BUEC-333
Statistical analysis of economic data

Summer 2016

Table of contents

Section 1

Administration

About me

- Chris Muris
  - office: WMC 3639
  - email: Come and see me in person!
  - www: chrismuris.github.io
  - oo: Thursday, 10:00-12:00

Course website

- Location: through chrismuris.github.io/buec333
- Contains all you need to know
  - syllabus
  - detailed readings/schedule
  - additional materials (slides, code, data)
  - problem sets (for tutorials, highly recommended)
- WARNING: old exams

The strange!

1. I don’t do email!
2. Tutorials are blackboard-free!
3. Two tests, no final (2x30%=60%)

Lectures

Goal: To help you read Stock and Watson.

- Review non-self-study material (90 minutes)
- Solve in-class exercises (60 minutes)
- R (30 minutes)

Tutorials

- Weekly **problem sets**
  - subject of the **tutorials**
  - strongly **indicative** of **midterm+exam** questions
  - not obligatory

Lab

- Get help with learning **R**
- Two **obligatory** hand-in assignments (2x20%=40%)
  - part theoretical, part practical (R)
  - dates: see course website
- Time and date: TBA

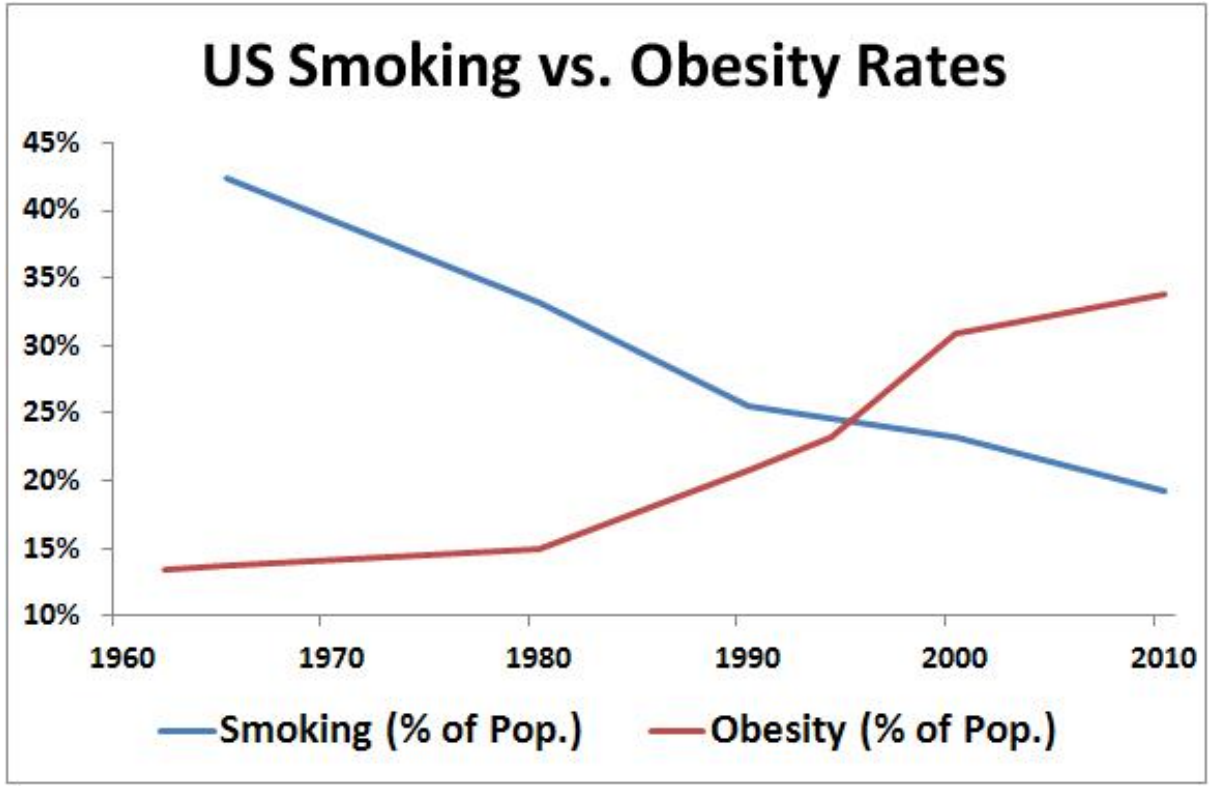
Section 2

Why take this course?

What will you learn?

