

Applied Statistics and Econometrics

Lecture 1

GIUSEPPE Ragusa

Luiss University

`gragusa@luiss.it`

`http://gragusa.org/`

February 14, 2017

Luiss University

- The use of mathematical and statistical methods:
 - to verify economic theories
 - to fit economic models to real data
 - to forecast future values of economic quantities
- Econometric techniques are also used by: sociologists, political scientists and other social scientists.

Why should you study econometrics?

Policy evaluation

Assist in evaluating effects of policies both before and after implementation

Financial markets

Forecasting, CAPM, APT, etc.

Strategic management

Inventory management and firm performances, analysis of divestiture, etc.

Marketing

Demand functions for industries, study of consumer behavior, etc.

Macroeconomics

Models and business cycles, models of the monetary policy, growth forecast, etc.

Industrial organization

Price discrimination theories, estimation of market power, etc.

Why studying econometrics?

*One hurdle is a talent and skills gap. The United States alone, McKinsey projects, will need **140,000** to **190,000** more people with “deep analytical” skills, typically experts in statistical methods and data-analysis technologies.*

McKinsey says the nation will also need 1.5 million more data-literate managers [...]. [...] the need for a sweeping change in business to adapt a new way of managing and making decisions that relies more on data analysis.

Source: McKinsey (2011), *Big data: The next frontier for innovation, competition, and productivity*

Question: how much does cigarettes consumption respond to a change in price?

Question: how much does cigarettes consumption respond to a change in price?

- We know from consumer theory that price elasticities are negative

$$\epsilon = \frac{dq}{dp} \frac{p}{q} < 0$$

Question: how much does cigarettes consumption respond to a change in price?

- We know from consumer theory that price elasticities are negative

$$\epsilon = \frac{dq}{dp} \frac{p}{q} < 0$$

- but theory does not tell us its exact value

Question: how much does cigarettes consumption respond to a change in price?

- We know from consumer theory that price elasticities are negative

$$\epsilon = \frac{dq}{dp} \frac{p}{q} < 0$$

- but theory does not tell us its exact value
- yet, from a policy point of view it is fundamental to know the exact magnitude of the elasticity.

Testing the validity of various versions of CAPMs attracts a lot attention in empirical finance. Various testing procedures and statistical methods have been proposed and studied.

CAPM

Fama and French (1993) form a three factor model to explain the expected excessive returns of assets. Broadly speaking, the three factors are:

- market index
- value equity of firms
- book-to-market value

Other examples

- What is the quantitative effect of reducing class size on student achievement?
- How does another year of education change earnings?
- Are “better” performing CEO payed more?
- What is the effect on output growth of a 1 percentage point increase in interest rates by the European Central Bank?
- What is the effect on housing prices on the environment?

In this course you will:

- Learn methods for estimating causal effects using observational data
- Learn some tools that can be used for other purposes, for example forecasting using time series data;
- Focus on applications - theory is used only as needed to understand the “why”s of the methods;
- Learn to evaluate the regression analysis of others - this means you will be able to read/understand empirical economics papers in other econ courses;
- Get some hands-on experience with regression analysis in your problem sets.

This course is “mostly” about using data to measure causal effects.

- Ideally, we would like an experiment
 - what would be an experiment to estimate the effect of class size on standardized test scores?
- But almost always we only have observational (nonexperimental) data.
 - returns to education
 - cigarette prices
 - monetary policy
- Most of the course deals with difficulties arising from using observational to estimate causal effects
 - confounding effects (omitted factors)
 - simultaneous causality
 - **correlation does not imply causation**

All you need is “data” ...



