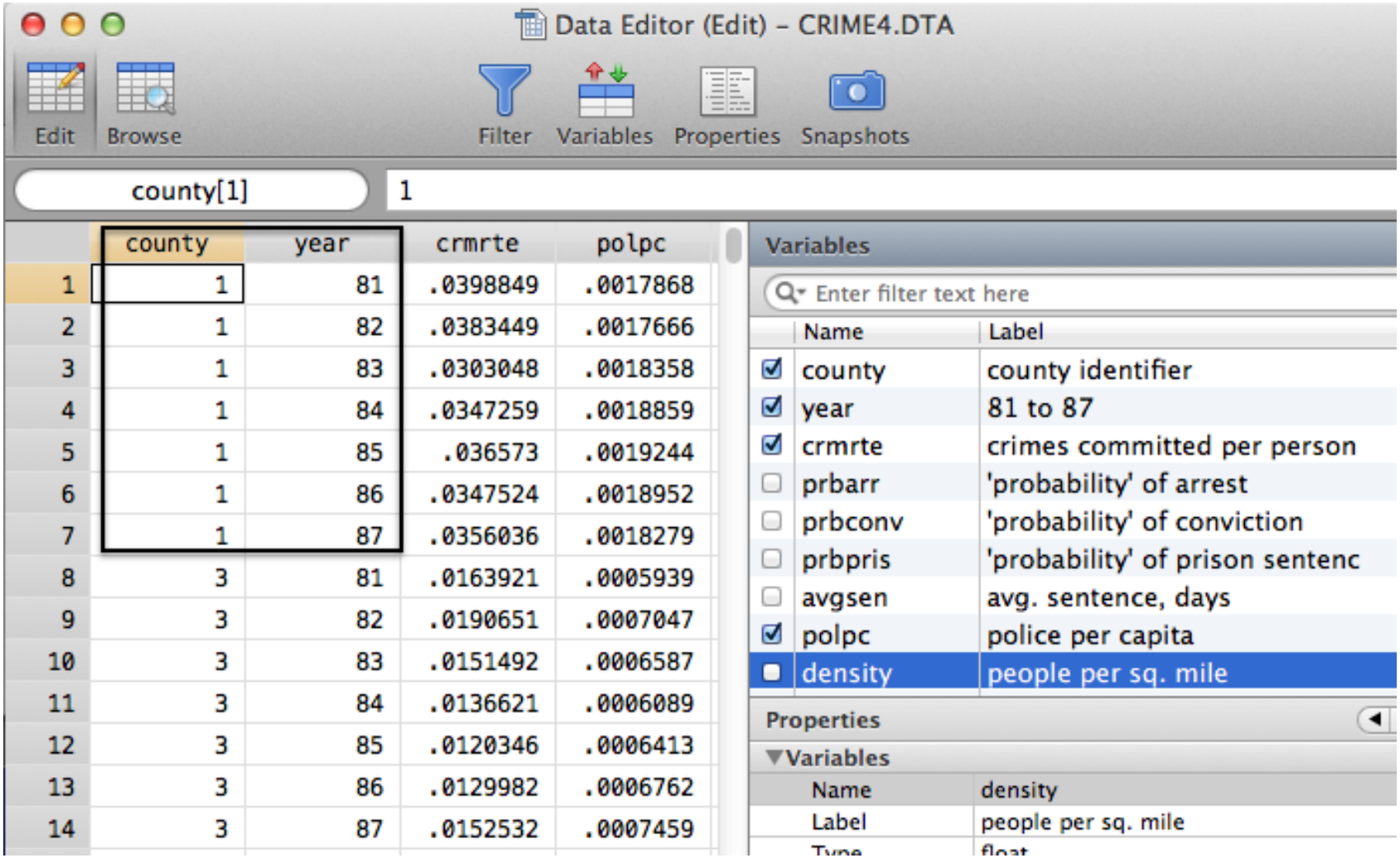
before

after

# Panel or longitudinal data

- The **same** cross-sectional units are followed over time
- Panel data have a **cross-sectional and a time series dimension**
- Panel data can be used to account for time-invariant unobservables
- Panel data can be used to model lagged responses
- Example
  - City crime statistics; each city is observed in two years
  - Time-invariant unobserved city characteristics may be modeled
  - Effect of police on crime rates may exhibit time lag

# Pooled cross sections: Crime across counties and years



# Overarching goals of this class

- Given a generic linear regression equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

  I want you to be able to:

  1. Estimate this model using Stata
  2. Interpret what the coefficients $\beta_1$, $\beta_2$ etc. **actually mean**
  3. Identify any omitted variables (contained in $u$) and explain how they affect your results

- These goals will be the focus of the in-class quizzes

# Statistics review: Summation

- **Summation operator** is useful for manipulating expressions involving the sums of many numbers, and is crucial in statistics and econometrics. If $\{x_i : i = 1, \ldots, n\}$ denotes a sequence of $n$ numbers (e.g. observations), we write the sum of these numbers as:

$$\sum_{i=1}^{n} x_i \equiv x_1 + x_2 + \cdots + x_n$$

- Properties:

$$\sum_{i=1}^{n} cx_i = c \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} (ax_i + by_i) = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} y_i$$

# Statistics review: Sample average

- Given $n$ numbers $\{x_i : i = 1, \ldots, n\}$, we compute their **average** or *mean* by adding them up and dividing by $n$:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Important property of averages:

$$\sum_{i=1}^{n} (x_i - \overline{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \overline{x} = \sum_{i=1}^{n} x_i - n\overline{x} = n\overline{x} - n\overline{x} = 0$$

- Example: $n = 5$ and $x_1 = 6$, $x_2 = 1$, $x_3 = -2$, $x_4 = 0$, and $x_5 = 5$. Then $\overline{x} = 2$, and demeaned sample is $\{4, -1, -4, -2, 3\}$. Adding these numbers gives 0.

