Introductory to econometrics

ESSEC
BUSINESS SCHOOL

Practical information

Contact: dufays@essec.edu

Grading

- Two homework to do in groups of at most 4 **(40%)**
  1. First homework to be hand in before May, 29th.
  2. Second homework to be hand in before June, 19th.
- Written examination during the final session **(60%)**

Book

- Brooks, C. (2019). *Introductory econometrics for finance*. Cambridge university press.

Available at the library for consultation

Statistical software

- R (and R studio)
- (Python)

Why several statistical softwares ?

R

- Academic developers in finance and econometrics
- Library to deal with volatility models (link)

Python

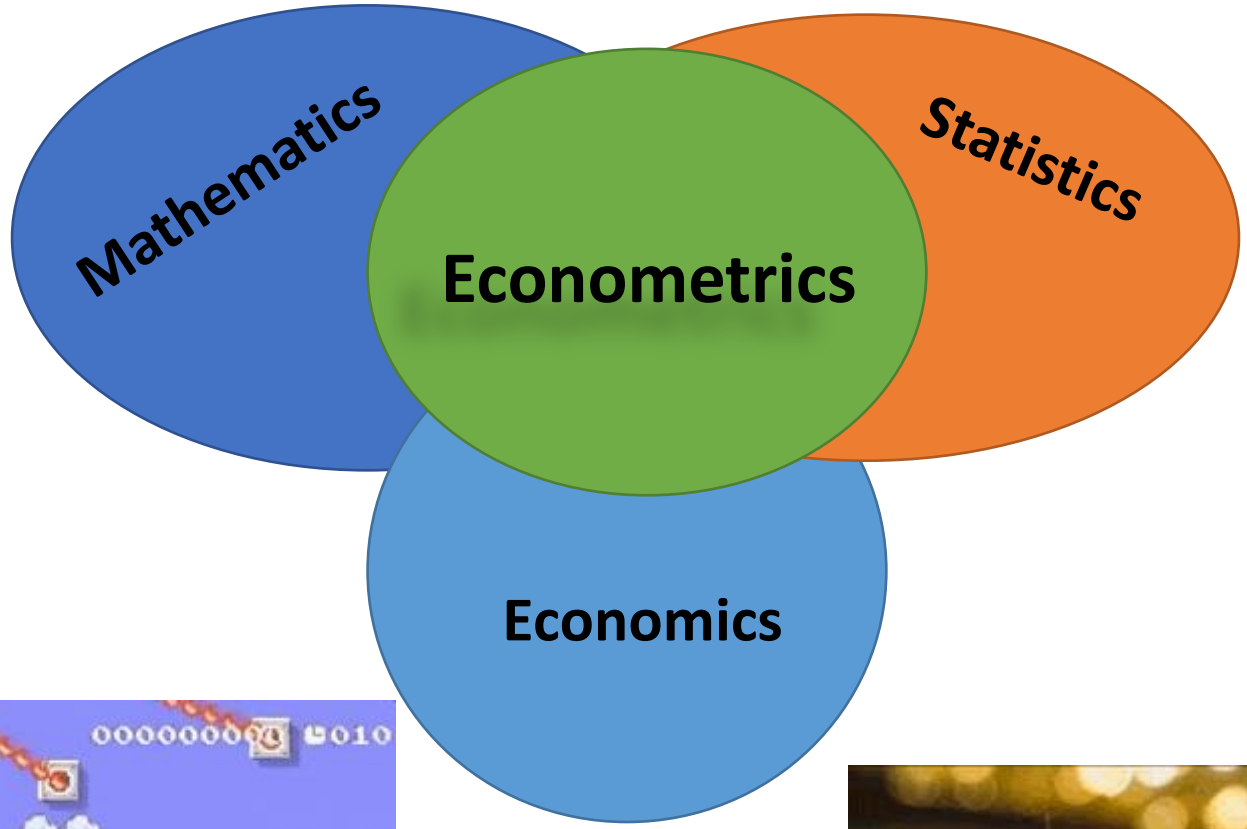
- Highly popular in machine learning
- Incredible growth in popularity

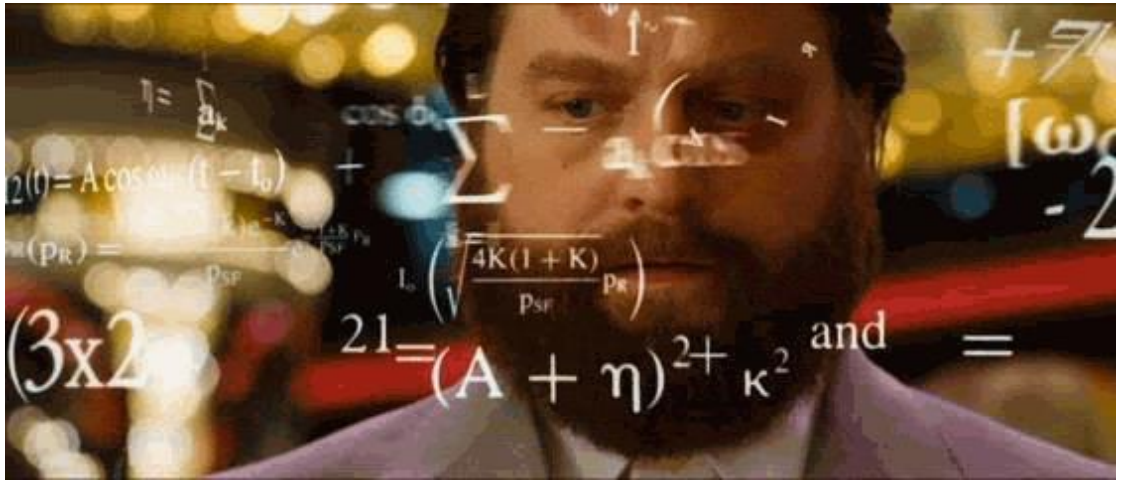
Python and R

- Open-source
- Knowing at least one of them is an asset on the job market.

All the codes are available in each language on the website

Why learning econometrics ?





Why econometrics is useful ?

Questions

Is this apartment worth the price ?

Is financial market efficient ?

Does this mutual fund beat the market or its competitors ?

Will this phone application become popular ?

Will this bank fail in one year ?

Spending hours on conjectures and qualitative arguments or... Applying an econometric model to quantify your answer

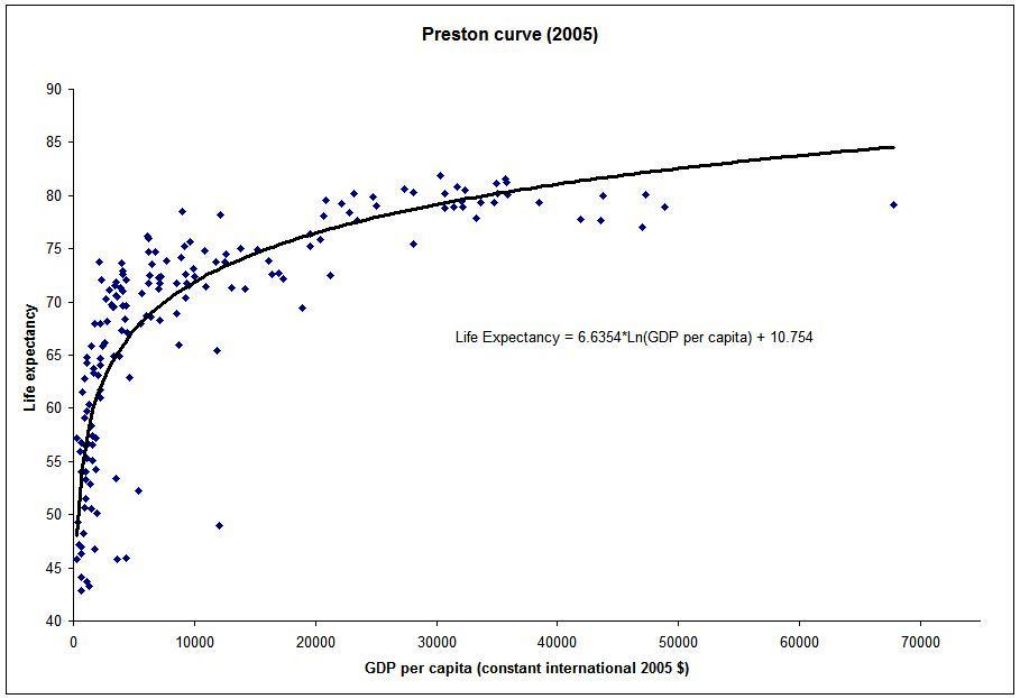
Econometrics helps to make a decision

Aims of the course

- Able to check affirmative sentences:

"Eating chocolate produces more Nobel prize winners" (link)

"Life expectancy is longer in countries with high GDP per capita" (link)

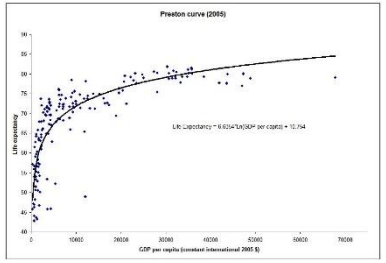


Aims of the course

- Able to check affirmative sentences:

"Eating chocolate produces more Nobel prize winners" (link)

"Life expectancy is longer in countries with high GDP per capita" (link)

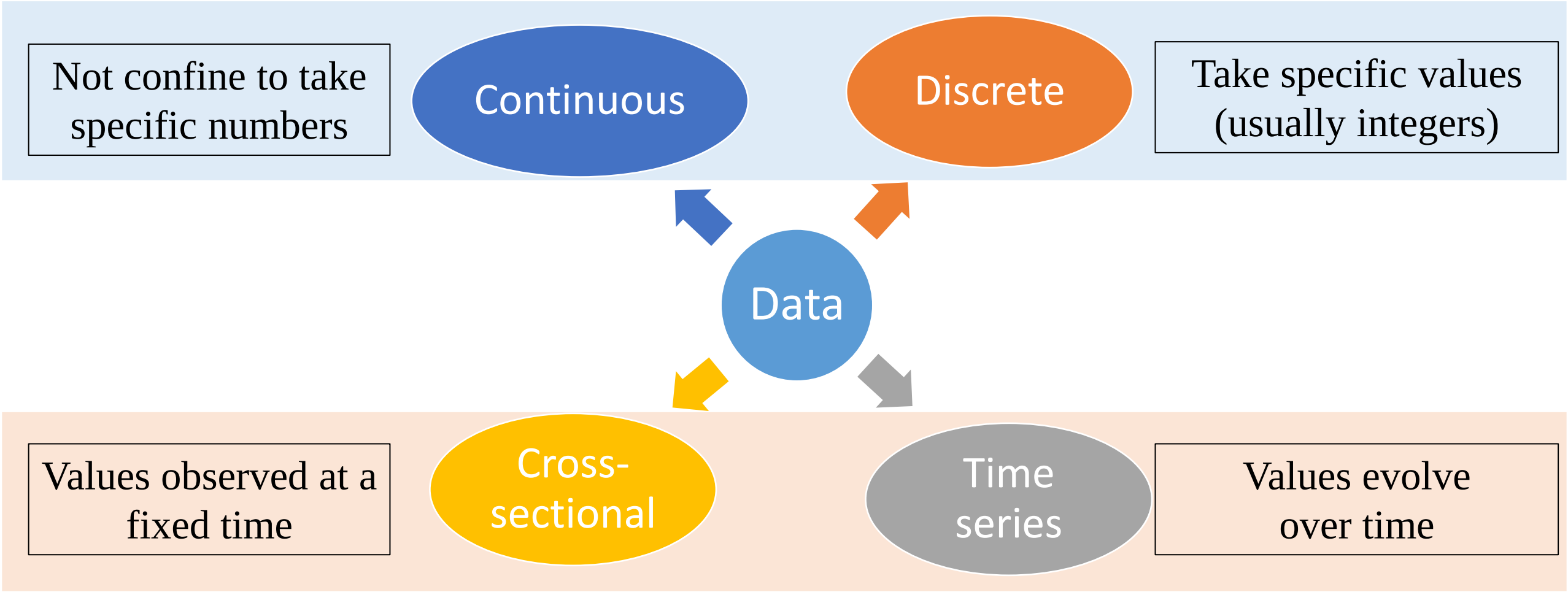


- Able to check the validity of a prediction:

"stock prices have reached 'what looks like a permanently high plateau." Irving Fisher, popular economist on October 16, 1929 (link)

Econometrics and Data

- **Data:** any collection of items
  - A list of prices, of eyes colors, of movie rankings from one to ten, …



Types of data

- **Cross-sectional data:** A list of prices, salaries, eyes colors, observed at a specific time.

Example:

Prices of all the assets included in the S&P 500 at the closing time on First September 2020.
Salary of the employees of a company in January 2020.

- **Time series data:** A list of prices, salaries, observed over time.

Example:

Prices of one financial asset at the closing time observed every day in 2020.
Salary of one employee over time.

We will develop statistical tools for both types of data

General outline of the course

1. Statistic foundations and introduction to linear regressions.
   - Statistical tests and predictions **of continuous variables**

2. How do we deal with discrete variables ?
   - Statistical tests and predictions **of discrete variables**

3. How do we deal with time series ?
   - How do we adapt the linear regression to **predict time series** ?

Outline of the first econometric framework

First statistical problem

Louis Bachelier (1870-1946)

Ph.d thesis:

- **Asset returns cannot be predicted!**

BUT

Every return of an asset is a realization of the same statistical distribution.

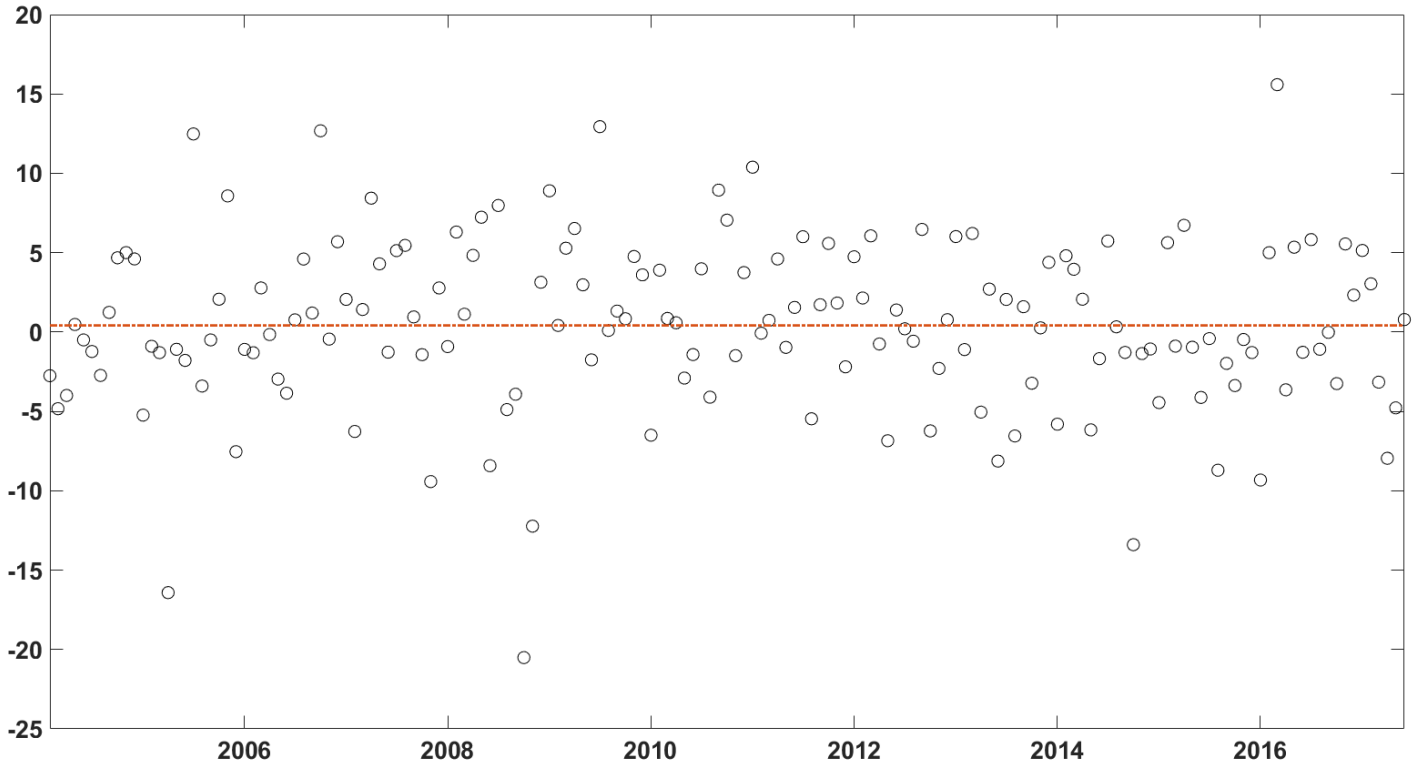
Normal distribution in finance

Louis Bachelier (1870-1946)

$X \sim N(\mu, \sigma^2)$

Every new return = a realization from a Normal distribution

Monthly IBM returns from Feb, 2004 to Jun, 2017 (T=161 observations).