Figure 1: The cover of The Economist, November 2010

RStudio

```

1 set.seed(100)
2
3 n_max <- 500
4 j <- 10000
5
6 Xt <- matrix(0, n_max, j)
7 St <- matrix(0, n_max, j)
8
9 tmp <- (rchisq(n_max, df = 2)-2)/sqrt(4)
10 Xt[,1] <- cumsum(tmp) / seq_along(Xt[,1])
11 St[,1] <- cumsum((tmp-Xt[,1])^2) / (seq_along(Xt[,1])-1)
12 plot(Xt[2:n_max,1], ylim = c(-4, 4), type = "l", lwd = 0.1, col = "grey", ylab = "Xbar", xlab = "t")
13
14
15 for (j in 1:j) {
16   tmp <- (rchisq(n_max, df = 2)-2)/sqrt(4)
17   Xt[,j] <- cumsum(tmp) / seq_along(Xt[,j])
18   St[,j] <- cumsum((tmp-Xt[,j])^2) / (seq_along(Xt[,j])-1)
19   lines(Xt[2:n_max,j], lwd = 0.1, col = "grey")
20 }
21 abline(h = 0, lwd = 2)
22 grid()
23
24 hist(sqrt(n_max)*Xt[n_max,], prob = TRUE, add = FALSE, col = "slategray", breaks = "Scott")
25 curve(dnorm, from = -4, to = 4, lwd = 3, col = "darkblue", add = TRUE)
26
27
28 hist(sqrt(25)*Xt[25,], prob = TRUE, add = FALSE, col = "slategray", breaks = "Scott", main = "",
29 curve(dnorm, from = -4, to = 4, lwd = 3, col = "darkblue", add = TRUE)
30 grid(col = "black")
31 hist(sqrt(25)*Xt[25,]/sqrt(St[25,]), prob = TRUE, add = FALSE, col = "slategray", breaks = "Scott",
32 curve(dnorm, from = -4, to = 4, lwd = 3, col = "darkblue", add = TRUE)
33
34
35 hist(sqrt(500)*Xt[500,]/sqrt(St[500,]), prob = TRUE, add = FALSE, col = "slategray", breaks = "Scott",
36 curve(dnorm, from = -4, to = 4, lwd = 3, col = "darkblue", add = TRUE)
37
38
39

```

output file: Lecture1.tex

```

[1] "Lecture1.tex"
> hist(sqrt(25)*Xt[25,], prob = TRUE, add = FALSE, col = "slategray", breaks = "Scott")
> curve(dnorm, from = -4, to = 4, lwd = 3, col = "darkblue", add = TRUE)
> hist(sqrt(25)*Xt[25,], prob = TRUE, add = FALSE, col = "slategray", breaks = "Scott", main = "")
> curve(dnorm, from = -4, to = 4, lwd = 3, col = "darkblue", add = TRUE)
> hist(sqrt(25)*Xt[25,]/sqrt(St[25,]), prob = TRUE, add = FALSE, col = "slategray", breaks = "Scott", main = "", xlab = "")
> curve(dnorm, from = -4, to = 4, lwd = 3, col = "darkblue", add = TRUE)
> grid()
> grid(col = "black")

```

Files Plots Packages Help Viewer

Zoom Export Publish

3.1.1 (Top Level) | R Script 1

Environment History Presentation

We will use R as a statistical language.

- <http://r-project.org>, the web site of the R project
- <http://rstudio.org>, the web site of the Rstudio IDE
- <http://github.com/gragusa/ase>, our own package

Types of data: Time Series, Cross Section, Panel Data

Time series

data for a **single entity** (person, firm, country) collected at multiple time periods

Cross section

data on different entity entity (workers, consumers, firms, etc.) collected at a single time period

Panel data

data for multiple entities in which each entry is observed at two or more time periods.

Time series

```
##          SP500
## 2014-11-28 2067.56
## 2014-12-31 2058.90
## 2015-01-30 1994.99
## 2015-02-27 2104.50
## 2015-03-31 2067.89
## 2015-04-30 2085.51
## 2015-05-29 2107.39
## 2015-06-30 2063.11
## 2015-07-31 2103.84
## 2015-08-31 1972.18
## 2015-09-30 1920.03
## 2015-10-30 2079.36
## 2015-11-30 2080.41
## 2015-12-31 2043.94
## 2016-01-29 1940.24
## 2016-02-19 1917.78
```

Cross section

```
salary  pcsalary  sales      roe        pcroe      ros        indus      finance
consprod utility  lsalary    lsales
```

Obs: 208

1. salary 1990 salary, thousands \$
2. pcsalary % change salary, 89-90
3. sales 1990 firm sales, millions \$
4. roe return on equity, 88-90 avg
5. pcroe % change roe, 88-90
6. ros return on firm's stock, 88-90
7. indus =1 if industrial firm
8. finance =1 if financial firm
9. consprod =1 if consumer product firm
10. utility =1 if transport. or utilities