# Statistics review: Sample average

- Additional algebraic facts involving deviations from sample averages:

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

- Which can be shown:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \overline{x})^2 &= \sum_{i=1}^{n}(x_i^2 - 2x_i\overline{x} + \overline{x}^2) \\ &= \sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + n\overline{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2n\overline{x}^2 + n\overline{x}^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2\end{aligned}$$

# Statistics review: Properties of linear functions

- Much of econometrics involves estimating and interpreting linear functions. If $x$ and $y$ are two variables related by:

$$y = \beta_0 + \beta_1 x,$$

then $y$ is a **linear function** of $x$, and $\beta_0$ and $\beta_1$ are two parameters (numbers) describing the relationship. The **intercept** is $\beta_0$, and the **slope** is $\beta_1$.

- The defining feature of linear function is that the change in $y$ is always equal to $\beta_1$ times the change in $x$:

$$\Delta y = \beta_1 \Delta x$$

where $\Delta$ is the "change" and $\beta_1$ is the constant **marginal effect** of $x$ on $y$.

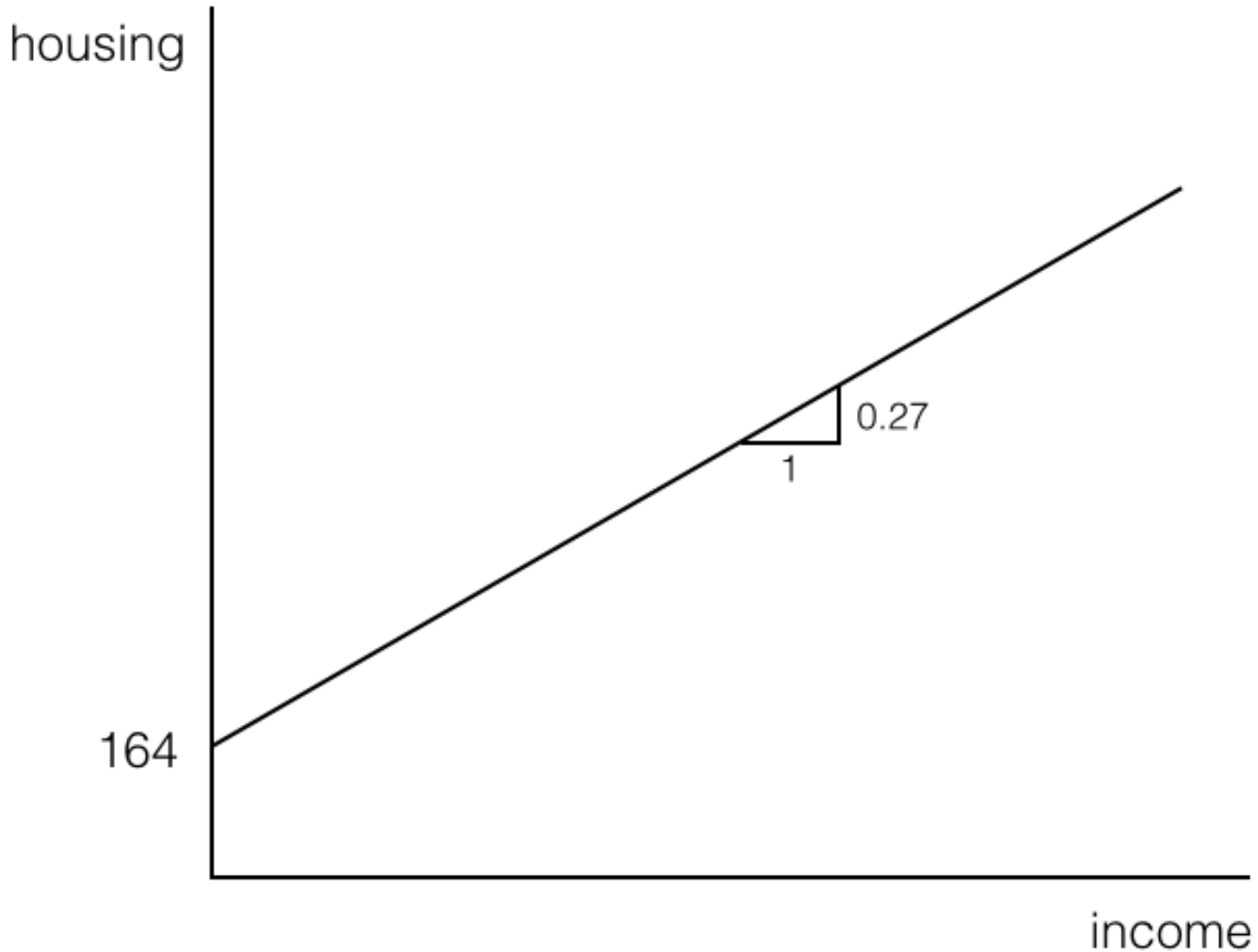
# Statistics review: Linear housing expenditure function

- Suppose the relationship between annual housing expenditure and annual income is

$$\textit{housing} = 164 + 0.27\ \textit{income}$$

so that each additional dollar of income increases housing expenditure by 27 cents (and housing expenditure is \$164 for those with zero income)