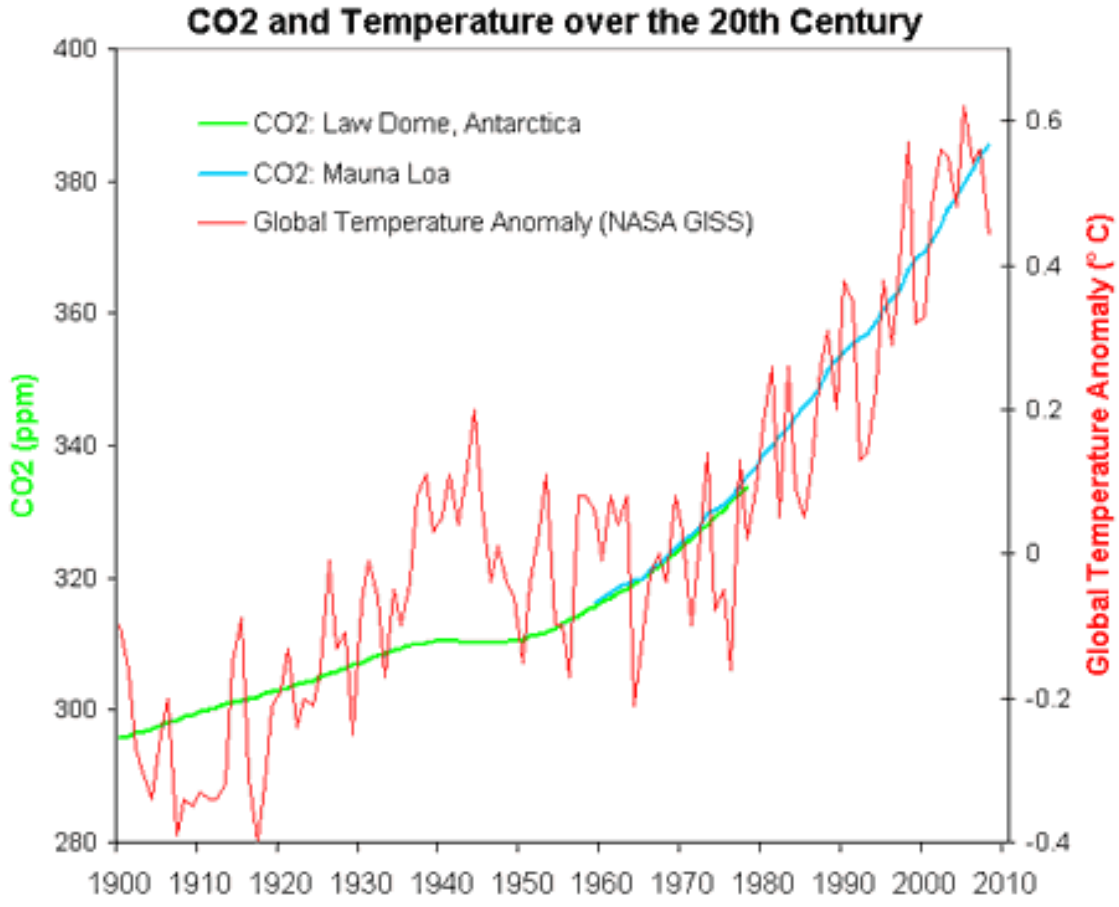
- Regression analysis
- **Quantifying** relationship between variables using **data**