How do I estimate the expectation and the variance ?

Normal distribution in finance

Louis Bachelier (1870-1946)

Every new return = a realization from a Normal distribution

- **Useful ?**

- Idea of the expected return.
- Idea of the range of future returns.
- Idea of how risky is the asset.

How do I estimate these quantities ?

Statistical foundations

A quick reminder

Summary statistics

Before any statistical analysis

Have a look to your data and compute **summary statistics**

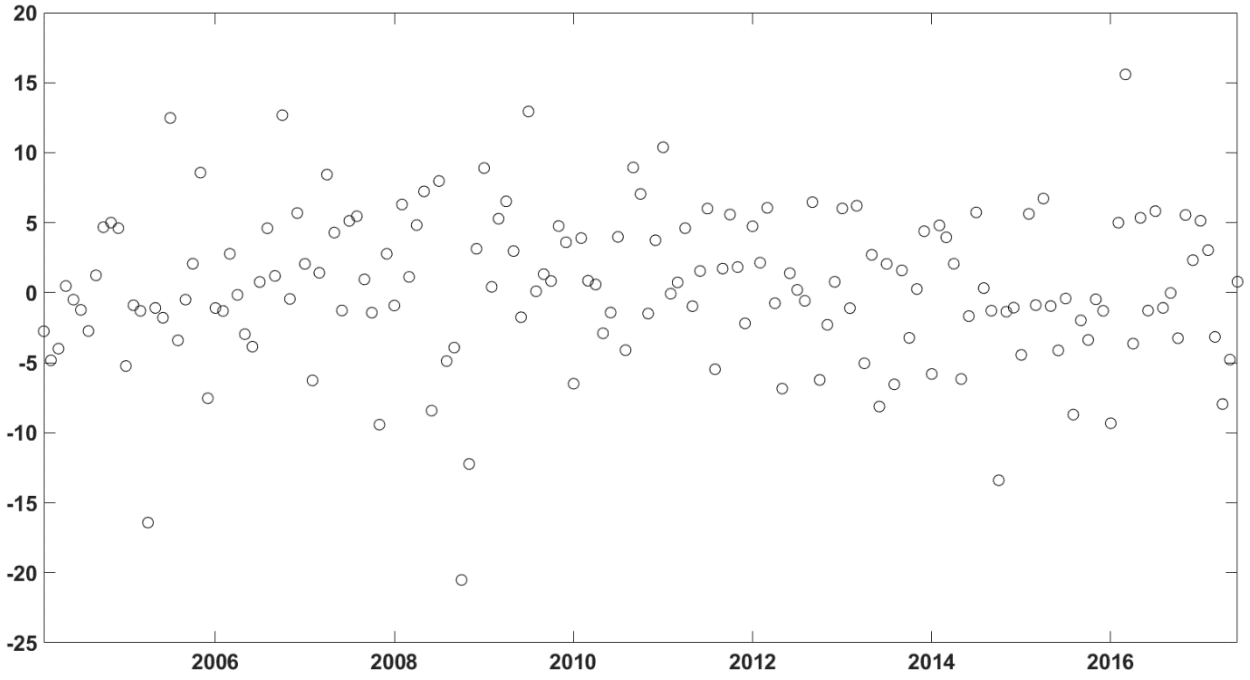
Summary statistics give answers to

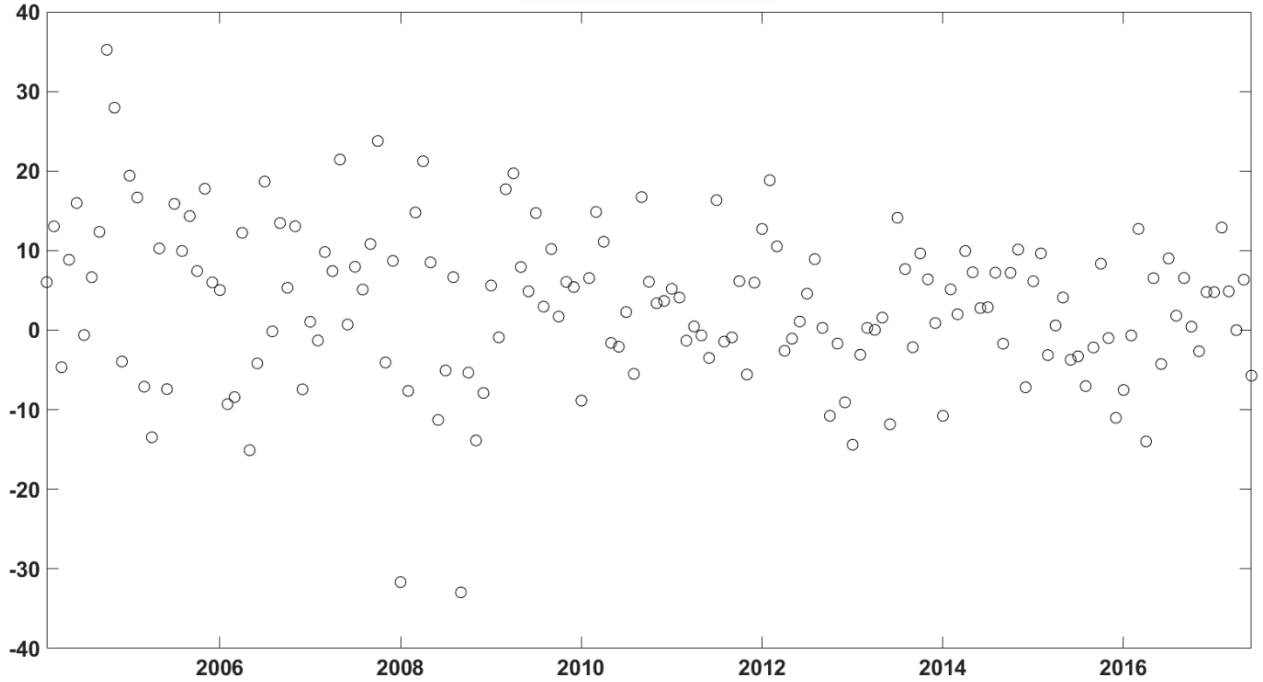
1. How do the data look like ?
2. What is the average value ?
3. How does it vary around the average value ?
4. Are the data symmetric around the average value ?

Summary statistics

Monthly returns from Feb, 2004 to Jun, 2017 (T=161 observations).

- **Graphic of the data**



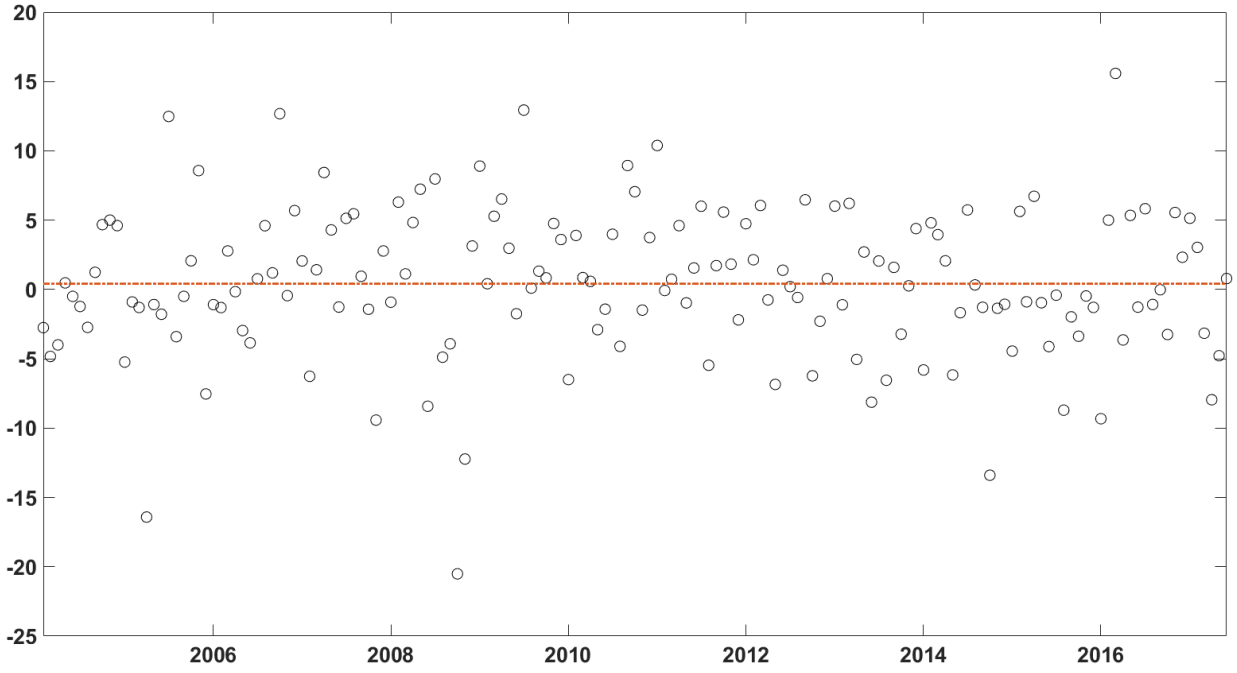


Summary statistics

Monthly returns from Feb, 2004 to Jun, 2017 (T=161 observations).

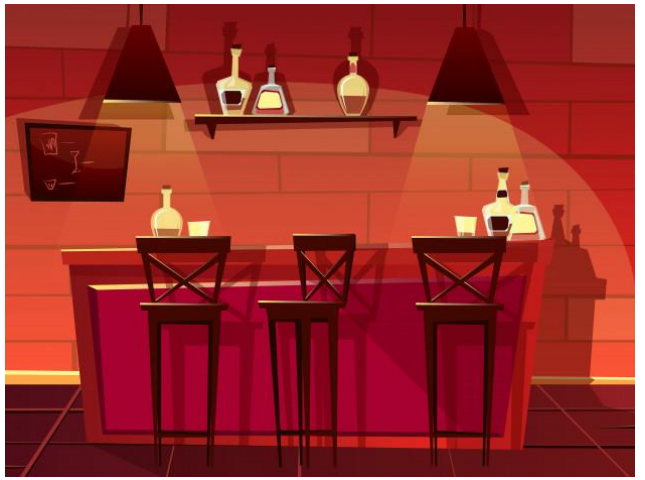
- **Monthly average returns:** $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$

$$\bar{y} = 0.42\%$$



$$\bar{y} = 3.31\%$$

Summary statistics



- Average salary: 1800 EUR

- What happens when Mark Zuckerberg enters in the bar ?

Hello! I am here to screw your average

Average salary: 100 000 EUR

Average sensitive to outliers.

Summary statistics

- **Median is insensitive to outliers:** $Q_{50\%}$
- **Quantiles:** $Q_{X\%}$

> Value which cuts the sorted sample into X% and 100-X%

Example:

10	20	8	2	9	1000	5	18	13	16	1

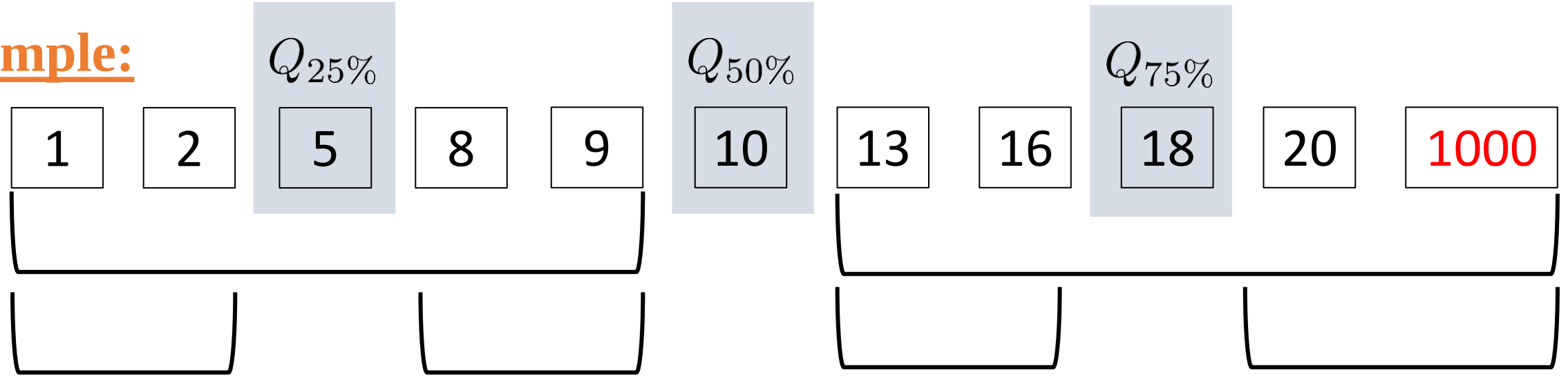
Average: 100.18

Quantiles: **First sort your series in ascending order**

Summary statistics

- **Quantiles:** $Q_{X\%}$

Example:



- **Median:** **cut the sample in two sample of exact same size**
- **1st quartile:** **cut the first split sample in two sample of exact same size**
- **3rd quartile:** **cut the 2nd split sample in two sample of exact same size**