# Statistics review: Linear housing expenditure function



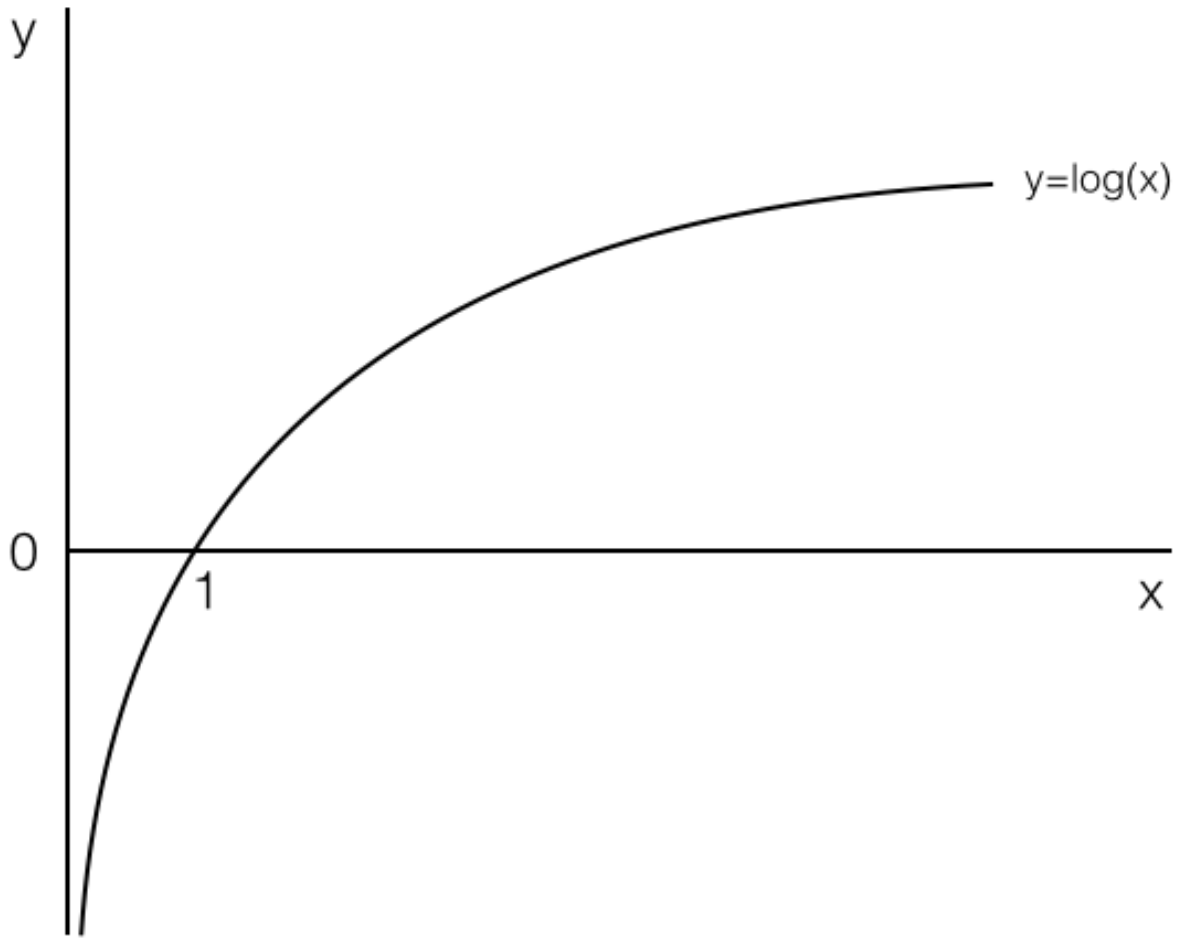
# Statistics review: The natural logarithm

- One of the most important nonlinear function in econometrics is the **natural logarithm**:

$$y = \log(x)$$

  or equivalently $\ln(x)$ or $\log_e(x)$.
- Relationship between $y$ and $x$ exhibits diminishing marginal returns: The effect of $x$ on $y$ decreases (but never becomes negative) as $x$ becomes larger.
  - The short-run relationship between employment and firm output exhibits diminishing marginal returns because (holding capital fixed) additional workers get in each other's way leading to lower output per worker

# Statistics review: The natural logarithm



# Statistics review: log-level model

- If the equation that is specified is:

$$\log(y) = \beta_0 + \beta_1 x$$

then approximate % change in $y$ is: $\%\Delta y \approx 100 \cdot \beta_1 \Delta x$.

  - An accurate prediction is $\%\Delta y = (e^{\beta_1 \Delta x} - 1) \cdot 100$. There will be little difference between this and the approximation if $\beta_1$ is small.

- Example: Suppose that hourly wage and years of education are related by:

$$\log(wage) = 2.78 + 0.094\ educ$$

then $\%\Delta wage = 100(0.094)\Delta educ = 9.4\ \Delta educ$.

# Statistics review: level-log model

- If the equation that is specified is:

$$y = \beta_0 + \beta_1 \log(x)$$

then approximate change in $y$ is $\Delta y \approx \beta_1 \frac{\Delta x}{x} = (\frac{\beta_1}{100})(\%\Delta x)$.

- Example: If labor supply of a worker can be described by:

$$hours = 33 + 45.1 \log(wage)$$

then $\Delta hours = (\frac{45.1}{100})(\%\Delta wage) = 0.451\ \%\Delta wage$.

# Statistics review: log-log model

- If the equation that is specified is:

$$\log(y) = \beta_0 + \beta_1 \log(x)$$

  then $\beta_1$ is the elasticity of $y$ with respect to $x$ (assuming $x$, $y \geq 0$).
- Example: If $q$ is quantity demanded and $p$ is price and demand is given by:

$$\log(q) = 4.7 - 1.25 \log(p)$$

  then the price elasticity of demand is $-1.25$ so a 1% increase in price leads to a 1.25% fall in the quantity demanded.

# Summary of interpreting $\beta_1$ with and without logs

| | | $y$ | |
|---|---|---|---|
| | | *level* | *log* |
| $x$ | *level* | $\Delta y = \beta_1 \Delta x$ | $\%\Delta y = 100 \cdot \beta_1 \Delta x$ ** |
| | *log* | $\Delta y = \frac{\beta_1}{100} \cdot \%\Delta x$ | $\frac{\%\Delta y}{\%\Delta x} = \beta_1$ |

** Accurate interpretation is: $\%\Delta y = (e^{\beta_1 \Delta x} - 1) \cdot 100$

# Statistics review: Sampling

- Let $Y$ be a random variable representing a population with probability density function $f(y;\theta)$, which depends on the single parameter of $\theta$. The probability density function (pdf) of $Y$ is assumed to be known (e.g. it might be normal) except for the value of $\theta$.
- If we take samples (e.g. random samples) from this population we can learn something about $\theta$.
  - Example 1: if $f$ is distributed $N(\mu, \sigma^2)$ then we seek to learn both $\mu$ and $\sigma^2$ so here $\theta = (\mu, \sigma^2)$
  - Example 2: If $Y$ is described by a linear function $y = \beta_0 + \beta_1 x$ then we seek to learn $\beta_0$ and $\beta_1$ by sampling the data on $y$ and $x$ and running regressions so here $\theta = (\beta_0, \beta_1)$

# Statistics review: Sample average and sample variance

- To estimate the population mean $\mu$, the sample average is defined as:

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

- For hypothesis testing, we need to estimate the variance $\sigma^2$ from a population with mean $\mu$. The sample variance is defined as:

$$S^2 = S^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2,$$

where the division by $n-1$ rather than $n$ accounts from the fact that the mean $\mu$ is estimated rather than known.