Dataset in R

```
data(ceo)
head(ceo, 4)
```

```
##      salary pcsalary      sales      roe      pcroe  ros  indus  finance  consprod
## 1    1001      32  9958.000  10.900010 -30.60002  13    1      0      0
## 2    1122      9  6125.905  23.500000 -16.30001  14    1      0      0
## 3     578     -9 16246.000   5.900005 -25.70002 -21    1      0      0
## 4    1368      7 21783.220  13.800010  -3.00000  56    1      0      0
##      utility
## 1         0
## 2         0
## 3         0
## 4         0
```

Cross section

dataset: CASchools

district school county grades students teachers calworks
lunch computer expenditure income english read math testscore

Obs: 420

1 school: School name
2 county: County name
3 grades: Grade span of district
4 students: Student enrollment
5 teachers: Number of teachers
6 calworks: % of qualifying for CalWorks (income assistance)
7 lunch: % qualifying for reduced-price lunch
8 computer: Number of computers
9 expenditure: Expenditure per student
10 income: District average income (in USD 1,000)
11 english: % of English learners
12 read: read test score
13 math: math test score
14 testscore: average of math and read
15 str: students/teachers

Cross section

```
data(CASchools)
head(CASchools, 2)

##           school county grades students teachers calworks  lunch computer
## 1 Sunol Glen Unified Alameda KK-08      195    10.90   0.5102  2.0408      67
## 2 Manzanita Elementary  Butte  KK-08      240    11.15  15.4167 47.9167     101
##  expenditure income  english  read  math testscore      str
## 1   6384.911 22.690 0.000000 691.6 690.0      690.8 17.88991
## 2   5099.381  9.824 4.583333 660.5 661.9      661.2 21.52466
```

Panel Data

dataset: Cigarettes

```
state      year      cpi      population packs      income
tax        price     taxes
```

1. state: State
2. year: Year
3. cpi: Consumer price index
4. population: State population
5. packs: Number of packs per capita
6. income: State personal income (total, nominal)
7. tax: Average state, federal and average local excise taxes for fiscal year
8. price: Average price during fiscal year, including sales tax
9. taxes: Average excise taxes for fiscal year, including sales tax

Panel Data

	state	year	cpi	population	packs	income	tax	price	taxs
01	AL	1985	1.076	3973000	116.48	46014968	32.5	102.18	33.34
02	AR	1985	1.076	2327000	128.53	26210736	37.0	101.47	37.00
03	AZ	1985	1.076	3184000	104.52	43956936	31.0	108.57	36.17
[...]									
49	AL	1995	1.524	4262731	101.085	83903280	40.5	158.37	41.90
50	AR	1995	1.524	2480121	111.042	45995496	55.5	175.54	63.85
51	AZ	1995	1.524	4306908	71.954	88870496	65.3	198.60	74.79
[...]									
94	WI	1995	1.524	5137004	92.466	115959680	62.0	201.38	71.58
95	WV	1995	1.524	1820560	115.568	32611268	41.0	166.51	50.42
96	WY	1995	1.524	478447	112.238	10293195	36.0	158.54	36.00

Formal definition of data

The data are

- a **sample** of size n , denoted

$$\{Y_1, Y_2, \dots, Y_n\}$$

- Y_1 is the first observation, Y_2 is the second observation, etc.
- for cross section the typical observation is the i^{th} observation, denoted Y_i
- for time series the typical observation is customarily denoted by Y_t
- if we have data on more than one variable, we have a **multivariate** sample,

$$\{(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)\}$$

Central tendency

(sample) mean, (sample) median

Dispersion

(sample) variance, (sample) standard deviation

Position

(sample) Percentiles, (sample) deciles, and (sample) quartiles

Central tendency

- The leading measure of central tendency is the **sample mean**, which is the arithmetic average of the data
- For a sample of size n , the sample mean

$$\bar{Y} = (Y_1 + Y_2 + \dots + Y_n)/n$$

- Often, this formula is abbreviated using the summation convention

$$\bar{Y} = \sum_{i=1}^n Y_i/n$$

Central tendency

- The other leading indicator of central tendency is the (sample) **median**, which is the value of the sample that divides the data after ordering into two halves, the median being the midpoint
- The median is relatively easy to calculate:
 - Odd Number of Data Values (n is odd)
 1. arrange data in order from smallest to largest
 2. Find the data value in the **exact** middle
 - Even Number of Data Values (n is even)
 1. arrange data in order from smallest to largest
 2. Find the mean of the **two** middle numbers