SP 1



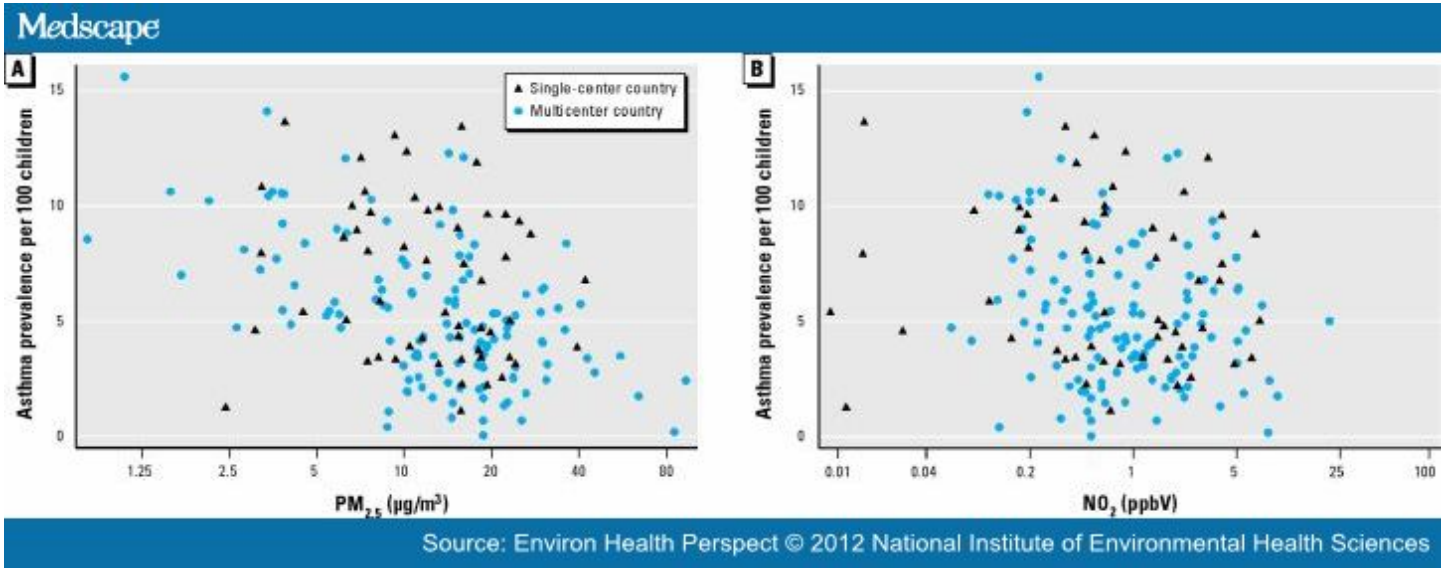
Are they related? How do we bend back the rise in obesity? [Source: Colorado Health Insurance Brokers]

SP 2



Did increasing carbon dioxide concentrations lead to global warming? How much warmer will it get? [Source: Skeptical Science]

SP 3



Does increased particulate matter affect health? How much should we invest in cleaner cars? [Source: MedScape]

What will you learn? (cont'd)

- Review probability and statistics
- **Regression analysis**
  - Understand it (lectures, tutorials)
  - Use it (computer lab)
  - Apply it to real-world questions (assignments)
- Extensions of regression analysis
- Correlation is not causation

Why take this course?

1. Increasing number of data/stats-related **jobs** that require statistical literacy
2. Regression analysis is the most **important tool** for quantitative research
3. Intellectual **curiosity**
4. ...

Reason 1: Increase in data use in industry

- Lots of new jobs require knowledge of statistics/data
- People in non-data jobs are increasingly encountering data
- NYTimes, February 2012, click here

Reason 1 (cont'd)

February 11, 2012

The Age of Big Data

By STEVE LOHR

GOOD with numbers? Fascinated by data? The sound you hear is opportunity knocking.

Mo Zhou was snapped up by I.B.M. last summer, as a freshly minted Yale M.B.A., to join the technology company's fast-growing ranks of data consultants. They help businesses make sense of an explosion of data — Web traffic and social network comments, as well as software and sensors that monitor shipments, suppliers and customers — to guide decisions, trim costs and lift sales. "I've always had a love of numbers," says Ms. Zhou, whose job as a data analyst suits her skills.

To exploit the data flood, America will need many more like her. A report last year by the McKinsey Global Institute, the research arm of the consulting firm, projected that the United States needs 140,000 to 190,000 more workers with "deep analytical" expertise and 1.5 million more data-literate managers, whether retrained or hired.

The impact of data abundance extends well beyond business. Justin Grimmer, for example, is one of the new breed of political scientists. A 28-year-old assistant professor at Stanford, he combined math with political science in his undergraduate and graduate studies, seeing "an

Reason 1 (cont'd)

- According to projections by McKinsey Global:

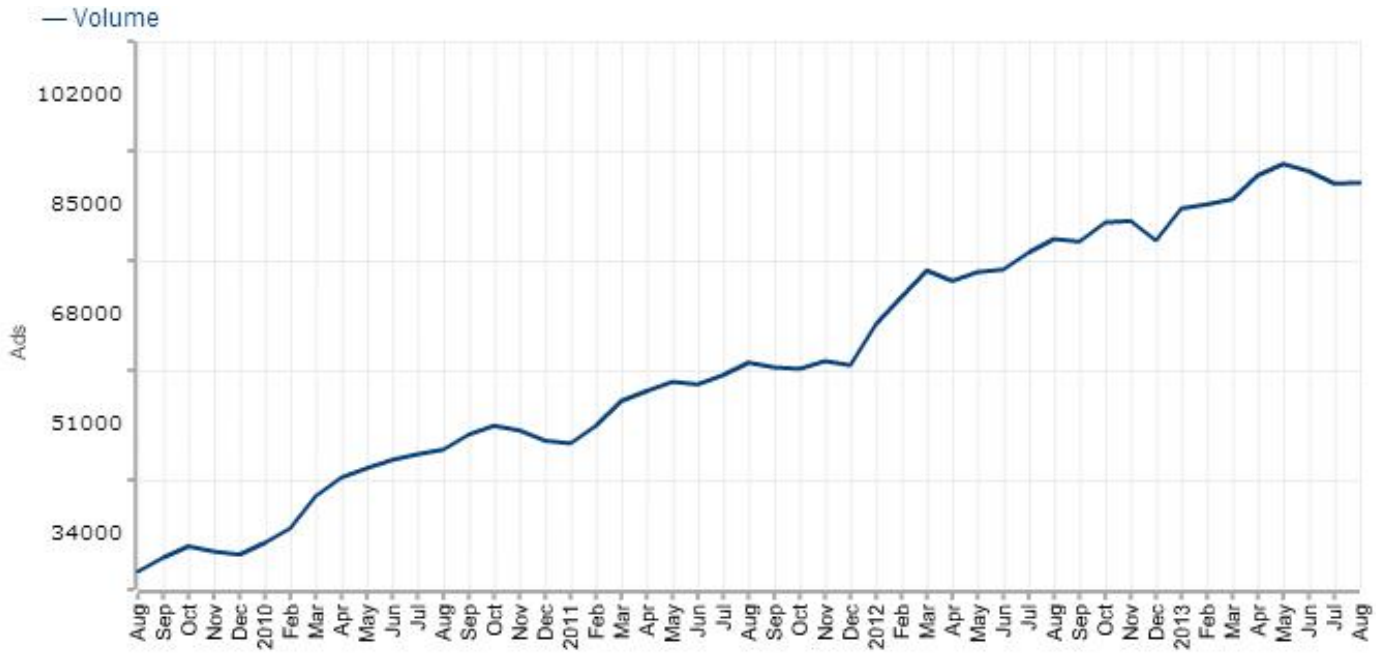
  "The U.S. needs

  - **140.000 to 190.000** more workers with "deep analytical" expertise, and
  - **1.5 million** more data-literate managers, whether retrained or hired." Source: NYTimes article.

Reason 1 (cont'd)

- An update from wantedAnalytics
- "In September [2013], the number of jobs that require data analytics skills sets increased 13% year-over-year compared to September 2012."
  - health and medical insurance carriers;
  - colleges, universities, and professional schools;
  - business support services;
  - computer systems design services;
  - and management consulting services

Reason 1 (cont'd)



Reason 2: Ability to answer interesting questions

- **Learn** about the most important tool for quantitative research
- **Ability** to quantify relationships
- **Understand** and **criticize** statistical results
  - Economics in the news
  - Election polls
  - Medical / health research
- See upcoming example, and the rest of the semester

Reason 2: Interesting/important questions

- What is the effect of a **texting ban** on highway **fatalities**? [info works]
- What is the effect of reducing **class size** on **student achievement**? [coming up]
- What is the effect of taking another year of **education** on your future **wages**? [5-10%]
- What is the best **place** to **open** a new **restaurant**?
- Can we use data from **Twitter** to predict **stock** market indicators? [yes]
- Can we use **weather data** to predict **wine prices**? [better than "experts"]

[Sources: various, including Stock and Watson, Chapter 4]

Reason 3: Intellectual/mathematical curiosity

- What is the **best** line you can draw through a **cloud** of points?
- How close is this estimated line to the truth?
- Can we uncover the truth if we obtain a large enough data set?
- How do I tell the difference between correlation and causality?

Reason 4

- It is a required course.