# Why divide by $n - 1$?

- Imagine you had 5 observations: $\{1, 2, 3, 4, 5\}$ with sample average, $\overline{y} = 3$.
- When calculating the sample variance we really only have 4 useful observations since the fifth observation can easily be recovered given $\overline{y}$.
  - If we know the first four observations: $\{1, 2, 3, 4\}$ then we can recover the fifth observation, $y_5$, from the following equation for calculating the sample mean (i.e. one equation, one unknown):

$$\overline{y} = 3 = \frac{1}{5}(1 + 2 + 3 + 4 + y_5)$$

# Statistics review: Confidence intervals

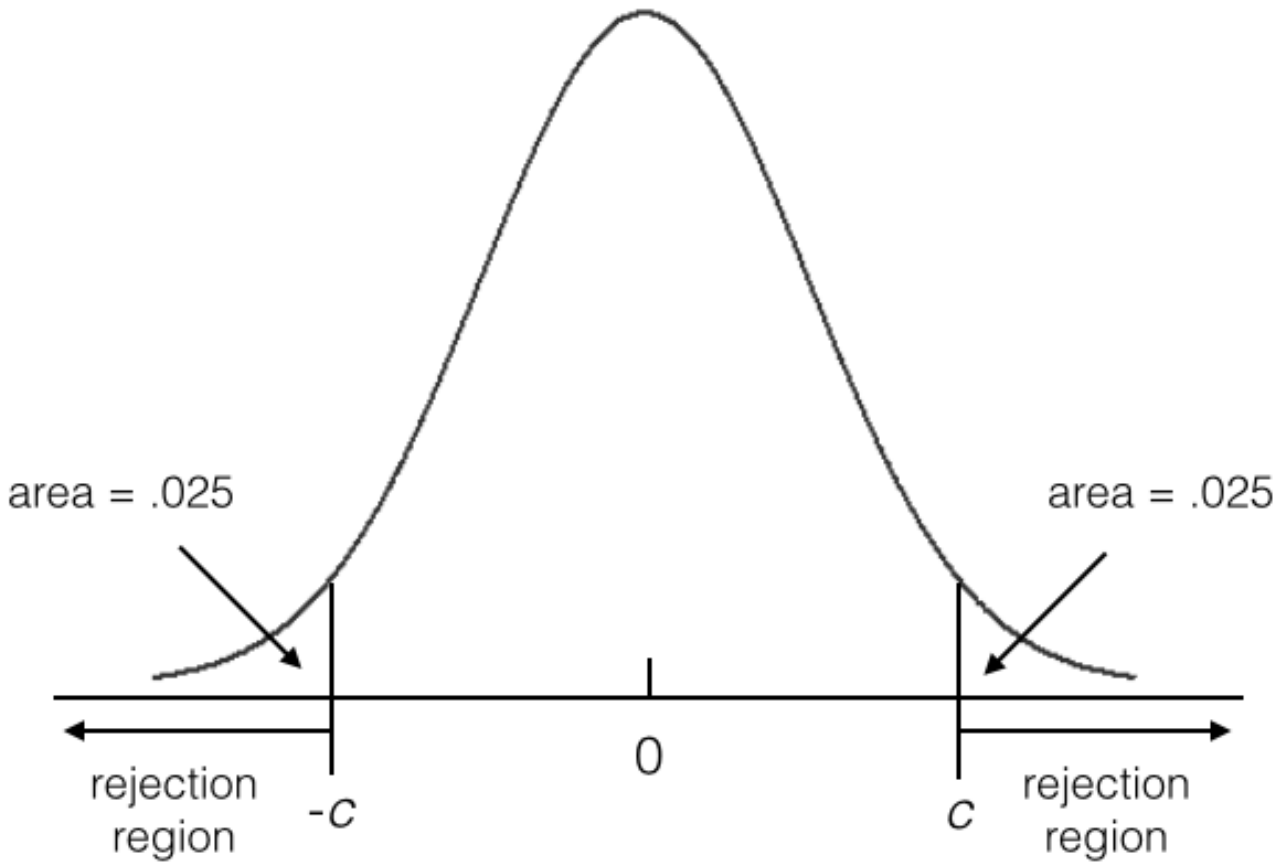
- Suppose you obtain an estimate of $\mu$ equal to $\overline{y}$. Can we make some statement about how close $\overline{y}$ is to $\mu$ assuming observations come from a normal distribution?
- The 95% confidence interval when $\sigma$ is unknown is:

$$[\overline{y} - c \cdot \frac{s}{\sqrt{n}}, \overline{y} + c \cdot \frac{s}{\sqrt{n}}],$$

where $c$ is the value such that 95% of the area in the $t_{n-1}$ distribution is between $-c$ and $c$.