- the **sample variance**

- is the average of the deviation of the data from the mean

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

- the **sample standard deviation**

- is the square root of the sample variance

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Measure of symmetry (skewness)

The **skewness** measures the symmetry of the distribution:

$$skew = \frac{1}{n} \sum_{i=1}^n \left(\frac{Y_i - \bar{Y}}{s} \right)^3$$

- A **positive skewed** or **right-skewed** data have a much longer tail on the right ($skew > 0$)
- A **negative skewed** or **left-skewed** data have a much longer tail on the left ($skew < 0$)

Measure of peakedness (kurtosis)

The peakedness of the distribution and fatness of the tails is measured by the **kurtosis**

$$kurt = \frac{1}{n} \sum_{i=1}^n \left(\frac{Y_i - \bar{Y}}{s} \right)^4 - 3$$

- **positive kurtosis** indicates data that is relatively flat
- **negative kurtosis** indicates data that is relatively peaked

The sample **quartiles** divides the data in 4 parts:

- the **lower quartile** (Q_1) is that point where one-quarter of the observed sample lies below and three-quarters of the order sample lies above
- the **middle quartile** (Q_2) is the sample median
- the **upper quartile** (Q_3) is that point where three-quarters of the ordered sample lies below and one-quarter of the order sample lies above.

Even more detailed divisions of the sample are possible.

- **Deciles** split the ordered sample into tenths and are used, for example, to summarize the distribution of individual income
- **Percentiles** split the order sample into hundredths. The p^{th} percentile is the value for which p percent of the observed values are equal to or less than the value

Summarizing data

Table 1: Descriptive Statistics, California Schools Dataset.

Statistic	Mean	St. Dev.	Pctl(25)	Median	Pctl(75)
students	2,629.0	3,913.0	379	950.5	3,008
teachers	129.1	187.9	19.7	48.6	146.4
calworks	13.2	11.5	4.4	10.5	19.0
lunch	44.7	27.1	23.3	41.8	66.9
computer	303.4	441.3	46	117.5	375.2
expenditure	5,312.0	633.9	4,906.0	5,214.0	5,601.0
income	15.3	7.2	10.6	13.7	17.6
english	15.8	18.3	1.9	8.8	23.0
read	655.0	20.1	640.4	655.8	668.7
math	653.3	18.8	639.4	652.4	665.8

Summarizing data in R

```
data(CASchools)
head(CASchools[, -c(1:4)], 4)

##   teachers calworks  lunch computer expenditure income  english  read  math
## 1    10.90   0.5102  2.0408         67    6384.911 22.690  0.000000 691.6 690.0
## 2    11.15  15.4167 47.9167        101    5099.381  9.824  4.583333 660.5 661.9
## 3    82.90  55.0323 76.3226        169    5501.955  8.978 30.000002 636.3 650.9
## 4    14.00  36.4754 77.0492         85    7101.831  8.978  0.000000 651.9 643.5
##   testscore      str
## 1    690.8 17.88991
## 2    661.2 21.52466
## 3    643.6 18.69723
## 4    647.7 17.35714
```

Summarizing data in R

```
summary(CASchools[c("students", "teachers", "math", "read")])
```

##	students	teachers	math	read
##	Min. : 81.0	Min. : 4.85	Min. :605.4	Min. :604.5
##	1st Qu.: 379.0	1st Qu.: 19.66	1st Qu.:639.4	1st Qu.:640.4
##	Median : 950.5	Median : 48.56	Median :652.5	Median :655.8
##	Mean : 2628.8	Mean : 129.07	Mean :653.3	Mean :655.0
##	3rd Qu.: 3008.0	3rd Qu.: 146.35	3rd Qu.:665.9	3rd Qu.:668.7
##	Max. :27176.0	Max. :1429.00	Max. :709.5	Max. :704.0