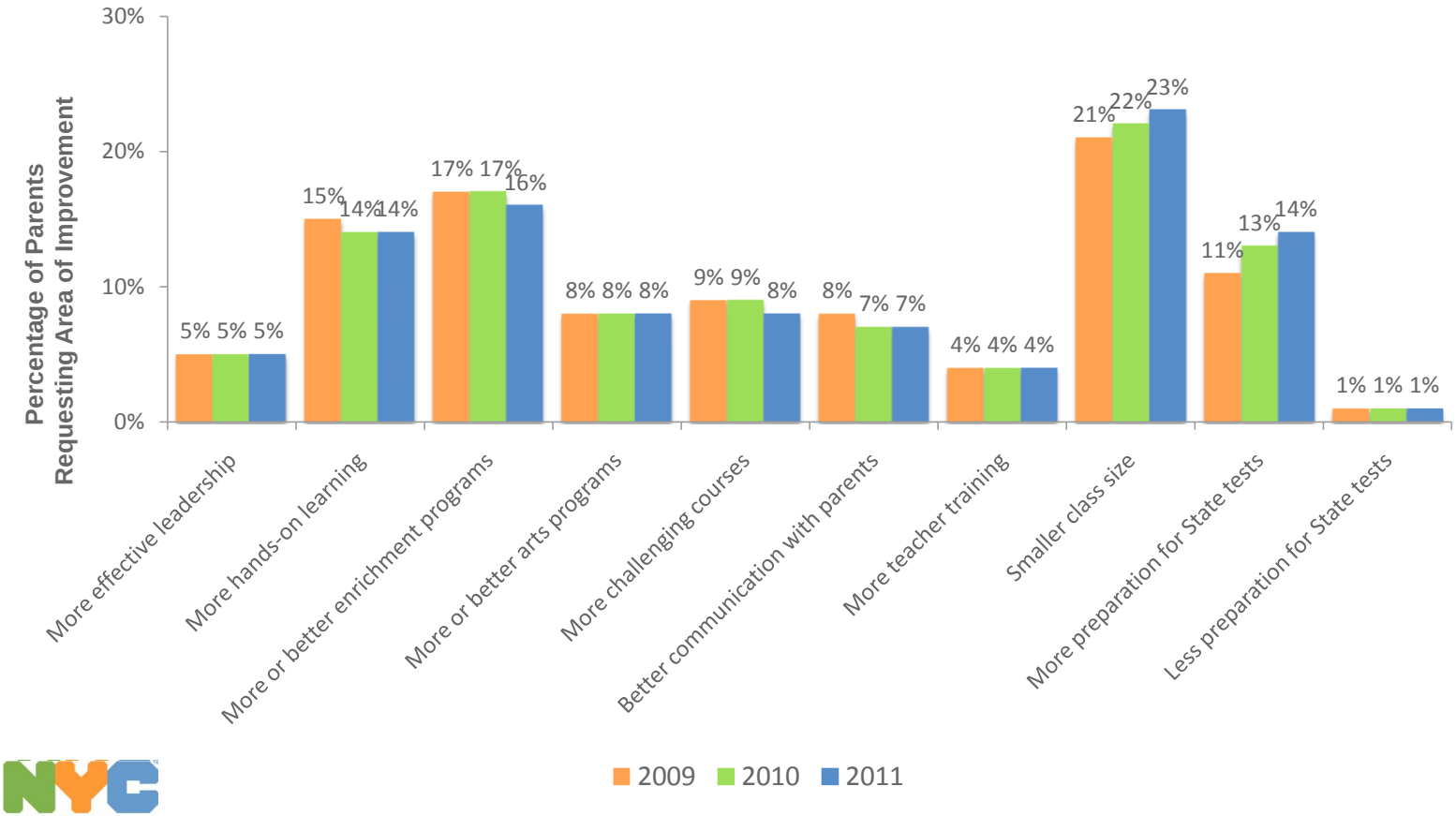
Class size and student performance: Debate

- There is ongoing, heated debate about school reforms
- Especially elementary schools

Debate (cont'd).

Parents want smaller class sizes. Source: NYC Dep of Education

Parent requests for school improvements remained consistent



Department of Education
Dennis M. Walcott, Chancellor

• *District 75 schools are not included here.*

Debate (cont'd).

Making classes smaller costs money. Source: NYT

March 6, 2011

Tight Budgets Mean Squeeze in Classrooms

By SAM DILLON

Millions of public school students across the nation are seeing their class sizes swell because of budget cuts and teacher layoffs, undermining a decades-long push by parents, administrators and policy makers to shrink class sizes.

Over the past two years, California, Georgia, Nevada, Ohio, Utah and Wisconsin have loosened legal restrictions on class size. And Idaho and Texas are debating whether to fit more students in classrooms.

Los Angeles has increased the average size of its ninth-grade English and math classes to 34 from 20. Eleventh- and 12th-grade classes in those two subjects have risen, on average, to 43 students.

"Because many states are facing serious budget gaps, we'll see more increases this fall," said

Reasoning

- Are smaller classes better?
- Smaller classes cost money!
- If you are interested in the academic discussion: http://nber.org/papers/w17632/, http://www.aypf.org/publications/rmaa/pdfs/ClassSizeSTAR.pdf

Questions

There are at least two different types of questions we can answer using statistical methods:

1. Do smaller class sizes lead to higher student performance? **(yes/no)**, testing
2. How does student performance change with class size? **(best guess)**, estimation
3. What is the **optimal** class size for student performance?