# Statistics review: Hypothesis testing about the mean from a Normal distribution

- Critical value $c$ is chosen such that 95% of the area in the $t_{n-1}$ distribution is between $-c$ and $c$



# Statistics review: Confidence intervals

- Suppose you obtain an estimate of $\mu$ equal to $\overline{y}$. Can we make some statement about how close $\overline{y}$ is to $\mu$ assuming observations come from a normal distribution?
- The 95% confidence interval when $\sigma$ is unknown is:

$$[\overline{y} - c \cdot \frac{s}{\sqrt{n}}, \overline{y} + c \cdot \frac{s}{\sqrt{n}}],$$

  where $c$ is the value such that 95% of the area in the $t_{n-1}$ distribution is between $-c$ and $c$.
- E.g., if $\overline{y} = 0.9$, $n = 21$, and $s = 0.2$, then we are 95% confident that $\mu$ lies in the range:

$$[0.9 - 2.086 \cdot \frac{0.2}{\sqrt{21}}, 0.9 + 2.086 \cdot \frac{0.2}{\sqrt{25}}] = [0.809, 0.991]$$

# If $n = 21$ (df $= n - 1 = 20$) and we want a $\leq 5\%$ probability of a Type I error then $c = 2.086$ (with two tails)

Critical Values of the *t* Distribution

| | | Significance Level | | | | |
|---|---|---|---|---|---|---|
| | **1-Tailed:** | **.10** | **.05** | **.025** | **.01** | **.005** |
| | **2-Tailed:** | **.20** | **.10** | **.05** | **.02** | **.01** |
| **Degrees of Freedom** | 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 |
| | 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 |
| | 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 |
| | 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 |
| | 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 |
| | 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 |
| | 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 |
| | 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 |
| | 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 |
| | 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 |
| | 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 |
| | 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 |
| | 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 |
| | 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 |
| | 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 |
| | 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 |
| | 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 |
| | 18 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 |
| | 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 |
| | 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 |

# Statistics review: Standard deviation vs. standard error

- $\frac{s}{\sqrt{n}}$ is the **standard error** of $\overline{y}$; our best estimate of $\frac{\sigma}{\sqrt{n}}$.
- What is the difference between $s$ and $\frac{s}{\sqrt{n}}$?
  - The sample standard deviation, $s$, says how far **observations** are from the sample average
  - The standard error, $\frac{s}{\sqrt{n}}$, says how far **sample averages** are from the population average

# Statistics review: Hypothesis testing about the mean from a Normal distribution

- With a sample mean, we can use hypothesis testing to answer "yes" or "no" questions related to that estimate, e.g. is there any evidence that the population mean $\mu$ is equal to $\mu_0 = 0.5$?

$$\mathrm{H}_0 : \mu = \mu_0$$

  where $\mu_0$ is a value we specify. In the majority of applications, $\mu_0 = 0$.
- A typical alternative hypothesis is

$$\mathrm{H}_1 : \mu \neq \mu_0.$$

- Intuitively, we reject $\mathrm{H}_0$ in favor of the $\mathrm{H}_1$ if the value of the sample average, $\overline{y}$, is "sufficiently" different than $\mu_0$.

# Statistics review: Hypothesis testing about the mean from a Normal distribution

- Test statistic $t$ is given by:

$$t = \frac{(\overline{y} - \mu_0)}{se(\overline{y})}$$

- Rejection rule is

$$|t| > c$$

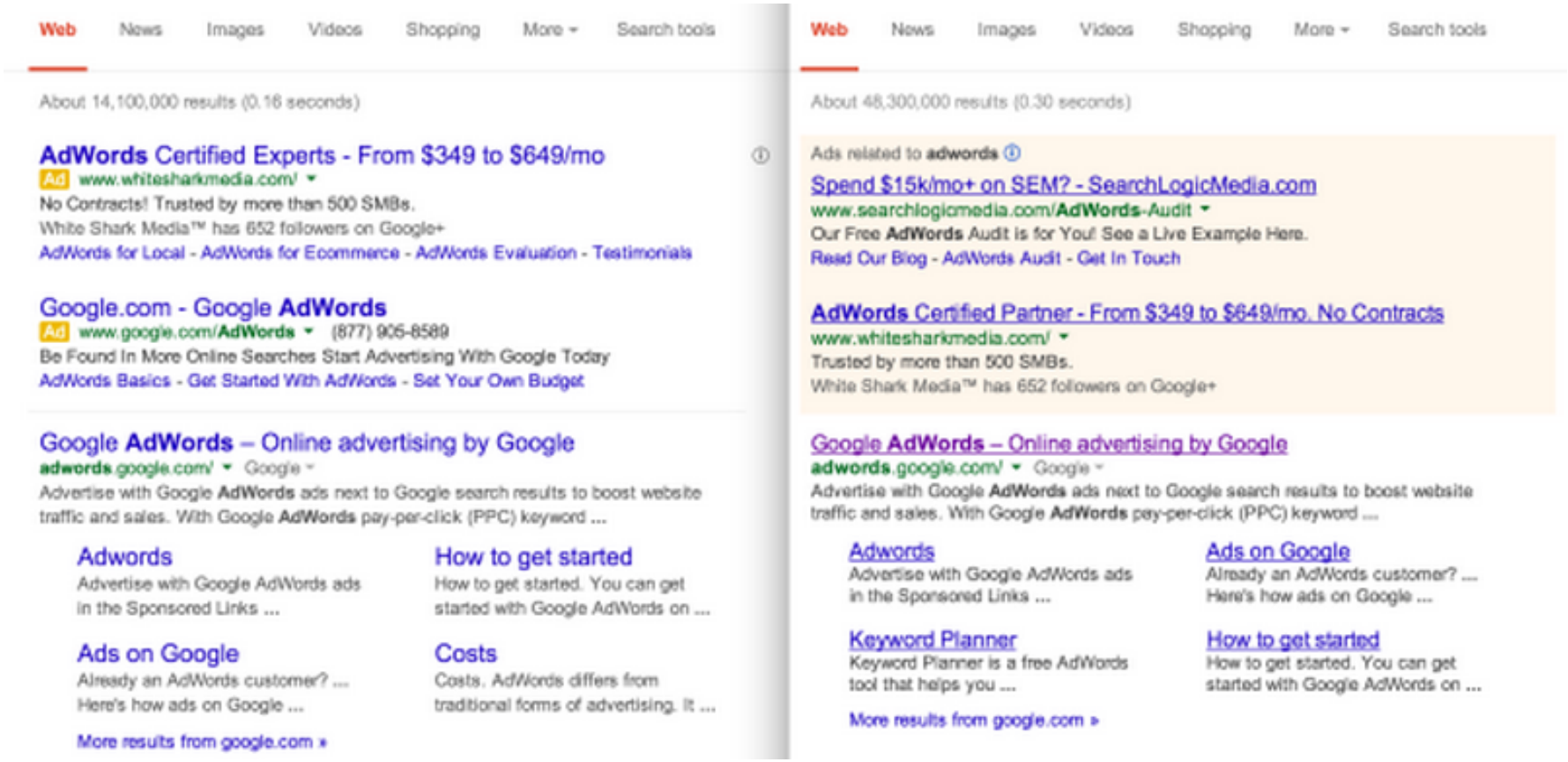
  where $c$ is chosen so that probability of a **Type I error**, rejecting the $H_0$ when it is true, is a certain size (usually 5%).
- Idea: if $\overline{y}$ and $\mu_0$ are very different then $|t|$ :
  - will be large,
  - will be more likely to exceed the chosen cutoff $c$, and
  - will be evidence that the null hypothesis is not true.