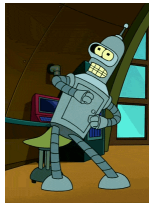
Summary of a series with 5 numbers

Minimum, 1st quartile, median, 3rd quartile, Maximum

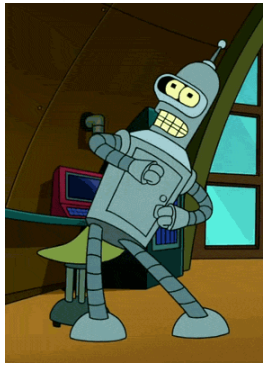
1	5	10	18	1000

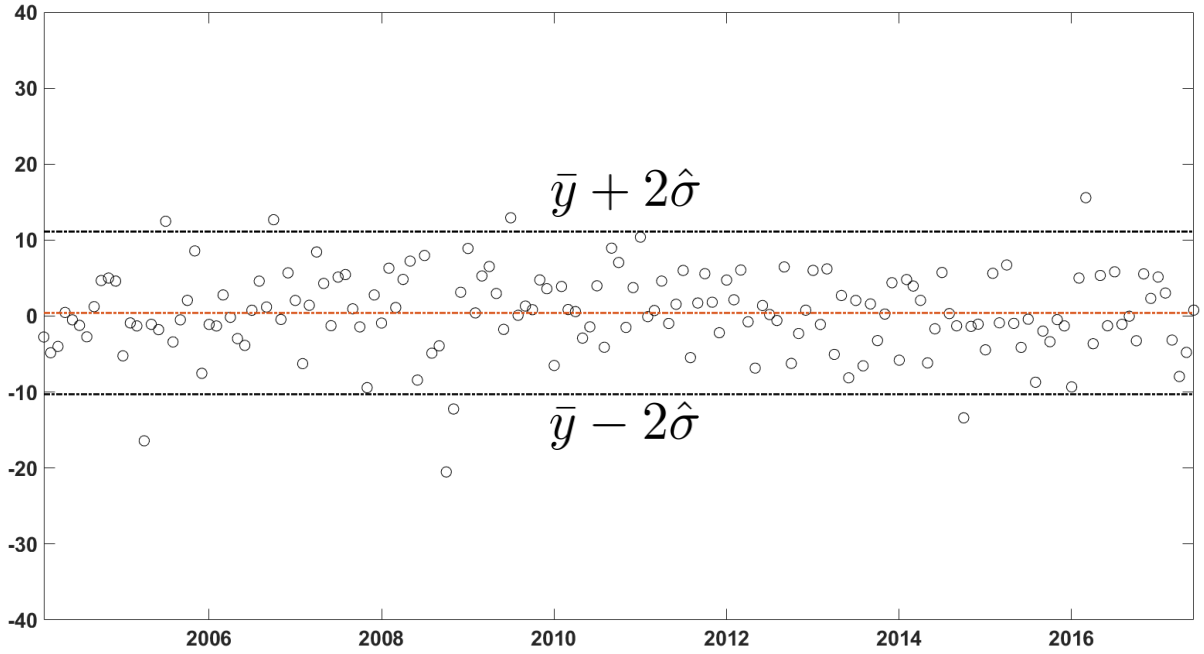
Summary statistics

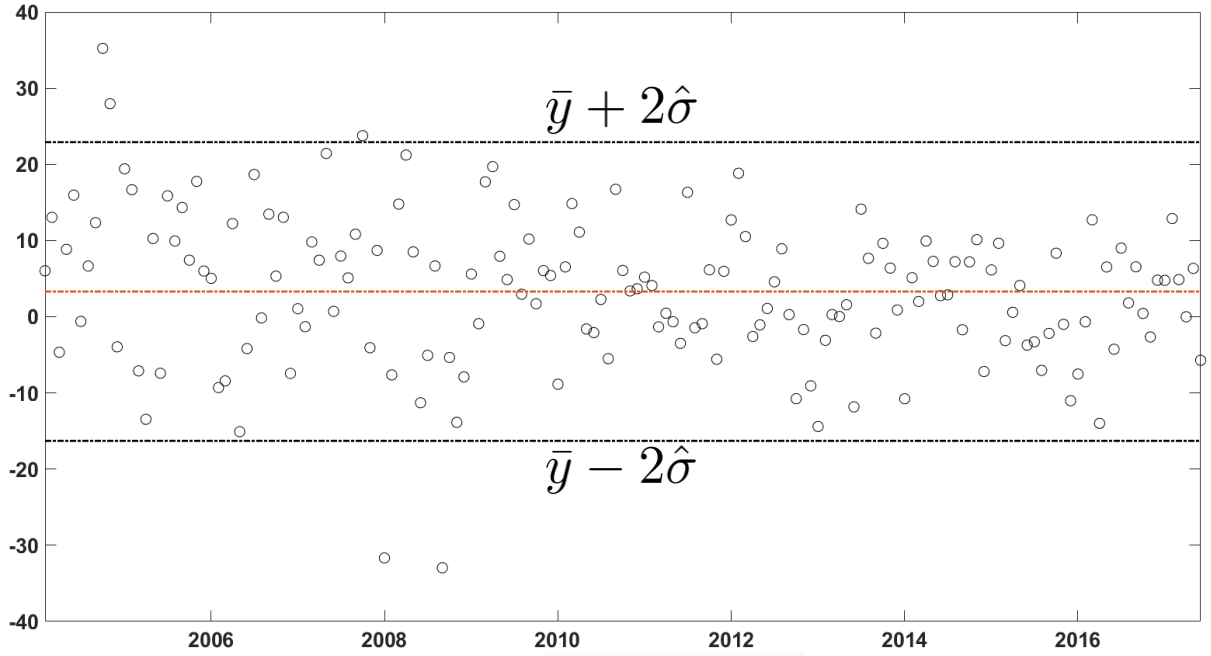
- **What variation do we expect around the average return ?**
- **Standard deviation:** $\hat{\sigma} = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(y_t - \bar{y})^2}$



$$\bar{y} = 0.42\% \quad \longleftrightarrow \quad \bar{y} = 3.31\%$$
$$\hat{\sigma} = 5.35\% \quad \longleftrightarrow \quad \hat{\sigma} = 9.80\%$$





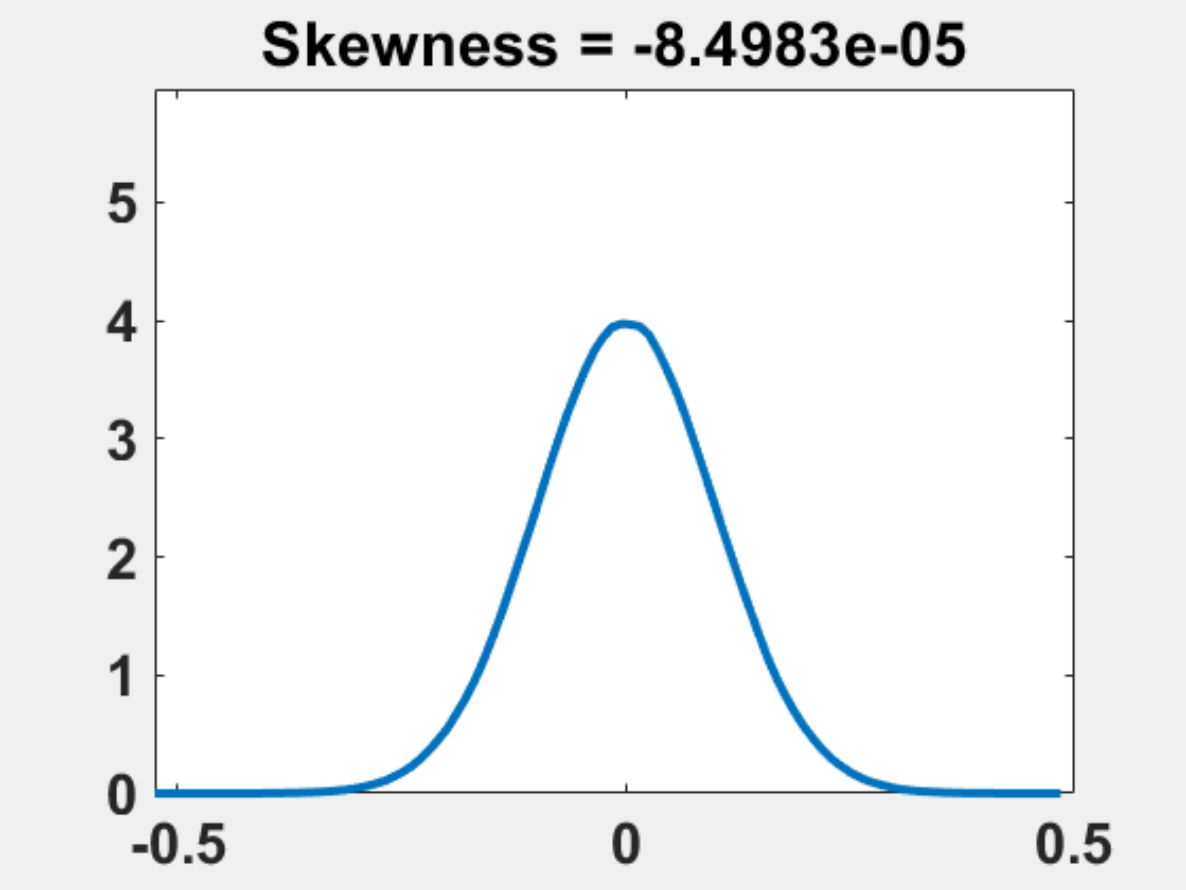


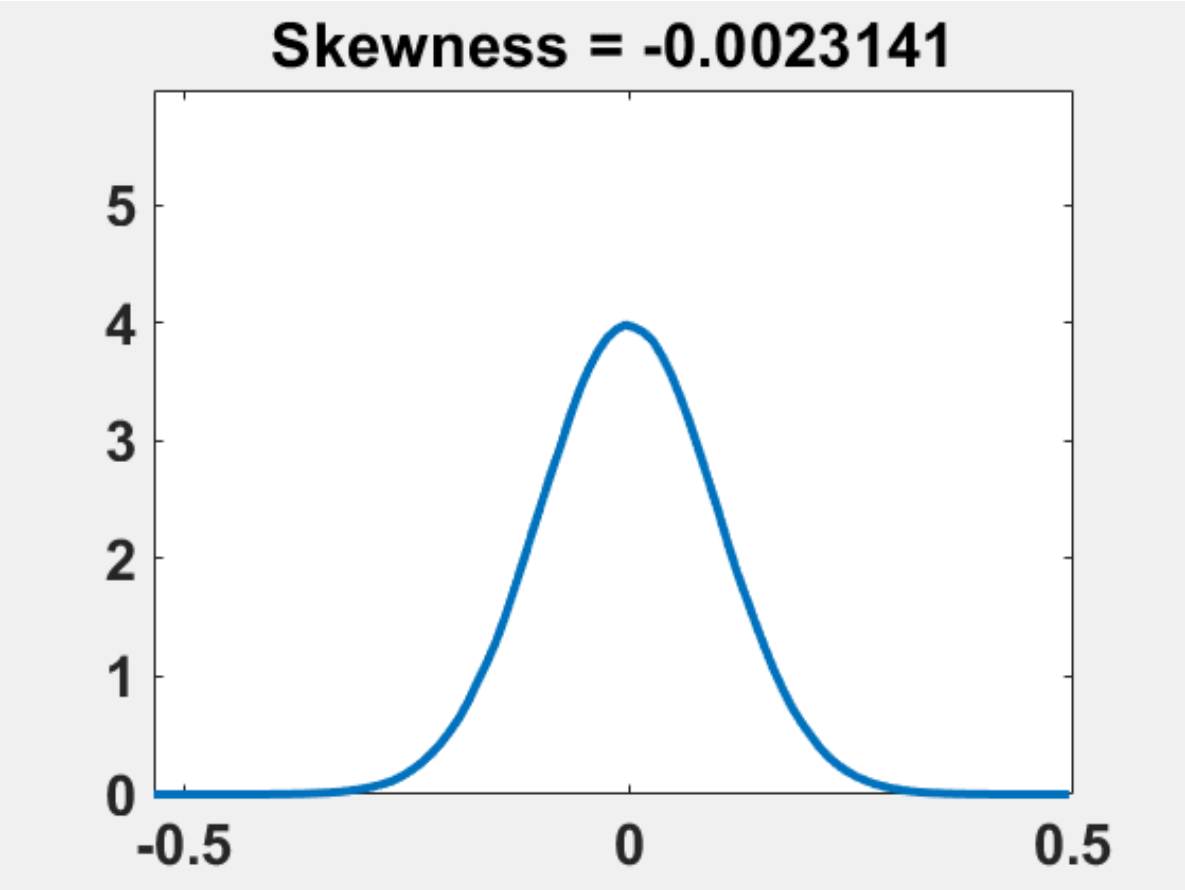
Summary statistics

- **Symmetric or asymmetric around the average ?**
- **Skewness:** $\kappa = \frac{\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^3}{(\hat{\sigma}^2)^{\frac{3}{2}}}$

IBM $\kappa = -0.42$

$\kappa = -0.25$





Summary statistics

- Mean, Median, Quartiles, standard deviation, skewness

Good way to start but… **are these statistics of any help ?**

Useful in a statistical framework!

Louis Bachelier (1870-1946)

Every new return = a realization from a Normal distribution

We need to relate summary statistics to random variables

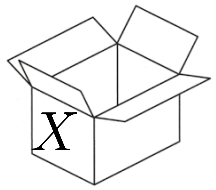
Random variables

- **Random variables:** take on any value from a given set and this value is determined at least in part by chance.

Notation

Random Variable: X

Realization: x

$x \sim$ 

Important quantities (*if they exist*):

Cumulative density function: $P[X \leq x] \in [0, 1]$

Probability density function: $f(X = x) \equiv f(x) \geq 0$

$$\frac{dP[X \leq x]}{dx} = f(x) \iff \int_{-\infty}^{x} f(z)dz = P[X \leq x]$$

Expectation: $E(X) = \int_{-\infty}^{\infty} xf(x)dx$

Variance: $V(X) = E(X^2) - E(X)^2$

Normal Distribution

Notation

$X \sim N(\mu, \sigma^2)$

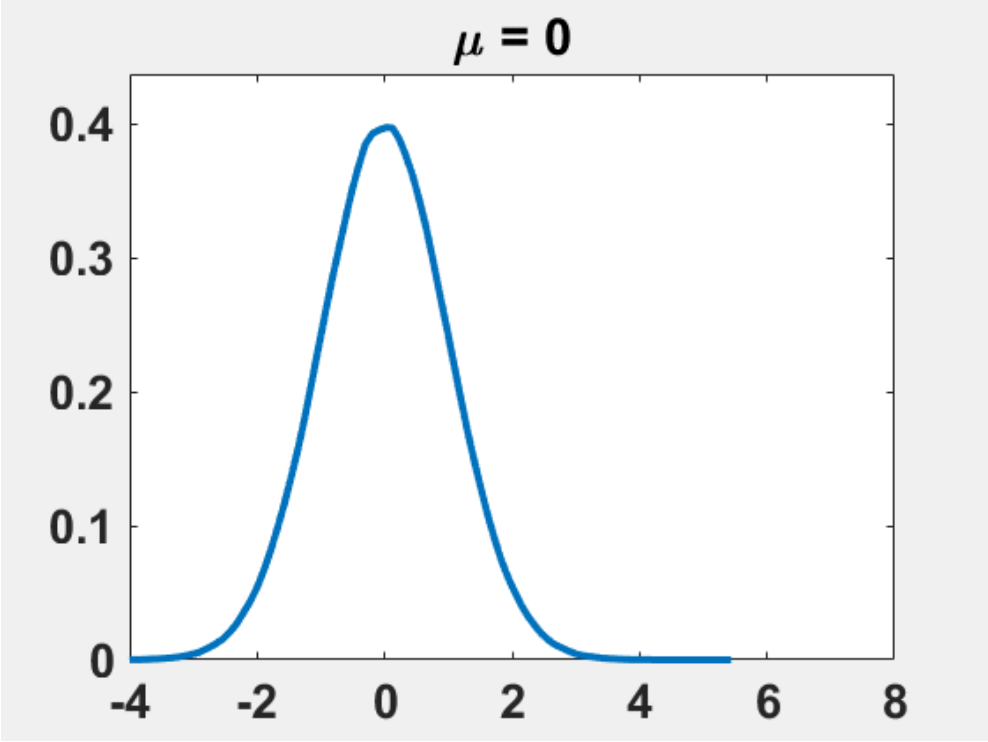
Expectation: μ

Variance: σ^2

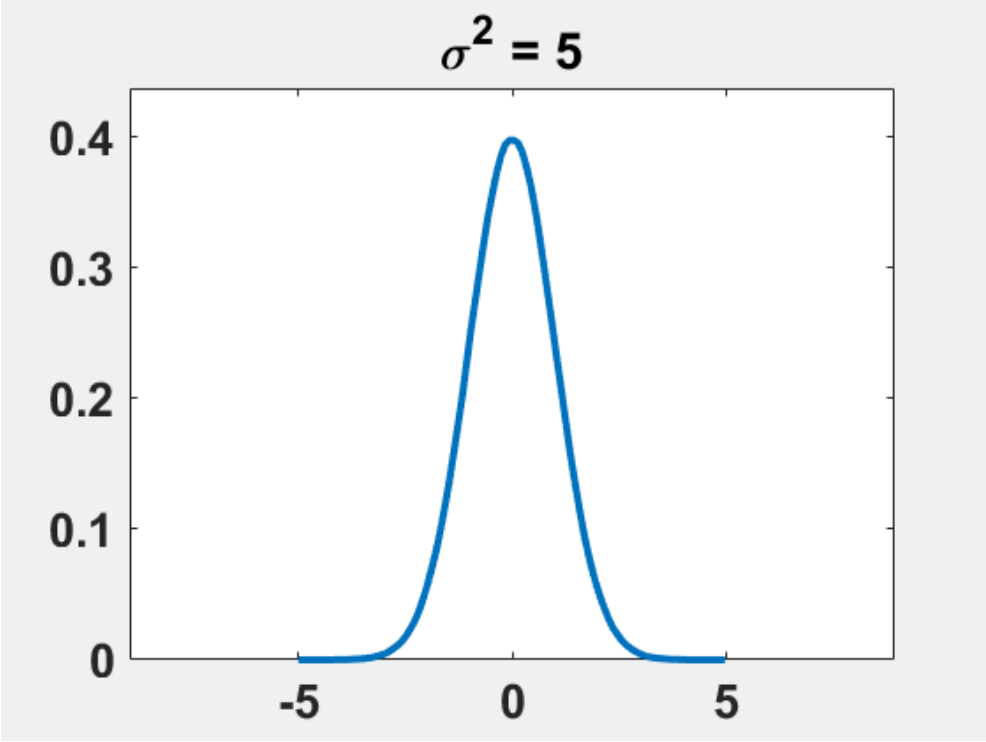
- **Normal distribution:** $X \sim N(\mu, \sigma^2)$