A priori reasoning

- I think it could go either way on the yes/no question:
  - In really large classes, there is no teacher-student interaction
  - In very small classes (1 student), students do not learn from each other

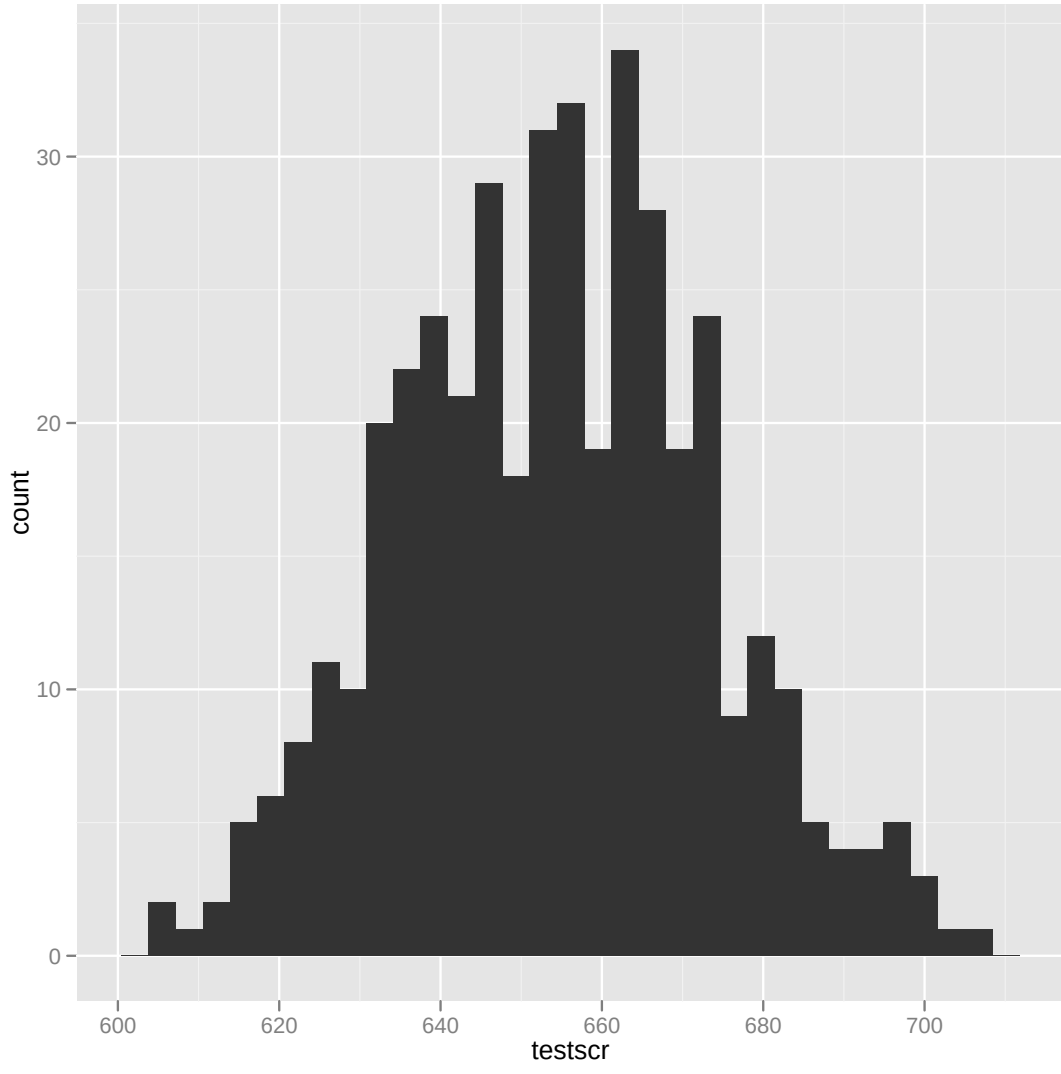
Empirical research

- Rather than theorizing what the effect may be...
- ...we can get data about school size and performance and use regression analysis
- We will use data from the STAR data in California.