Summarizing data in R

```
summary(ceo[c("salary", "sales", "roe", "ros")])
```

##	salary	sales	roe	ros
##	Min. : 223	Min. : 175.2	Min. : 0.5	Min. : -58.00
##	1st Qu.: 735	1st Qu.: 2200.3	1st Qu.: 12.4	1st Qu.: 20.75
##	Median : 1032	Median : 3693.3	Median : 15.5	Median : 52.00
##	Mean : 1282	Mean : 6824.4	Mean : 17.2	Mean : 61.18
##	3rd Qu.: 1408	3rd Qu.: 7017.2	3rd Qu.: 20.0	3rd Qu.: 81.00
##	Max. : 14822	Max. : 97649.9	Max. : 56.3	Max. : 418.00

Graphical representation of data

Graphical methods used vary with the type of univariate data

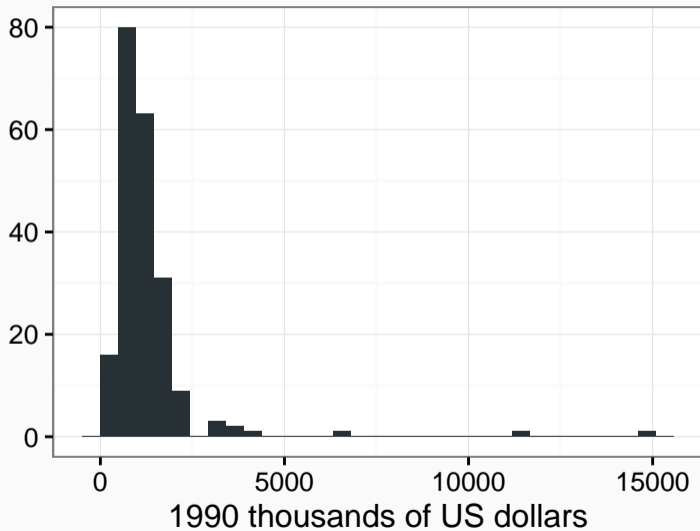
Cross-section

histogram, boxplot, pie-chart

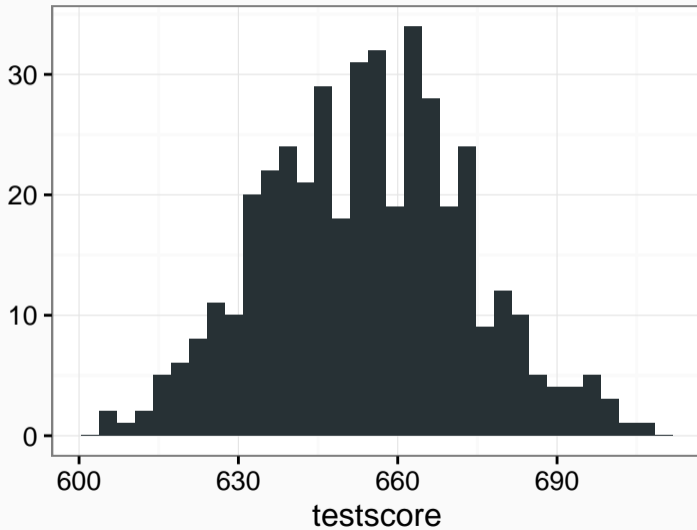
Time series

line chart

Cross section



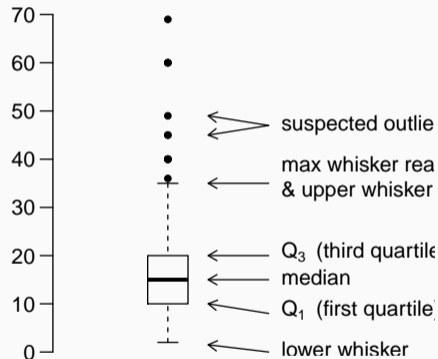
Cross section