Probability density function: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$

Expectation:



Variance:



Useful trick: $P[X \in (\mu - 2\sigma; \mu + 2\sigma)] \approx 0.95$

Normal Distribution

Notation

$$X \sim N(\mu, \sigma^2)$$

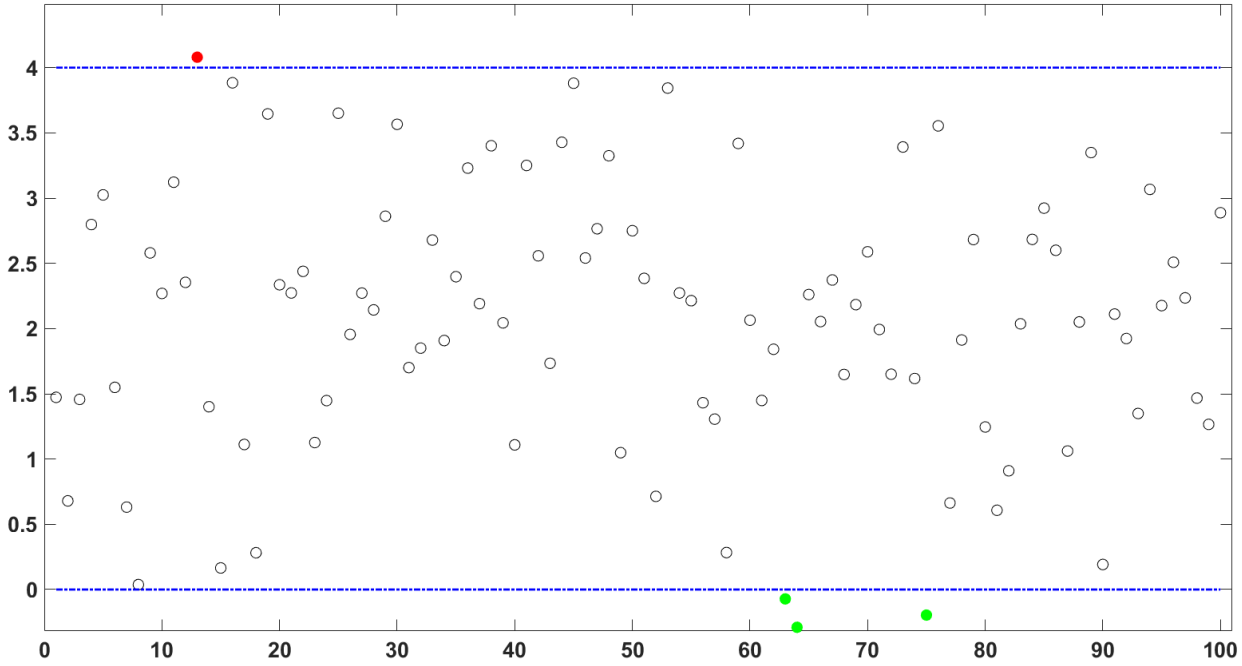
Expectation: μ

Variance: σ^2

- **Normal distribution:** $X \sim N(\mu, \sigma^2)$

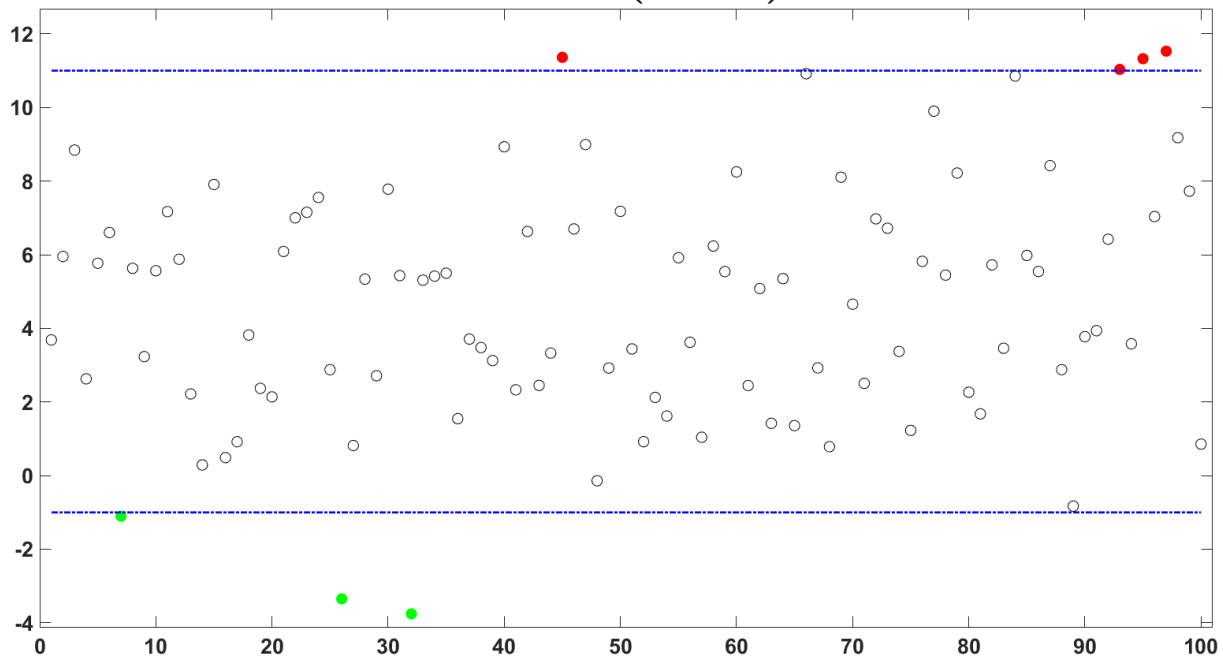
Useful trick: $P[X \in (\mu - 2\sigma; \mu + 2\sigma)] \approx 0.95$

$N(2, 1)$



#Out: 4

$N(4, 9)$



#Out: 7

Normal Distribution

Notation

$X \sim N(\mu, \sigma^2)$

Expectation: μ

Variance: σ^2

- **Normal distribution:** $X \sim N(\mu, \sigma^2)$
- **Useful property:**

$$Z = X + a \sim N(\mu + a, \sigma^2)$$

$$Z = \frac{X}{a} \sim N(\frac{\mu}{a}, \frac{\sigma^2}{a^2})$$

$$Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$$

Normal Distribution

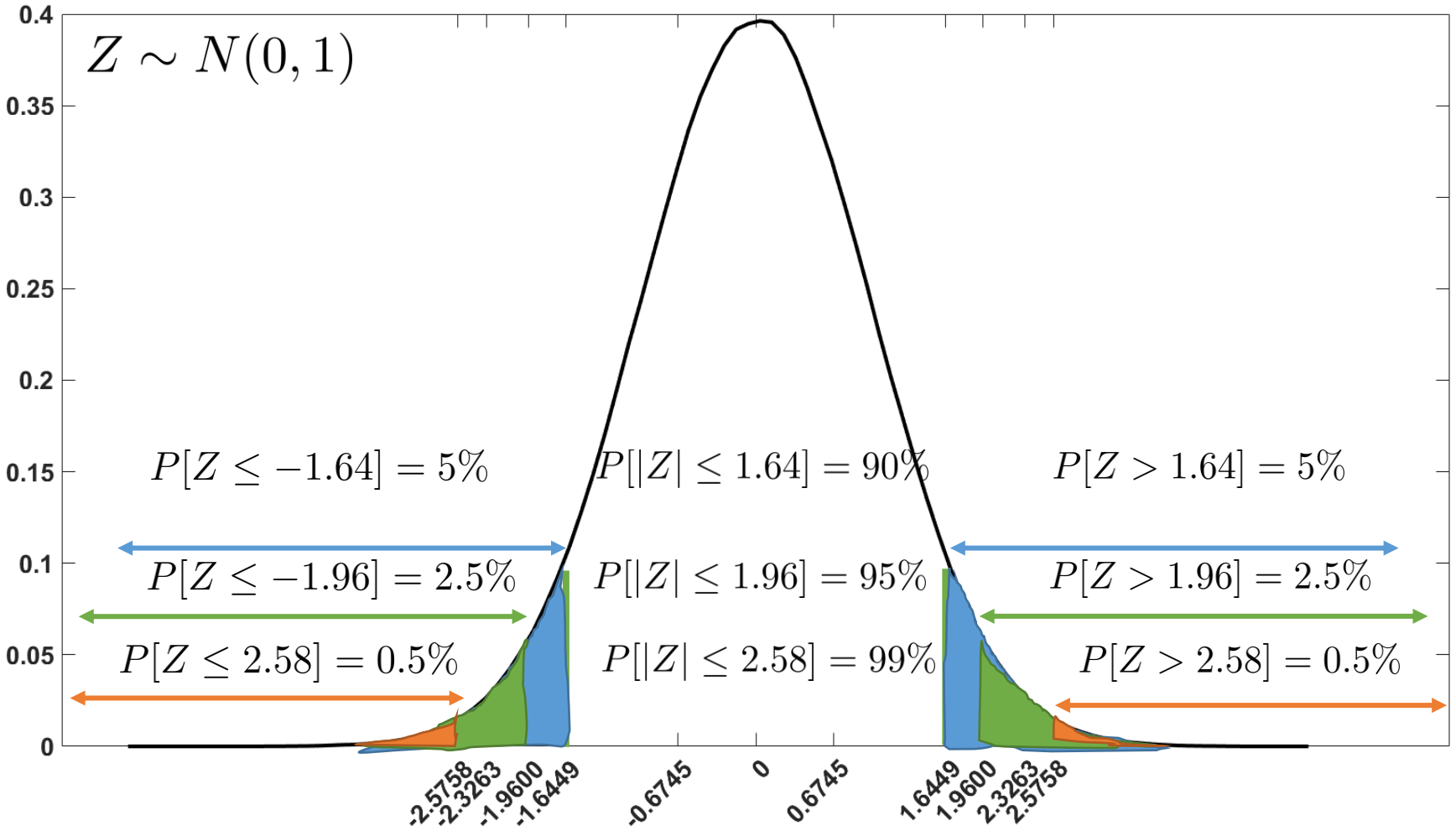
$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

Notation

$X \sim N(\mu, \sigma^2)$

Expectation: μ

Variance: σ^2



Fluctuation interval:

90%: [-1.64;1.64]

95%: [-1.96;1.96]

99%: [-2.58;2.58]

Normal Distribution

Notation

$X \sim N(\mu, \sigma^2)$

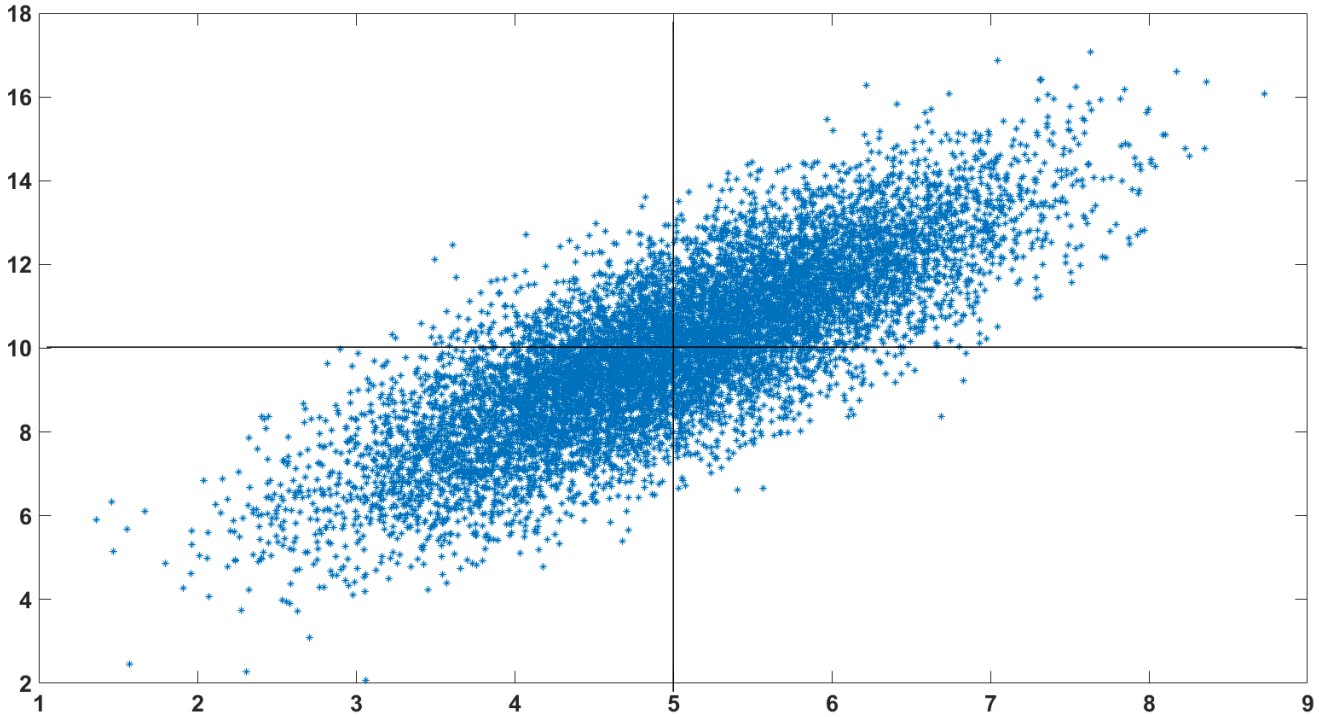
Expectation: μ

Variance: σ^2

- **Two random variables:** $X_1 \sim N(\mu_1, \sigma_1^2),\ X_2 \sim N(\mu_2, \sigma_2^2)$

Correlation: How two variables are linearly related

Positive correlation:



Correlation: Between -1 and 1

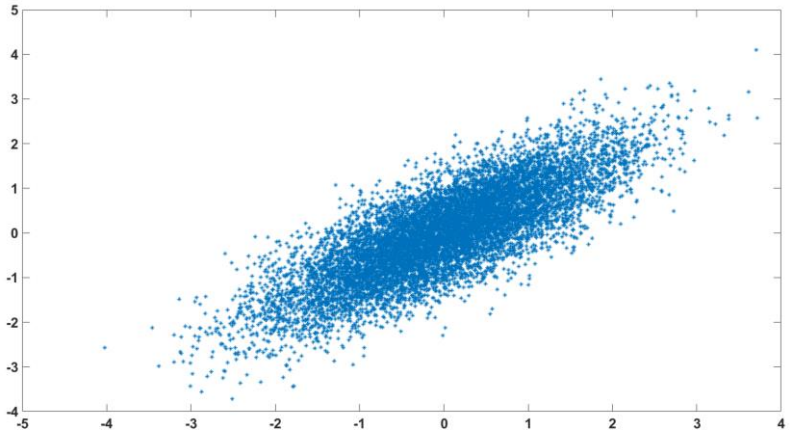
Normal Distribution

Notation

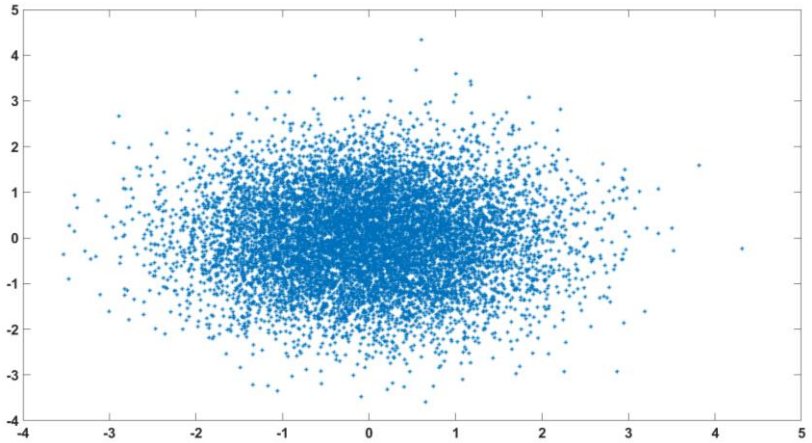
$X \sim N(\mu, \sigma^2)$
Expectation: μ
Variance: σ^2

- **Two random variables:** $X_1 \sim N(\mu_1, \sigma_1^2),\ X_2 \sim N(\mu_2, \sigma_2^2)$

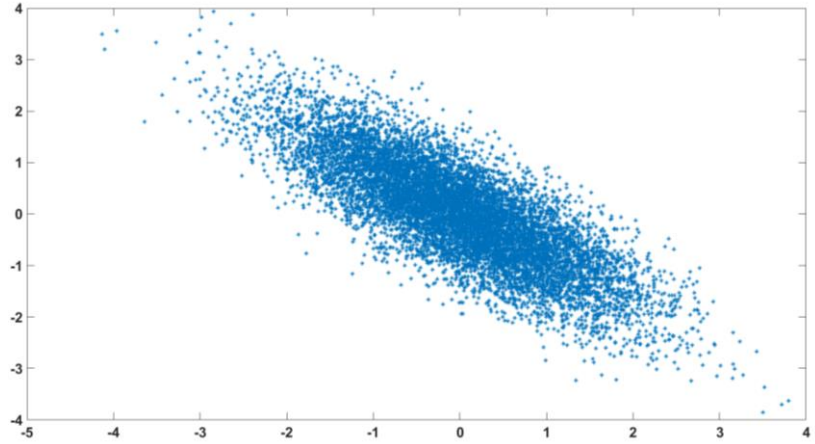
Positive correlation:



No correlation:



Negative correlation:



- **Useful property:**

Linear transformation of r.v. jointly normally distributed:

$$Z = aX_1 + bX_2 \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\mathrm{Cov}(X_1, X_2))$$

No correlation:

$$Z = aX_1 + bX_2 \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$$

Normal distribution in finance

Every new return = a realization from a Normal distribution

IBM $Y_t \sim N(\mu, \sigma^2)$

Every new return = cannot be predicted (independent)

Summary Statistics:

Average: $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$

one draw
one realization

Random variable:

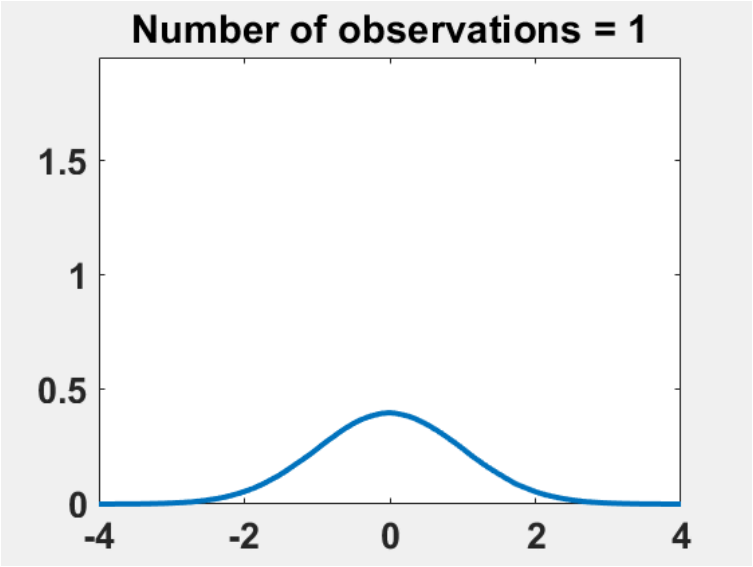
$\bar{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t$

No correlation

$\bar{Y} \sim N(\mu, \frac{\sigma^2}{T})$

Unbiasedness

$E(\bar{Y}) = \mu$



Normal distribution in finance

Black swan

Every new return = a realization from a Normal distribution

IBM $Y_t \sim N(\mu, \sigma^2)$

Every new return = cannot be predicted (independent)

Summary Statistics:

$$\hat{\sigma}^2 = \frac{1}{T-1}\sum_{t=1}^{T}(y_t - \bar{y})^2$$

one draw
one realization

Random variable:

$$Z = \frac{1}{T-1}\sum_{t=1}^{T}(Y_t - \bar{Y})^2$$

Unbiasedness

$$E(Z) = \sigma^2$$

$$Z = \frac{1}{T-1}\sum_{t=1}^{T}(Y_t - \mu)^2 + (\bar{Y} - \mu)^2 - 2T(\bar{Y} - \mu)^2$$

$$E(Z) = \frac{1}{T-1}[T\sigma^2 + \frac{T\sigma^2}{T} - 2T\frac{\sigma^2}{T}]$$

General result

Law of large numbers (LLN)

Let $\{Z_t\}_{t=1}^{T}$ be independent identically distributed (i.i.d.) with $E(Z_t) = \mu$.

Then $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} Z_t = \mu$ (a.s.)

$\bar{z} = \frac{1}{T}\sum_{t=1}^{T} z_t$ $\quad E(\bar{z}) = \frac{1}{T}\sum_{t=1}^{T} \mu$

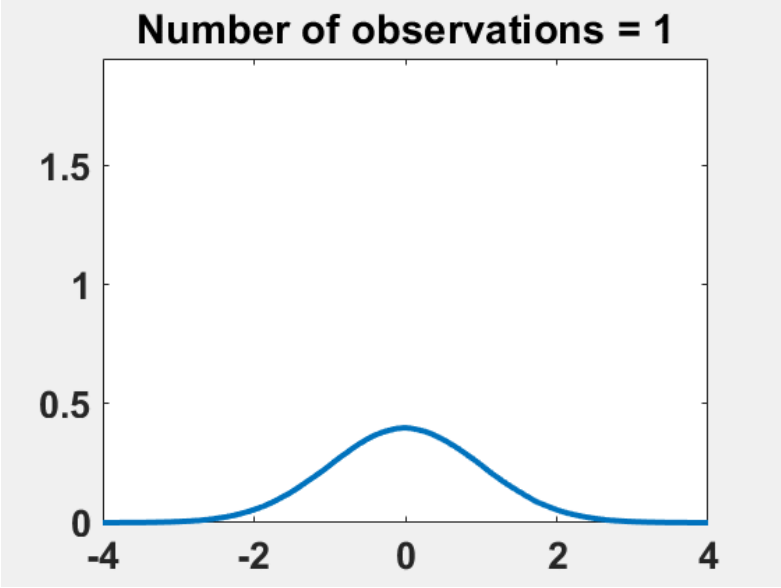
One realization is unpredictable, the average of the realizations is predictable

Average of Normal distributions:

$Z_t \sim N(\mu, \sigma^2) \Rightarrow \frac{1}{T}\sum_{t=1}^{T} Z_t \sim N(\mu, \frac{\sigma^2}{T})$

Average = -0.66016

$x_1 = -0.66016$



LLN: practical example

Law of large numbers:

Z_t : Result of the roulette wheel

$$P[Z_t = ODD] = \frac{18}{37}, P[Z_t = \text{EVEN}] = \frac{18}{37}, P[Z_t = 0] = \frac{1}{37}$$

Strategy:

At every roulette wheel, I bet 10EUR on EVEN numbers.

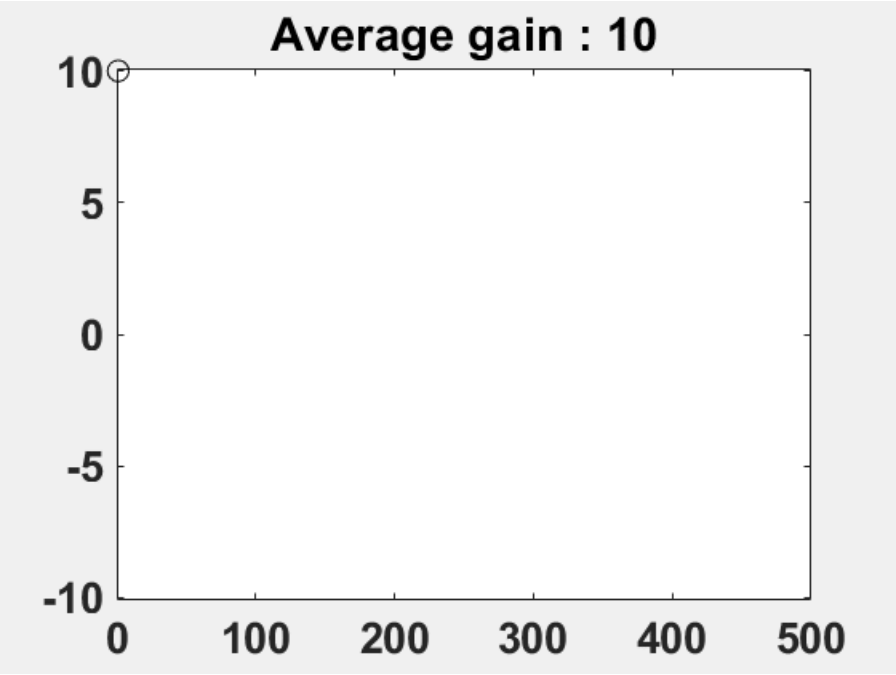
Gain and loss:

If EVEN: I win 10EUR ⇒ $G_t = 10$ if Z_t is EVEN.

If ODD or zero: I loose 10EUR ⇒ $G_t = -10$ if Z_t is ODD or 0.

Expected Gain per game: $E(G_t) = 10\frac{18}{37} - 10\frac{19}{37} = -0.27$

LLN:



After 100 wheel spinning, Expected loss of 27 EUR

Law of Large Numbers

How to estimate the expectation and the variance ?

Expectation: $E(Z) = \mu$ → $\bar{z} = \frac{1}{T}\sum_{t=1}^{T} z_t$

Expectation: $E(Z^2)$ → $\frac{1}{T}\sum_{t=1}^{T} z_t^2$

Variance: $\sigma^2 = E(Z^2) - E(Z)^2$ → $\frac{1}{T}\sum_{t=1}^{T} z_t^2 - \bar{z}^2$

Law of large numbers

Complex expectation: $E(f(Z_t))$ → Good estimate: $\frac{1}{T}\sum_{t=1}^{T} f(z_t)$

Application: Minimum variance portfolio

Financial portfolio with 2 assets:

$$X_p = \omega_1 X_1 + \omega_2 X_2 \text{ with } \omega_1 + \omega_2 = 1 \Rightarrow \boxed{\omega_2 = 1 - \omega_1}$$

Portfolio expected return: $E[X_p] = \omega_1 E[X_1] + \omega_2 E[X_2] = \omega_1 \mu_1 + \omega_2 \mu_2$

Portfolio variance:

$$\begin{aligned} V[X_p] &= E[(\omega_1 X_1 + \omega_2 X_2 - \omega_1 \mu_1 - \omega_2 \mu_2)^2] \\ &= \omega_1^2 V[X_1] + \omega_2^2 V[X_2] + 2\omega_1 \omega_2 \mathrm{Cov}(X_1, X_2) \end{aligned}$$

Minimum variance: $\boxed{\omega_2 = 1 - \omega_1}$

$$\begin{aligned} V[X_p] &= \omega_1^2 V[X_1] + (1-\omega_1)^2 V[X_2] + 2\omega_1(1-\omega_1)\mathrm{Cov}(X_1, X_2), \\ &= \omega_1^2 (V[X_1] + V[X_2] - 2\mathrm{Cov}(X_1, X_2)) + 2\omega_1(\mathrm{Cov}(X_1, X_2) - V[X_2]) + V[X_2] \end{aligned}$$

Application: Minimum variance portfolio

Financial portfolio with 2 assets:

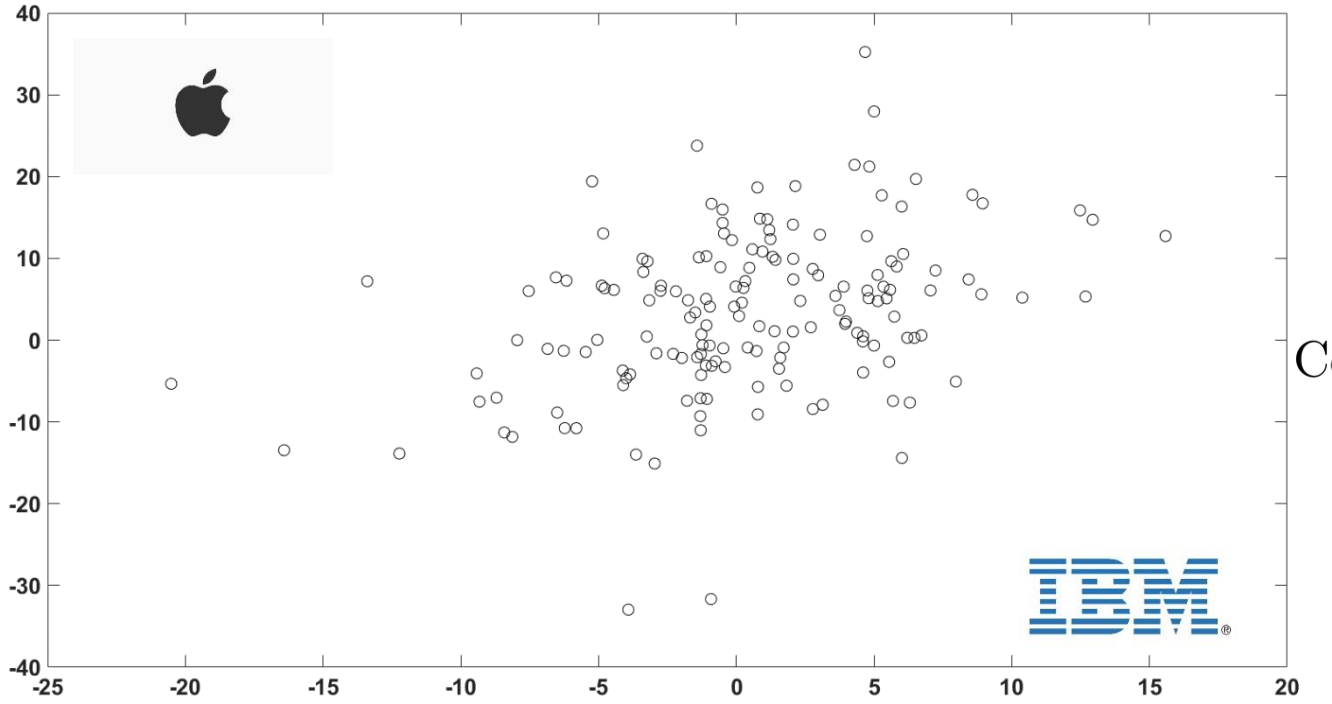
Minimum variance:

$$\begin{aligned} V[X_p] &= \omega_1^2 V[X_1] + (1-\omega_1)^2 V[X_2] + 2\omega_1(1-\omega_1)\mathrm{Cov}(X_1, X_2), \\ &= \omega_1^2(V[X_1] + V[X_2] - 2\mathrm{Cov}(X_1, X_2)) + 2\omega_1(\mathrm{Cov}(X_1, X_2) - V[X_2]) + V[X_2] \end{aligned}$$