# Statistics review: Example (based on Wooldridge, p. 793)

- One setting in which experimental data is more likely is in the tech world where companies are able to perform A/B testing
- E.g. Google claims that a new look search results page has increased ad clicks. Test this claim.

| User | Ad clicks before | Ad clicks after |
|---|---|---|
| 1 | 8.30 | 9.25 |
| 2 | 9.40 | 9.00 |
| 3 | 9.00 | 9.25 |
| 4 | 10.50 | 10.00 |
| 5 | 11.40 | 12.00 |
| 6 | 8.75 | 9.50 |
| 7 | 10.00 | 10.25 |
| 8 | 9.50 | 9.50 |
| 9 | 10.80 | 11.50 |
| 10 | 12.55 | 13.10 |

# Statistics review: Example (based on Wooldridge, p. 793)



# Statistics review: Example (based on Wooldridge, p. 793)

- Difference is $D_i \equiv C_i^a - C_i^b$. Assume that the $D_i$ are a random sample from the Normal $N(\mu, \sigma^2)$ distribution.

| $i$ | $c_i^b$ | $c_i^a$ | $d_i$ |
|---|---|---|---|
| 1 | 8.30 | 9.25 | 0.95 |
| 2 | 9.40 | 9.00 | -0.40 |
| 3 | 9.00 | 9.25 | 0.25 |
| 4 | 10.50 | 10.00 | -0.5 |
| 5 | 11.40 | 12.00 | 0.60 |
| 6 | 8.75 | 9.50 | 0.75 |
| 7 | 10.00 | 10.25 | 0.25 |
| 8 | 9.50 | 9.50 | 0.0 |
| 9 | 10.80 | 11.50 | 0.70 |
| 10 | 12.55 | 13.10 | 0.55 |

# Statistics review: Example (based on Wooldridge, p. 793)

- The sample average increase in clicks is

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{10}\sum_{i=1}^{10} d_i = \frac{3.15}{10} = 0.315.$$

- The sample variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2 = \frac{2.16025}{9} = 0.240.$$

- The sample standard deviation is $\sqrt{s^2} = s = 0.4899$ and the standard error of $\overline{d}$ is $se(\overline{d}) = \frac{s}{\sqrt{n}} = \frac{0.4899}{\sqrt{10}} = 0.1549.$

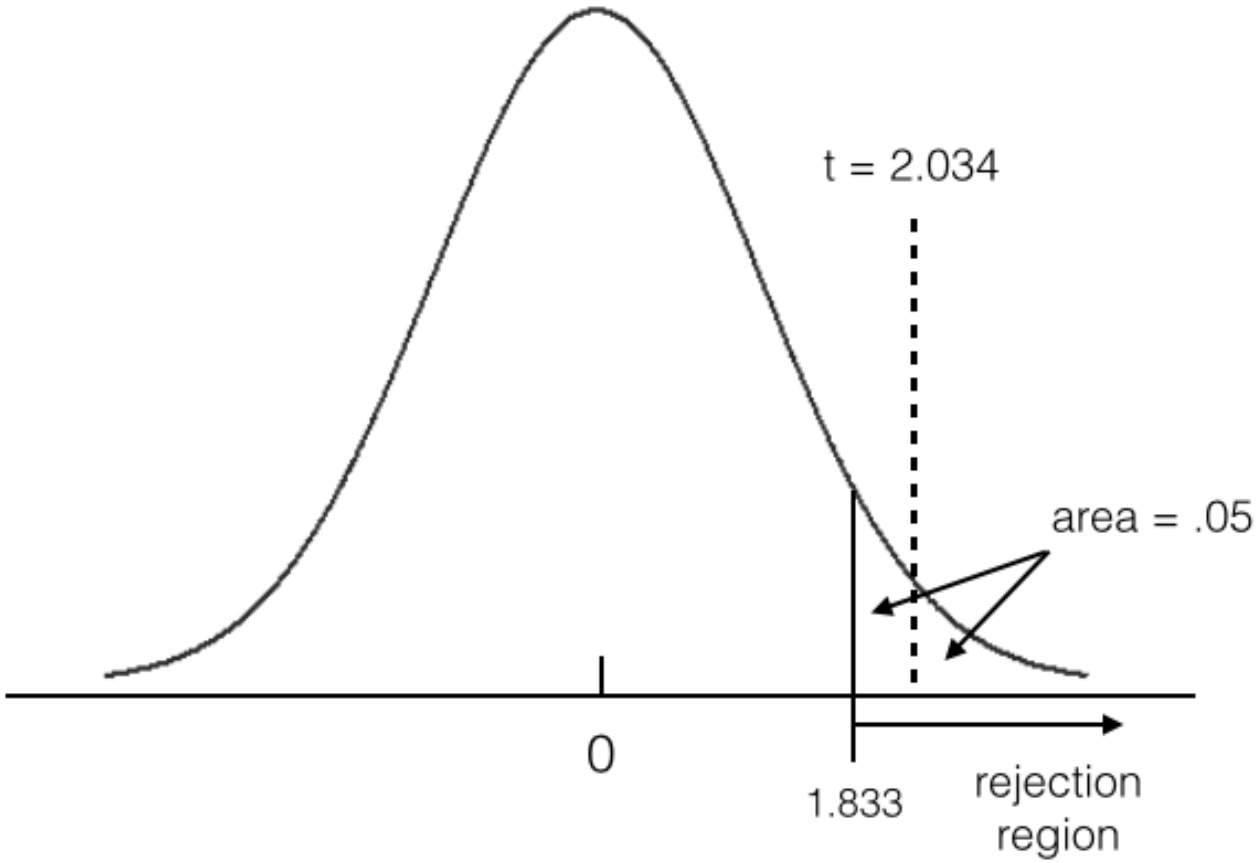
# Statistics review: Example (based on Wooldridge, p. 793)