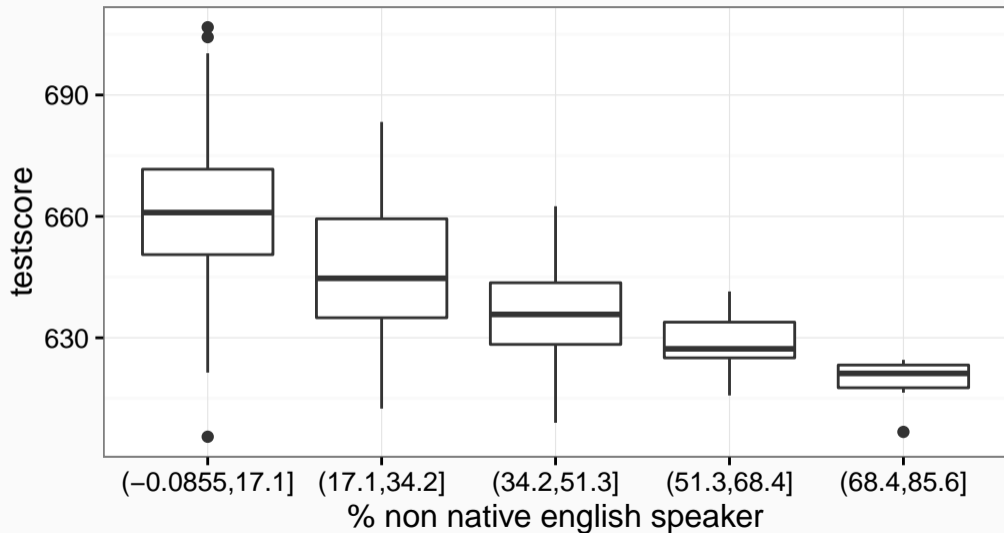
Box plots

A **box plot** or boxplot is a convenient way of graphically depicting groups of numerical data through their quartiles.

Box plots have lines extending vertically from the boxes (**whiskers**) indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot.

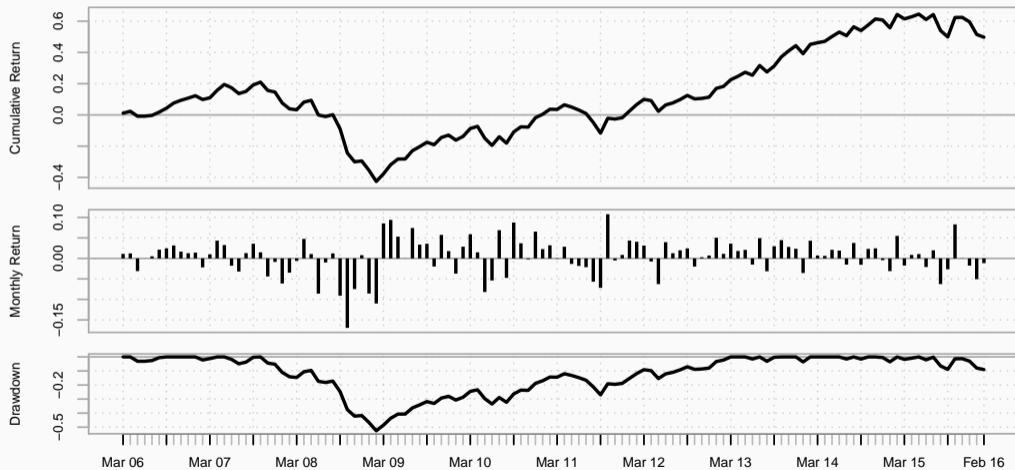


Cross section



Time series

S&P 500



Bivariate data analysis

Bivariate data analysis considers the relationship between two variables, such as education and income, or price and house size, or test score and str.

Data summary tools:

- scatterplot (graphical)
- covariance and correlation (numerical)

Scatter plot

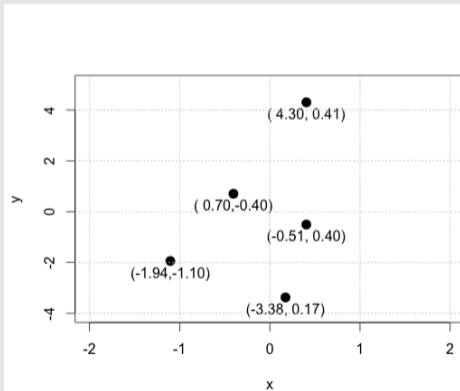
The data is displayed as a collection of points.