IBM

$$\bar{y} = 0.42\%$$
$$\hat{\sigma}^2 = 28.65$$

$$\bar{y} = 3.31\%$$
$$\hat{\sigma}^2 = 96.06\%$$



$$\mathrm{Cov}(App, IBM) = 19.31$$

$$\mathrm{Corr}(App, IBM) = \frac{19.31}{\sqrt{28.65}\sqrt{96.06}} = 0.37$$

Application: Portfolio

$\bar{y} = 0.42\%$ $\bar{y} = 3.31\%$

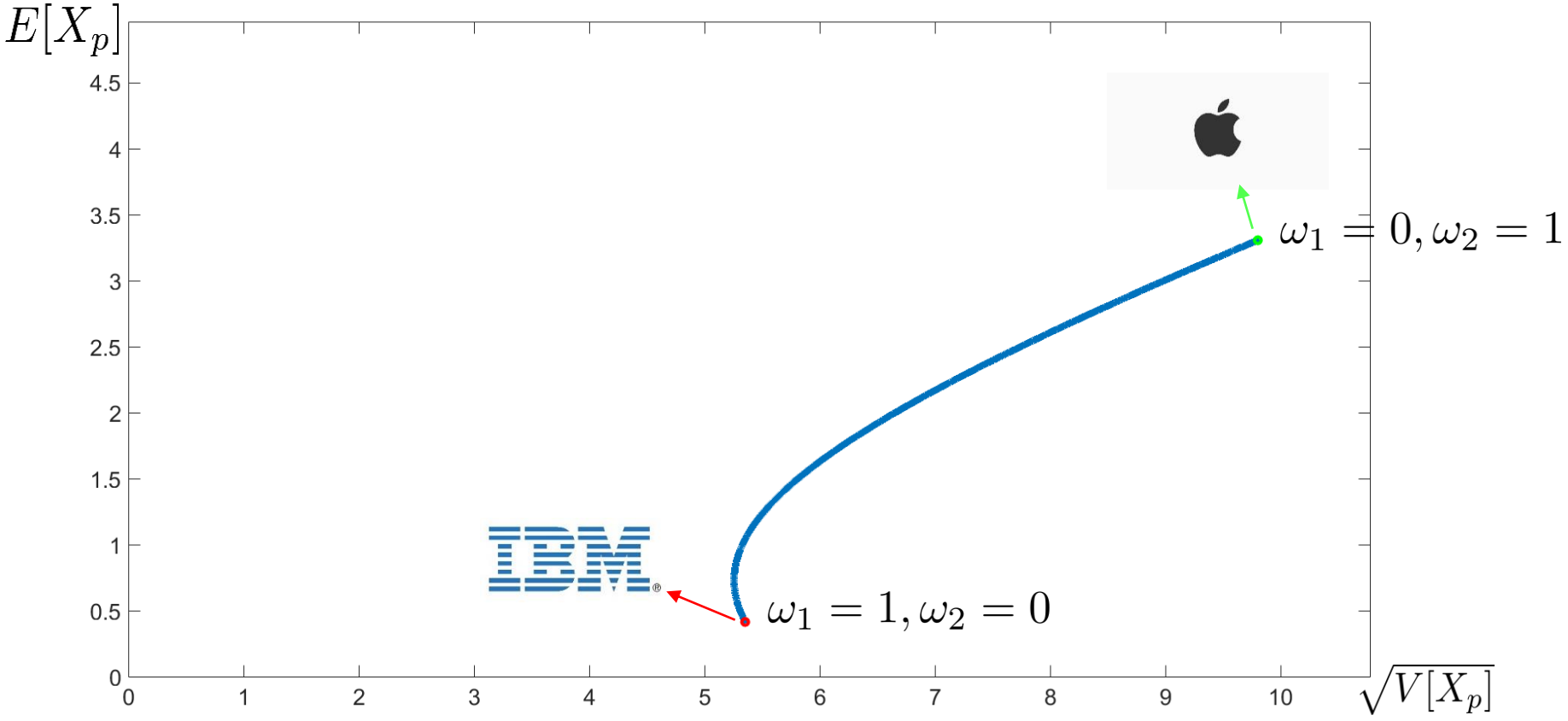
$\hat{\sigma}^2 = 28.65$ $\hat{\sigma}^2 = 96.06$

$\mathrm{Cov}(App, IBM) = 19.31$

Financial portfolio with 2 assets:

Variance of the portfolio:

$$V[X_p] = \omega_1^2(V[X_1] + V[X_2] - 2\mathrm{Cov}(X_1, X_2)) + 2\omega_1(\mathrm{Cov}(X_1, X_2) - V[X_2]) + V[X_2]$$

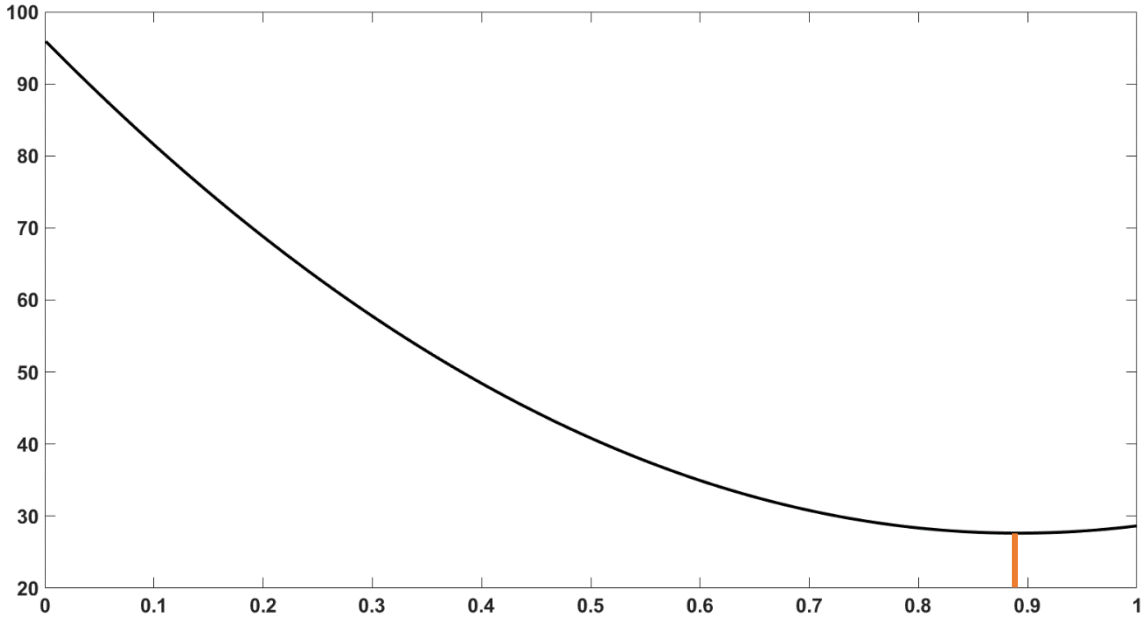


Application: Minimum variance portfolio

Financial portfolio with 2 assets:

Minimum variance:

$$\begin{aligned} V[X_p] &= \omega_1^2(V[X_1] + V[X_2] - 2\text{Cov}(X_1, X_2)) + 2\omega_1(\text{Cov}(X_1, X_2) - V[X_2]) + V[X_2], \\ &= \omega_1^2(96.06 + 28.65 - 2(19.31)) + 2\omega_1(19.31 - 96.06) + 96.06, \\ &= 86.09\omega_1^2 - 153.50\omega_1 + 96.06 \end{aligned}$$



$$\begin{aligned} \hat{\omega} &= \frac{V[X_2] - \text{Cov}(X_1, X_2)}{V[X_1] + V[X_2] - 2\text{Cov}(X_1, X_2)}, \\ &= \frac{76.75}{86.06}, \\ &= 0.89 \end{aligned}$$

Application: Minimum variance portfolio

Financial portfolio with N assets:

$$X_p = \sum_{i=1}^{N} \omega_i X_i \text{ with } \sum_{i=1}^{N} \omega_i = 1 \; \forall i \in [1, N]$$

Portfolio expected return: $E[X_p] = \sum_{i=1}^{N} \omega_i E[X_i] = \sum_{i=1}^{N} \omega_i \mu_i$

Portfolio variance:

$$V[X_p] = V[\underbrace{(\omega_1 \quad \omega_2 \quad \dots \quad \omega_N)}_{\omega'} \underbrace{\begin{pmatrix} X_1 \\ X_2 \\ \dots \\ X_N \end{pmatrix}}_{X}]$$

$$= \omega' V[X] \omega \quad (= \omega' \Sigma \omega)$$

Reminder: $V(X) = E\big((X - E(X))(X - E(X))'\big)$

Application: Minimum variance portfolio

Risk minimization:

$$\hat{\omega} = (\hat{\omega}_1, ..., \hat{\omega}_N)' = \text{argMin}_\omega V[X_p] \quad (= \omega' \Sigma \omega) \text{ such that } \sum_{i=1}^{N} \omega_i = 1$$

Solution: $\hat{\omega} = \frac{\Sigma^{-1}\mathbb{1}}{\mathbb{1}'\Sigma^{-1}\mathbb{1}}$ with $\mathbb{1} = \underbrace{(1, 1, \ldots, 1)'}_{N}$

Optimal portfolio requires an estimation of the variance

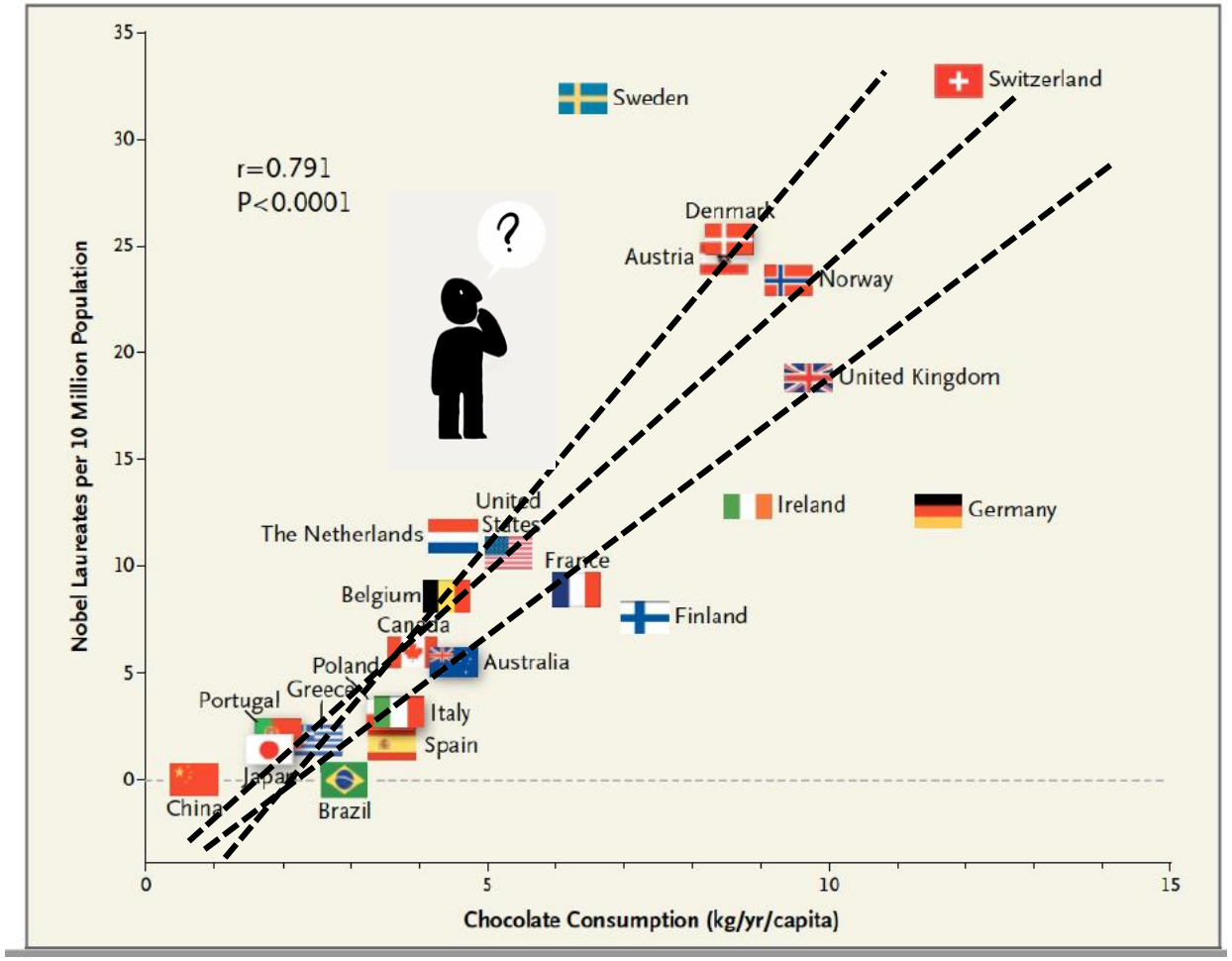
Linear regression with one explanatory variable

Mathematical framework

(see also Brooks, chapter 3)

Linear regression

"Eating chocolate produces more Nobel prize winners" (link)



Goals of the linear regression:

- Drawing a line between the variables

$$\text{Nobel} = \beta_1 + \beta_2 \# \text{ Chocolate}$$

- Statistical testing for specific values of the intercept and the slope:

$$\beta_1 = 0 \; ? \qquad \beta_2 = 0.5 \; ?$$

- Predict the dependent variable:

Predicted number of laureates for Marocco?