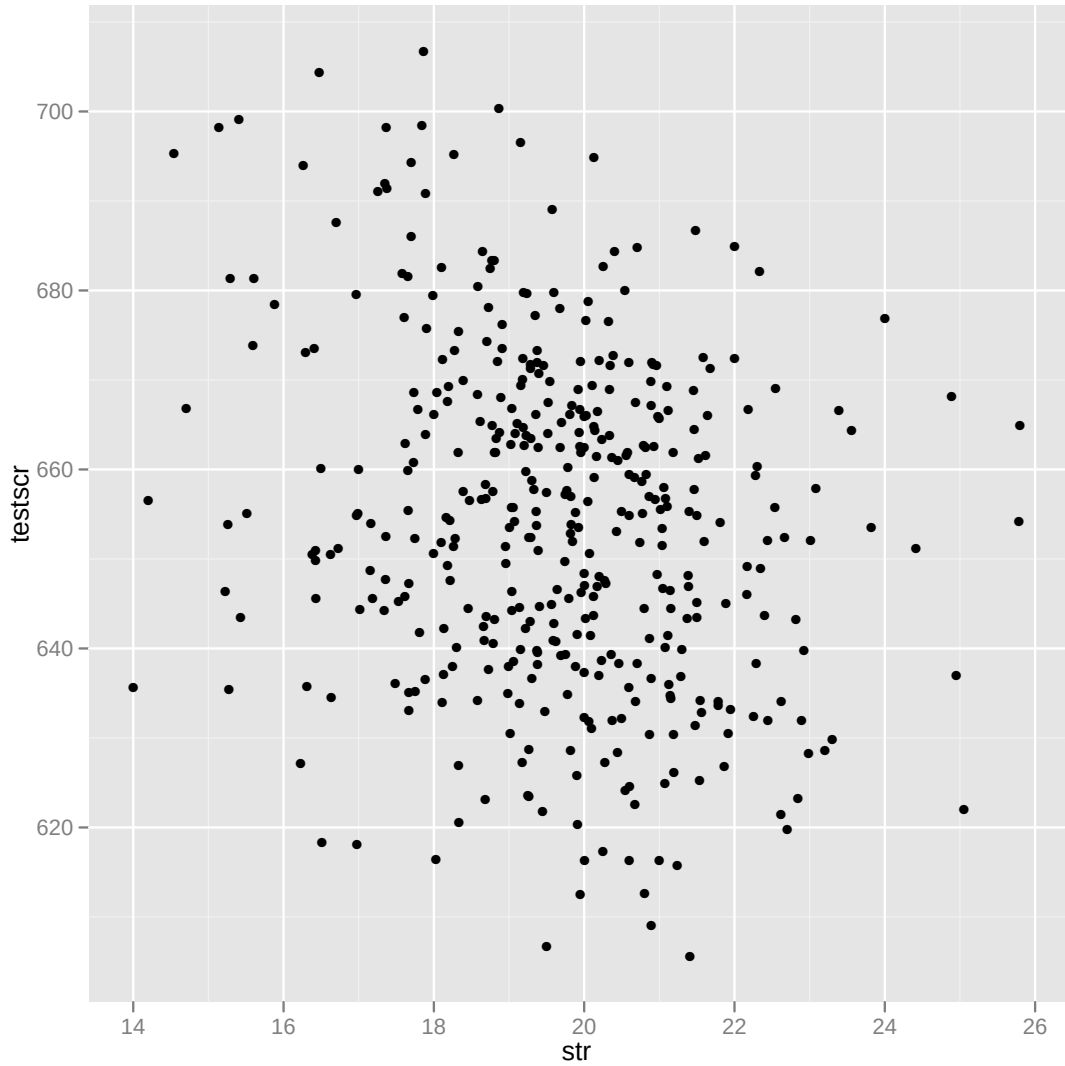
STAR experiment

- Two years of data: 1998-1999
- All 420 elementary school districts in California
- *testscores*: average of a reading and math score (5th grade)
- *str*: student-to-teacher ratio