- Formal hypothesis test:

$$\mathrm{H}_0 : \mu = 0$$
$$\mathrm{H}_1 : \mu > 0.$$

- $t = \frac{(\bar{d}-0)}{se(\bar{d})} = \frac{0.315}{0.1549} = 2.03357$. The critical value for a one-tailed test at the 5% level is 1.833.
- Since $|t| = 2.03357 > 1.833 = c$, we can reject the null hypothesis that clicks have remained the same at the 5% level.

# Statistics review: Example (based on Wooldridge, p. 793)



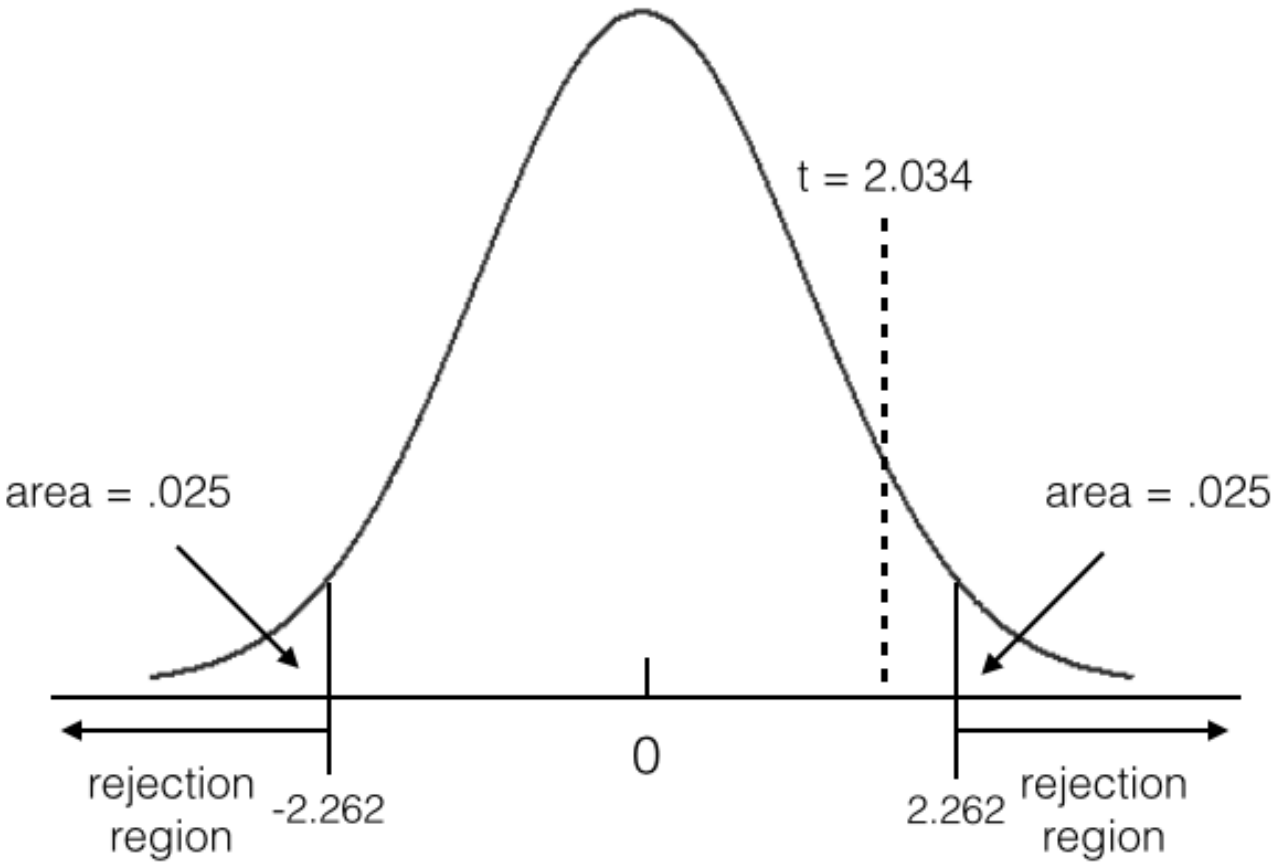
# Statistics review: Example (based on Wooldridge, p. 793)

- If Google was agnostic about the direction that the new look page affected clicks then we would need a two-tailed test.
- Formal hypothesis test:

$$\begin{aligned} \mathrm{H}_0 &: \mu = 0 \\ \mathrm{H}_1 &: \mu \neq 0. \end{aligned}$$

- $t = \frac{(\bar{d}-0)}{se(\bar{d})} = \frac{0.315}{0.1549} = 2.03357$. The critical value for a two-tailed test at the 5% level is 2.262.
- Since $|t| < c$, we fail to reject the null hypothesis that clicks have remained the same at the 5% level.
- This is consistent with the fact that the 95% CI includes zero
  - The 97.5th percentile in the $t_{n-1} = t_9$ distribution is 2.262 so the 95% CI is $[-0.035, 0.665]$.