Example (Data)

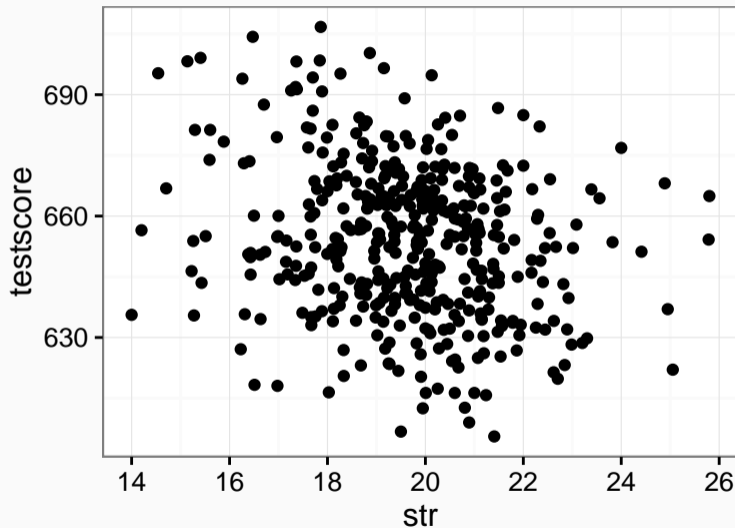
Suppose to have the following bivariate sample:

	Y	X
1	-0.51	0.40
2	0.70	-0.40
3	-3.38	0.17
4	-1.94	-1.10
5	4.30	0.41

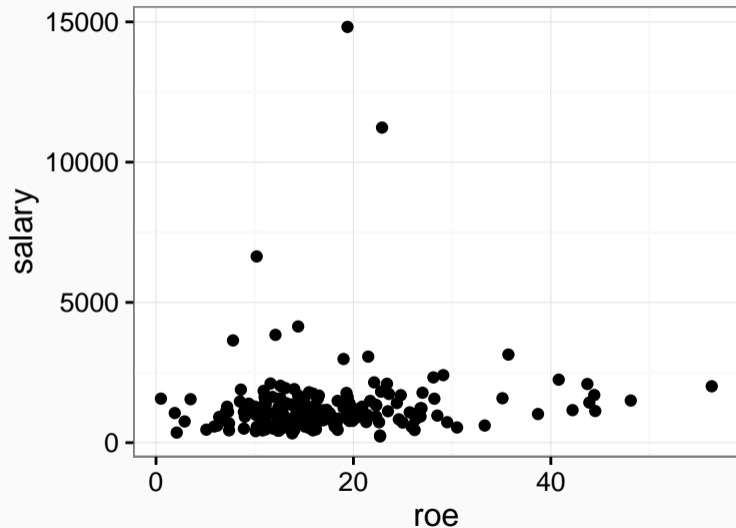
Scatterplot



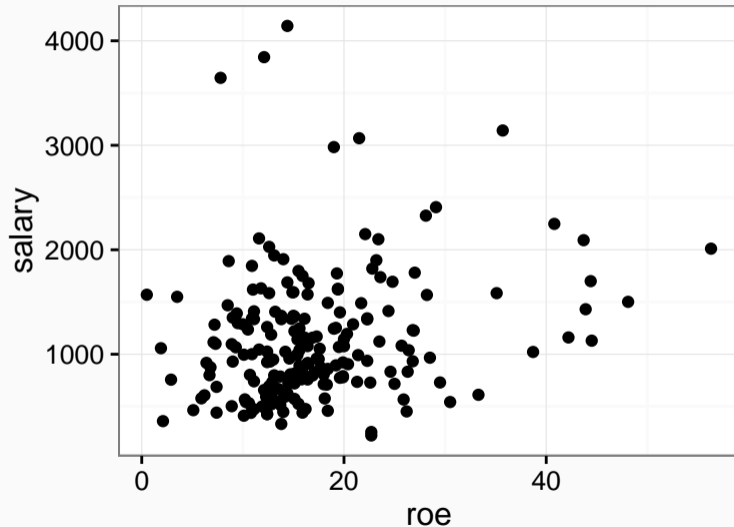
Scatter plot: testscore and str



Scatter plot: salary and roe



Scatter plot: salary and roe



Covariance

A measure of association between two variables, say x and y is the **covariance**:

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

A measure of association between two variables, say x and y is the **covariance**:

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

- $s_{XY} > 0$: if x and y tend to move together in the **same** direction

Covariance

A measure of association between two variables, say x and y is the **covariance**:

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

- $s_{XY} > 0$: if x and y tend to move together in the **same** direction
- $s_{XY} < 0$: if X and Y tend to move together in the **opposite** direction

A measure of association between two variables, say x and y is the **covariance**:

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

- $s_{XY} > 0$: if x and y tend to move together in the **same** direction
- $s_{XY} < 0$: if X and Y tend to move together in the **opposite** direction
- $s_{XY} = 0$: no association

Covariance

A measure of association between two variables, say x and y is the **covariance**:

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

- $s_{XY} > 0$: if x and y tend to move together in the **same** direction
- $s_{XY} < 0$: if X and Y tend to move together in the **opposite** direction
- $s_{XY} = 0$: no association

The covariance measures only linear association between y and x

Covariance between testscore and str

The covariance between Test Score and str is negative:

```
with(CASchools, cov(str, testscore))
```

```
## [1] -8.159323
```

- the unit of measure of the covariance are **difficult** to interpret

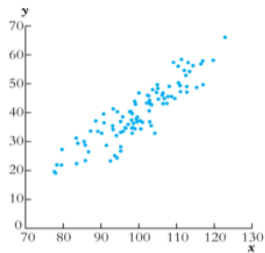
Correlation coefficient

The correlation coefficient is defined in terms of the covariance:

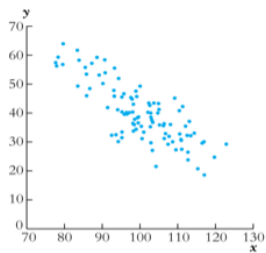
$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

- $-1 \leq r_{XY} \leq 1$
- $r_{XY} = 1$ mean perfect **positive** linear association
- $r_{XY} = -1$ means perfect **negative** linear association
- $r_{XY} = 0$ means **no** linear association

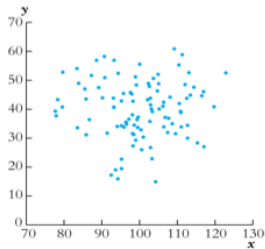
The correlation coefficient measures linear association



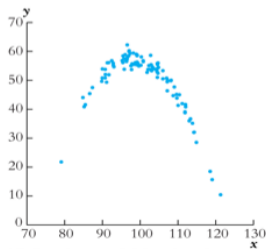
(a) Correlation = +0.9



(b) Correlation = -0.8



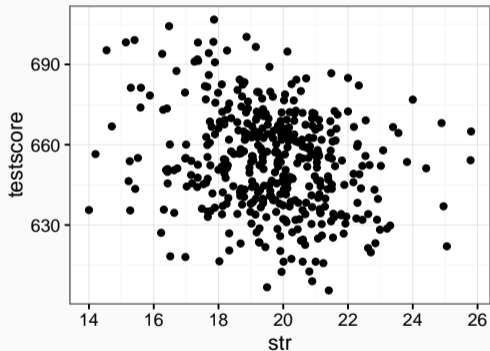
(c) Correlation = 0.0



(d) Correlation = 0.0 (quadratic)

Correlation testscore vs. str

How big do you think it is?



```
with(CASchools, cor(str, testscore))  
  
## [1] -0.2263627
```


Correlation coefficient

The correlation coefficient is defined as

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

where s_X and s_Y are the sample standard deviations of X and Y , respectively.

Remarks

- treat X and Y symmetrically: $r_{XY} = r_{YX}$
- while r_{XY} detects (linear) association, it is neutral on whether it is X that is causing Y or Y that is causing X
- it can be shown that r_{XY} measure the number of standard deviations that Y changes by when X changes by one standard deviation