Linear regression in a nutshell

Other terminology

A sample

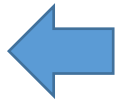
Notation: $\{y_t, x_{t,1}, \ldots, x_{t,K}\}_{t=1}^{T}$

- Statistical model typically relates variables to each others

Variable of interest

Dependent variable

Notation: y_t



Explanatory variables

Regressors

Independent variables

Explanatory variables

Notation: $x_{t,1}, \ldots, x_{t,K}$

- Linear regressions relate variables linearly:

$$y_t = \beta_1 + \beta_2 x_{t,2} + \ldots + \beta_K x_{t,K} + \epsilon_t$$

Additional terminologies

Regression coefficient

Notation: $\beta_1, \beta_2, \ldots, \beta_K$

Coefficient estimate

Notation: $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K$

Coefficient estimator

Notation: $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K$

Example: $\hat{\beta}_1 = \frac{1}{2}(y_1 + y_2)$

Random variables

- Dependent and explanatory variables are observed *at random*

  → No control over their values

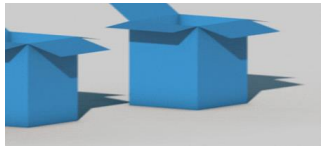
- We postulate a linear relation between the explanatory and the dependent variable:

$$y_t = \beta_1 + \beta_2 x_{t,2} + \ldots + \beta_K x_{t,K} + \epsilon_t$$

Error term
=
What is not explained by the linear function

Conclusion:

- The explanatory variable is assumed to be a random variable
- The error term is a random variable
- The dependent variable is a random variable: $y_t = f(x_{t,2}, \ldots, x_{t,K}, \epsilon_t)$



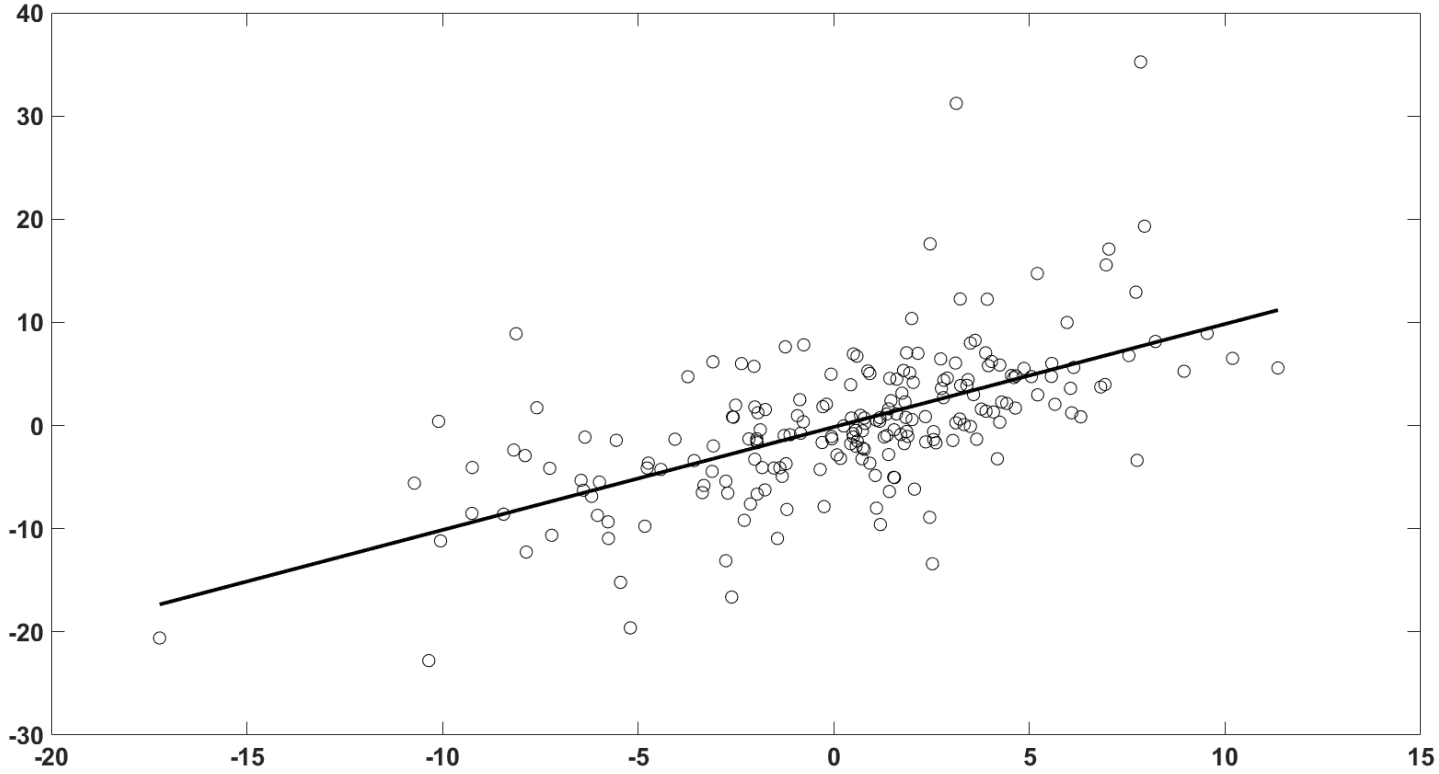
Linear regression

Capital asset pricing model:

$$y_t = \beta_1 + \beta_2 x_t + \epsilon_t$$

- Drawing a line between the variables

$$y_t = -0.138 + 0.998 x_t + \epsilon_t$$



Dependent variable

y_t: excess returns of a financial asset (IBM).

Explanatory variable

x_t : excess return of the market.

- Test Hypothesis for the beta market

$$H_0 : \beta_2 = 1$$

- IBM Return if market equals 2% ?

Determining the line

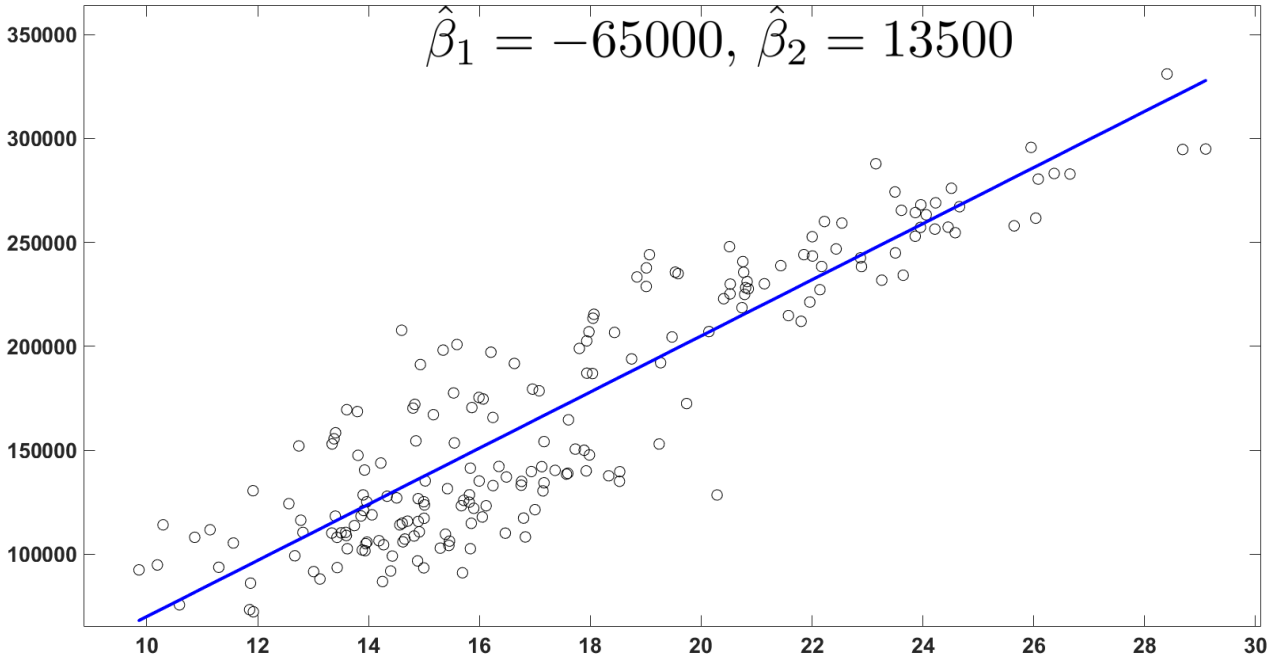
Dependent variable

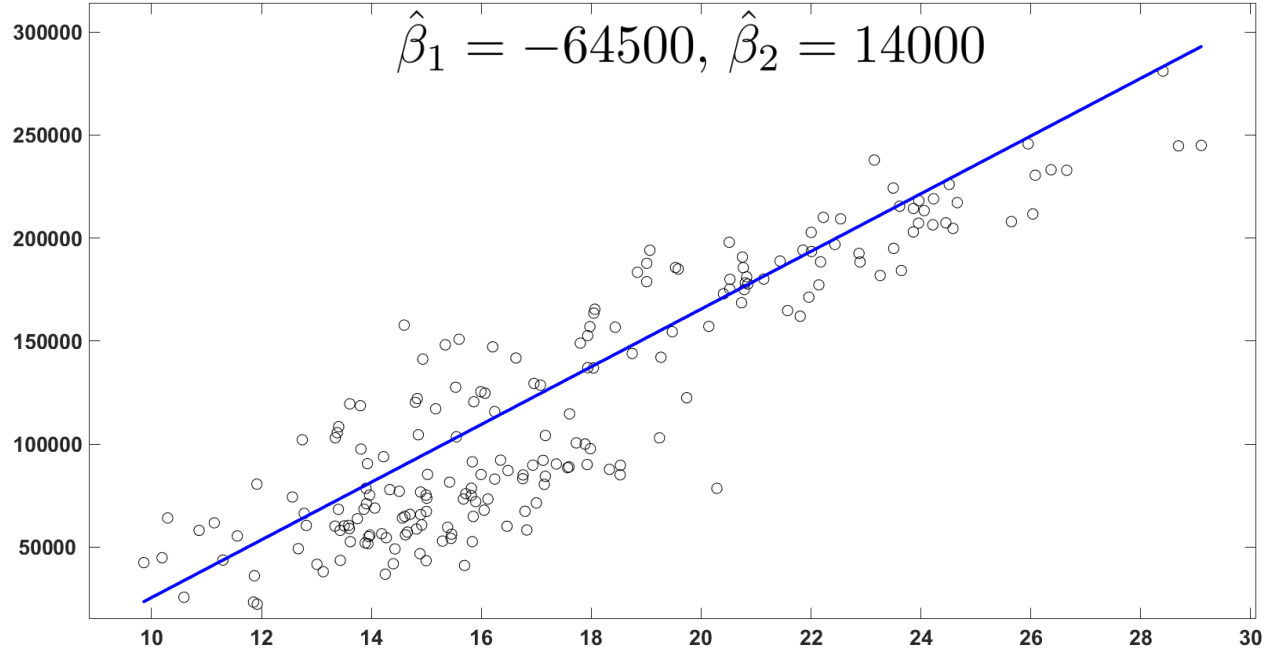
y_t: Salary of the worker.

Explanatory variable

x_t : experience.

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$





Which one fits the best ?

Choosing the line

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

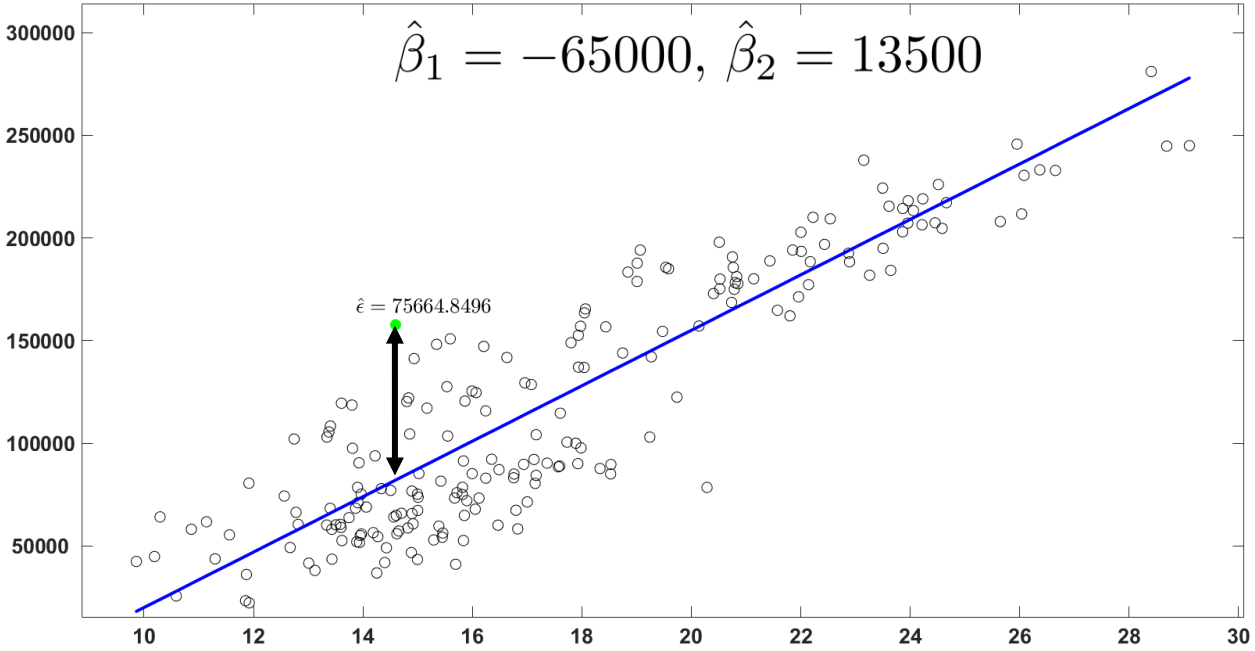
Dependent variable

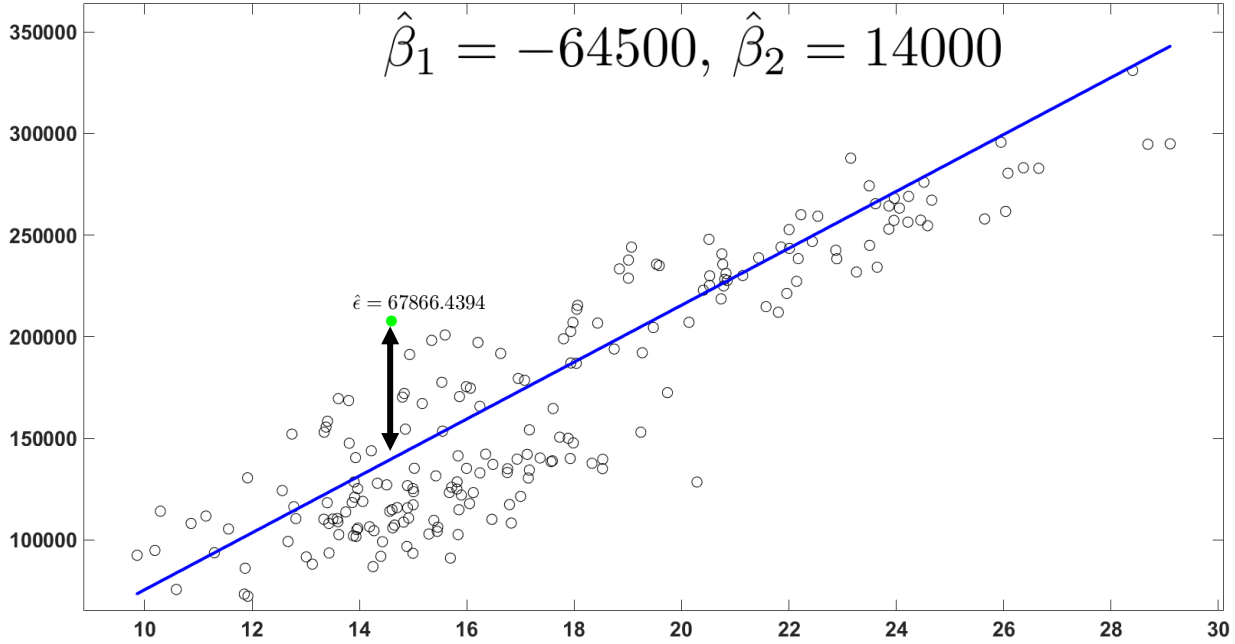
y_t: Salary of the worker.

Explanatory variable

x_t : experience.