# Statistics review: Example (based on Wooldridge, p. 793)



# Statistics review: Covariance

- The sample covariance is defined as:

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}).$$

  where the division by $n-1$ rather than $n$ again accounts from the fact that the mean $\mu$ is estimated rather than known.
- Positive covariance indicates that the two variables move in the same direction.
- If $X$ and $Y$ are independent then (the population covariance) $Cov(X, Y) = \sigma_{XY} = 0$.
- $Cov(a_1X + b_1, a_2Y + b_2) = a_1a_2Cov(X, Y)$
  - Covariance can be altered simply by multiplying one or both random variables by a constant (e.g. $a_1$, $a_2$), which creates a problem if different units of measurement are used!

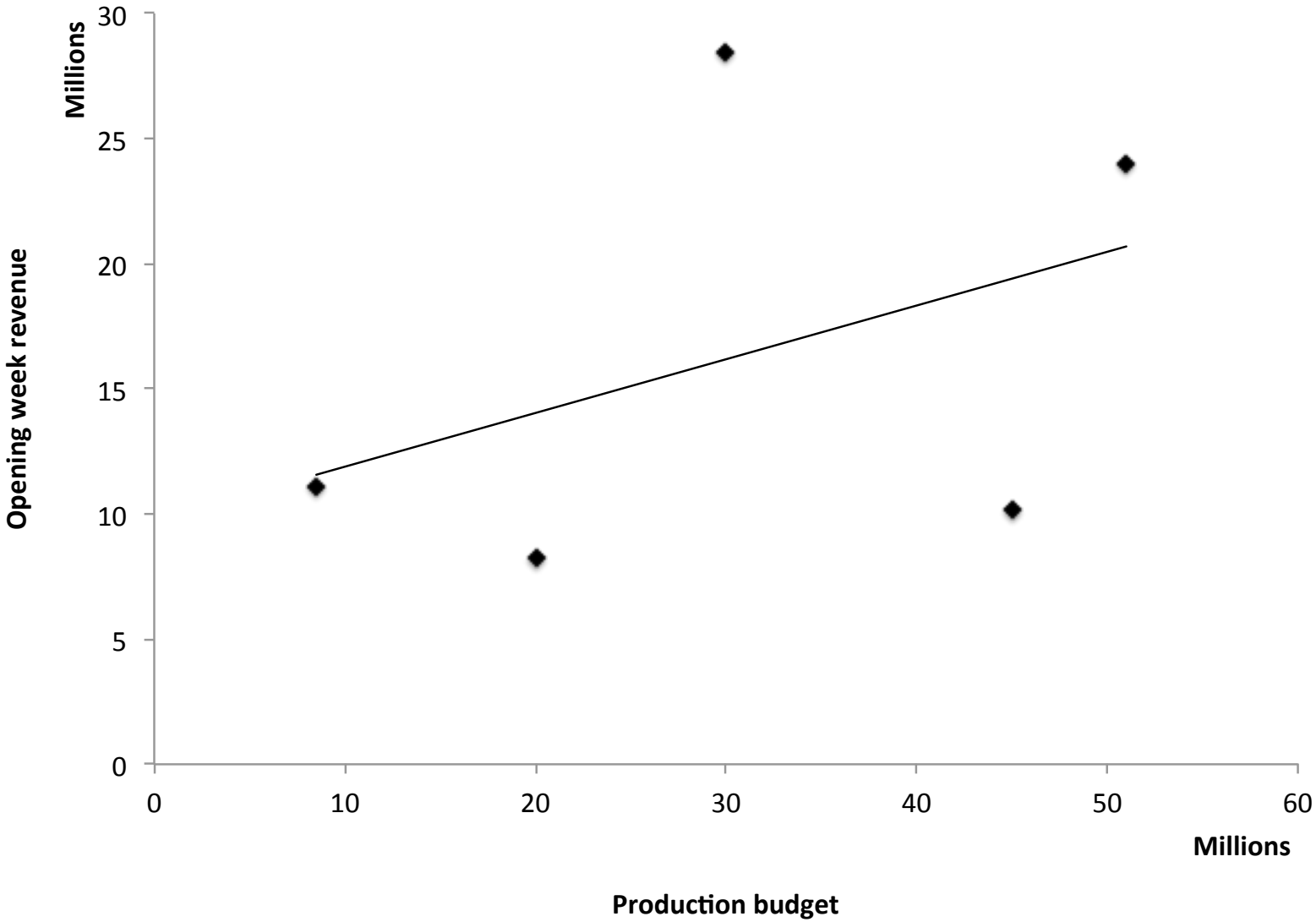
# Statistics review: Correlation

- The sample correlation coefficient (or sample correlation):

$$R_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^{\frac{1}{2}}\left(\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right)^{\frac{1}{2}}}$$

- Varies from $-1$ to 1.
- Not dependent on units of measurement.
  - Population correlation $Corr(a_1 X + b_1, a_2 Y + b_2) = Corr(X, Y)$

# Statistics review: Covariance and correlation



# Statistics review: Covariance and correlation

| title | rev (y) | bdgt (x) | y-(Σy/n) | x-(Σx/n) | [y-(Σy/n)] * [x-(Σx/n)] | [y-(Σy/n)]^2 | [x-(Σx/n)]^2 |
|---|---|---|---|---|---|---|---|
| bangkokdangerous | 10.13 | 45 | -6.25 | 14.10 | -88.06 | 39.01 | 198.81 |
| bankjob | 8.20 | 20 | -8.17 | -10.90 | 89.09 | 66.81 | 118.81 |
| barbershop2 | 28.47 | 30 | 12.10 | -0.90 | -10.89 | 146.33 | 0.81 |
| barnyard | 24.02 | 51 | 7.64 | 20.10 | 153.55 | 58.36 | 404.01 |
| baggageclaim | 11.06 | 8.5 | -5.32 | -22.40 | 119.09 | 28.27 | 501.76 |
| | | | | | 262.79 | 338.78 | 1224.20 |
| **Σ** | 81.88 | 154.50 | | | | | |
| **Σ/n** | 16.38 | 30.90 | | | | | |
| **Sxy** | 65.70 | | | | | | |
| **Rxy** | 0.41 | | | | | | |

# Next week

- Simple regression model (Chapter 2)