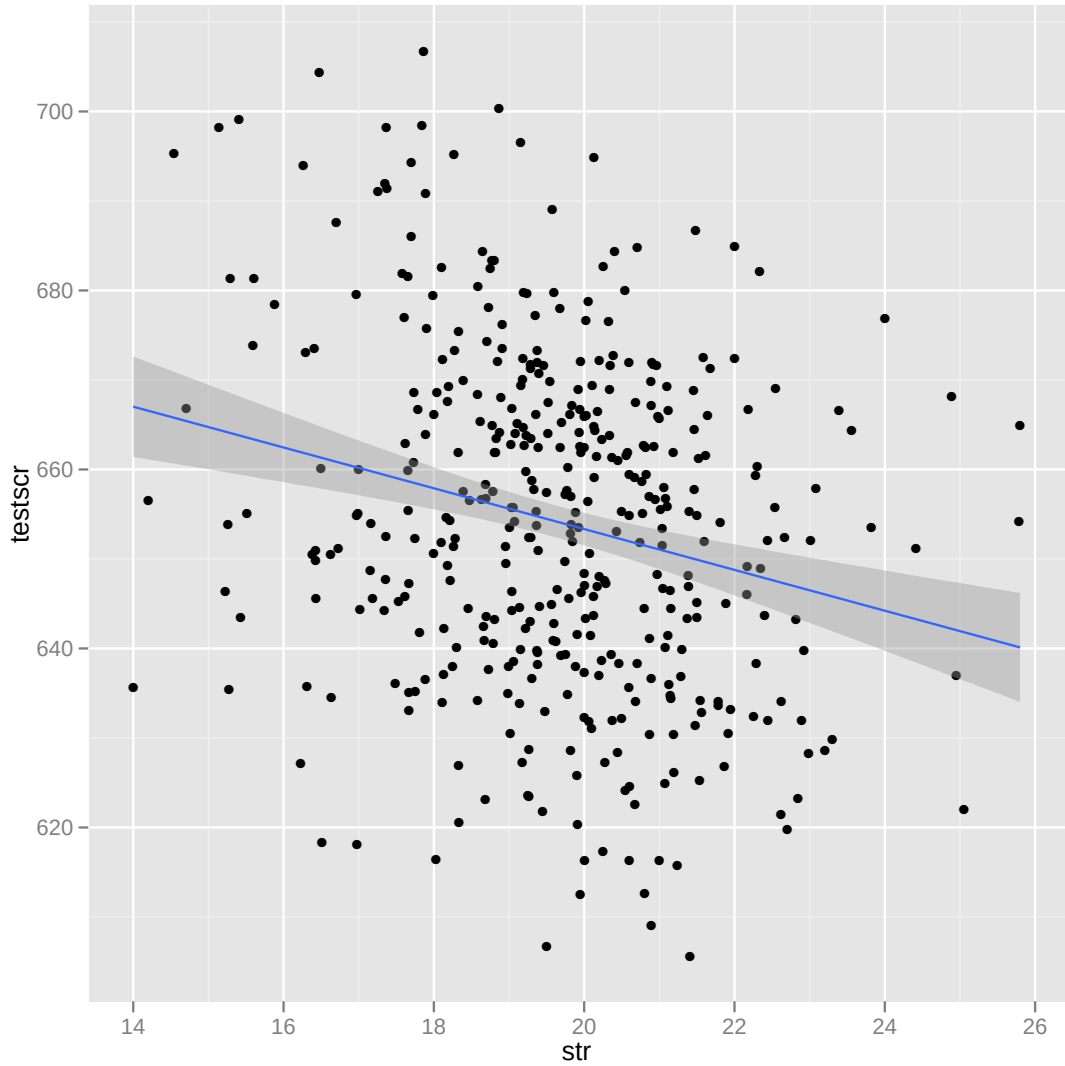
Data plot: testscr



Data plot: scatter plot



Data plot: fitted line



Remaining questions

1. How did I get that line? (Part 1)
2. Other factors may play a role: how do I include them? (P2)
3. How do I know that they do not move together by chance? (P2)
4. How do I know that the relationship is not due to some other variable? (P2)
5. How do I find a school that will give my child a good performance? (P1/2)
6. What if the relationship is more complicated? (P2)

Section 3

Review: Probability theory

Probability and statistics

- First: probability; then: statistics
- The theory of probability gives you the tools to **describe** things around you that are random
- Statistics is about **learning** about these things around you by using data

Overview of probability review

1. Random variables **- self-study**
2. Expected value, variance **- self-study**
3. Two random variables: joint, conditional, and marginal distributions
4. Continuous random variables, and special distributions

Random variables

- A **random variable** (RV) is a **numerical** measure of a random outcome
- It describes an event of which you do not know (yet) know the outcome
- Examples:
  - coin toss / die throw
  - tomorrow's temperature
  - bus waiting time (w/o Translink App)
  - winner World Cup 2020
  - height of **randomly chosen** SFU student

Concepts

- A possible realization of a RV X is called an **outcome**
- The **sample space** is the set of all possible outcomes
- The **probability** of an outcome is the proportion of the time the outcome occurs in the long run
- The **probability distribution** lists outcomes in sample space and their probabilities
- An **event** is a subset of the sample space