- Computing the errors: $\hat{\epsilon}_t = y_t - \hat{\beta}_1 - \hat{\beta}_2 x_t$





Choosing the line

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

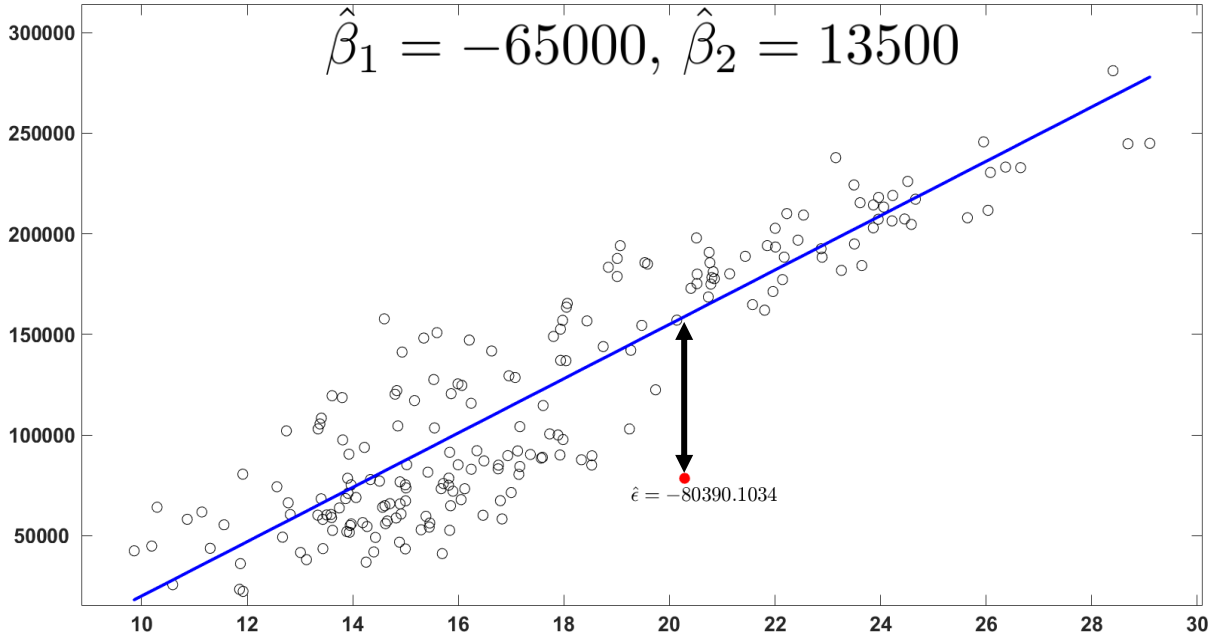
Dependent variable

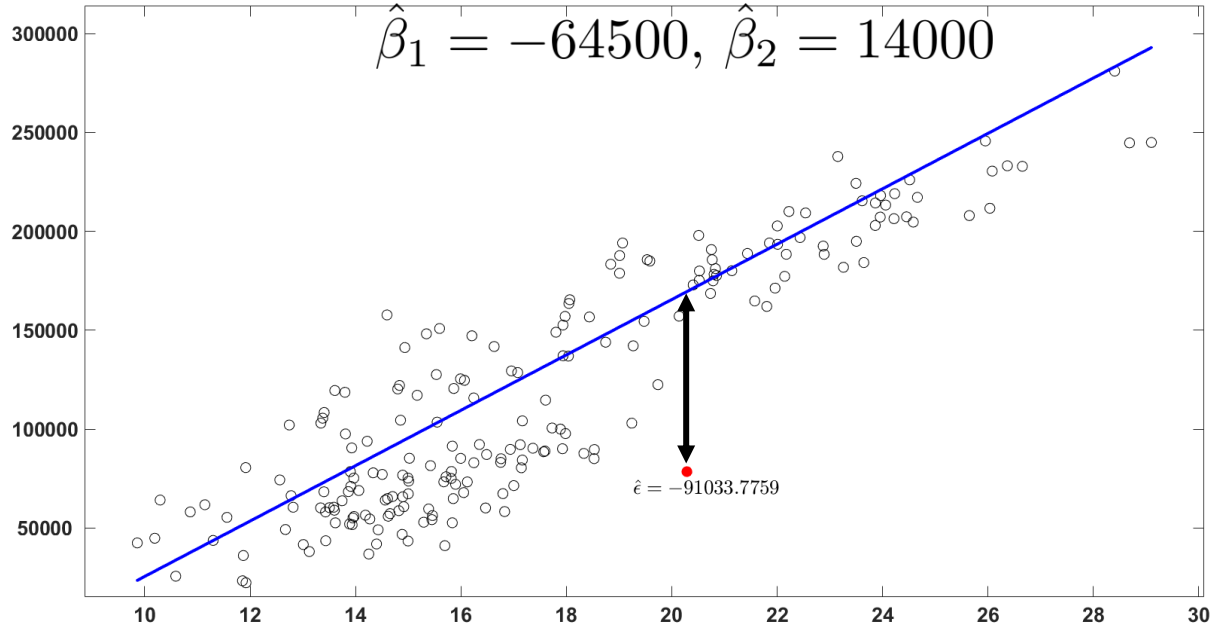
y_t: Salary of the worker.

Explanatory variable

x_t : experience.

- Computing the errors: $\hat{\epsilon}_t = y_t - \hat{\beta}_1 - \hat{\beta}_2 x_t$





Choosing the line

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

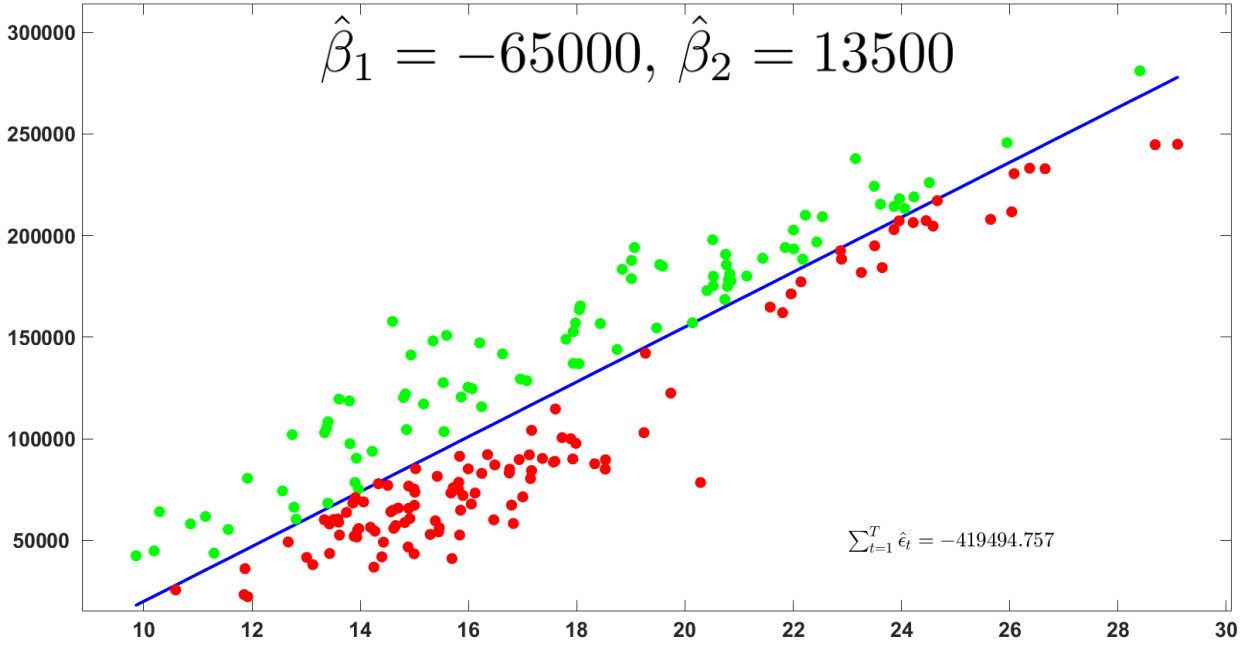
Dependent variable

y_t: Salary of the worker.

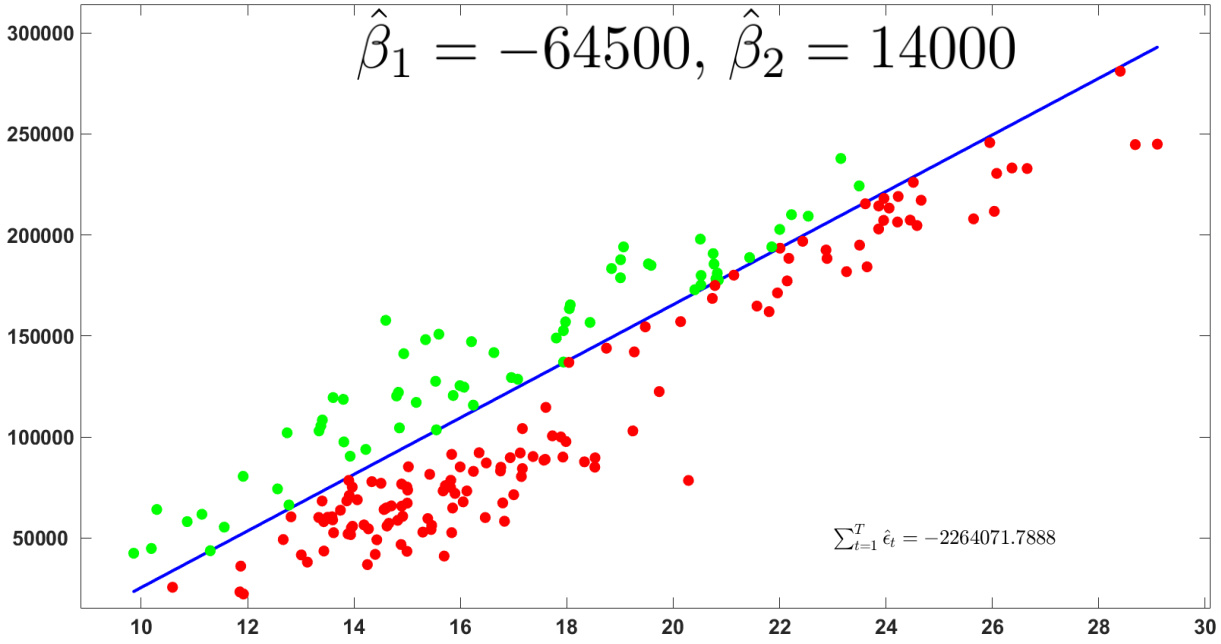
Explanatory variable

x_t : experience.

- Computing the errors: $\hat{\epsilon}_t = y_t - \hat{\beta}_1 - \hat{\beta}_2 x_t$



Errors closer to zero



Errors more negative

Choosing the line

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Dependent variable

y_t: Salary of the worker.

Explanatory variable

x_t : experience.

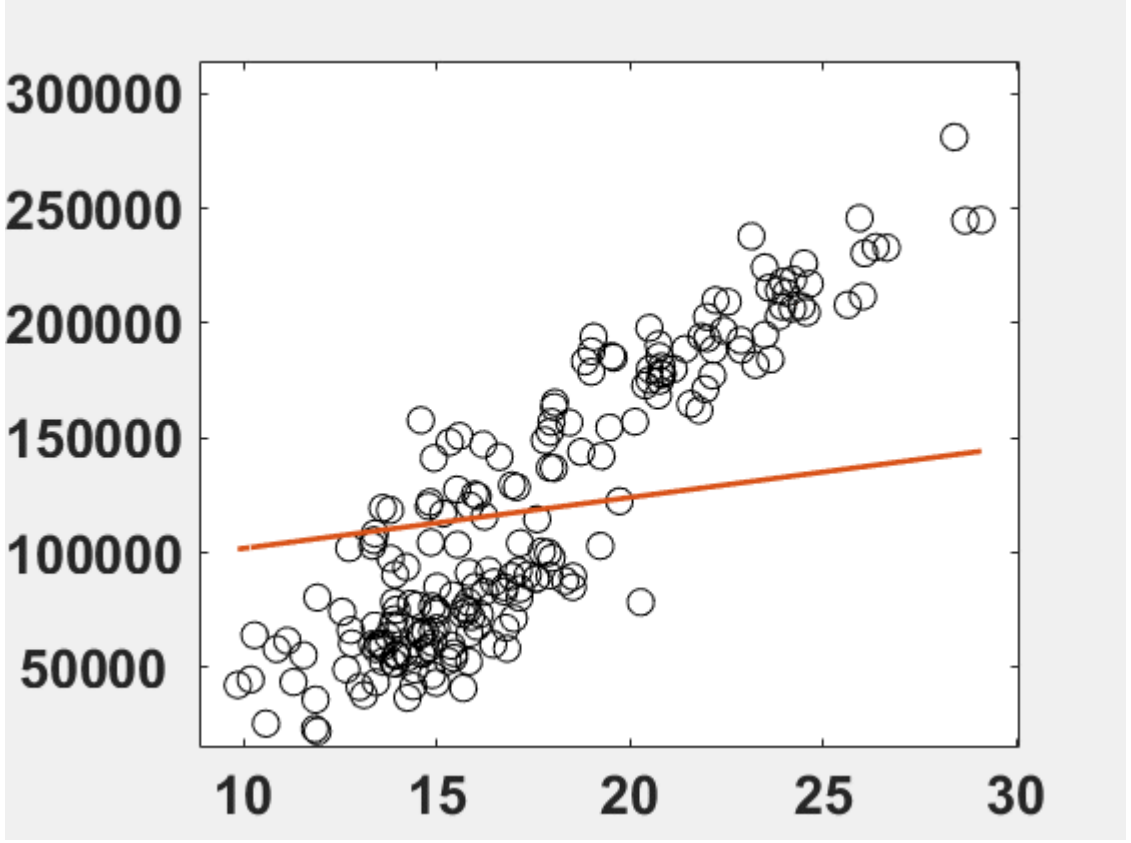
- First algorithm to find the parameter values:

1. Find the parameter values that lead to a sum of the error terms equal to zero: $\sum_{t=1}^{T} \hat{\epsilon}_t = 0$

$$\sum_{t=1}^{T} \hat{\epsilon}_t = 0$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

Identification problem



Choosing the line

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Possible criterion:

$(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} |\epsilon_t|$

Pierre-Simon Laplace (1745-1827)

Possible criterion:

$(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \epsilon_t^2$

Friedrich Gauss (1777-1855)

Choosing the line

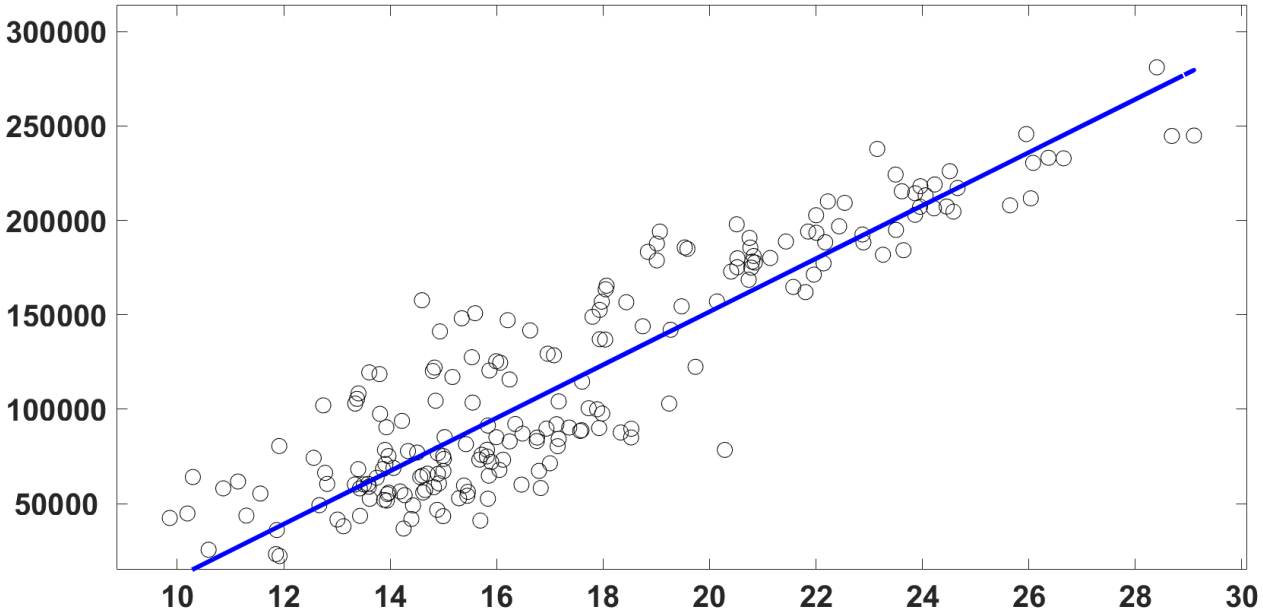
Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

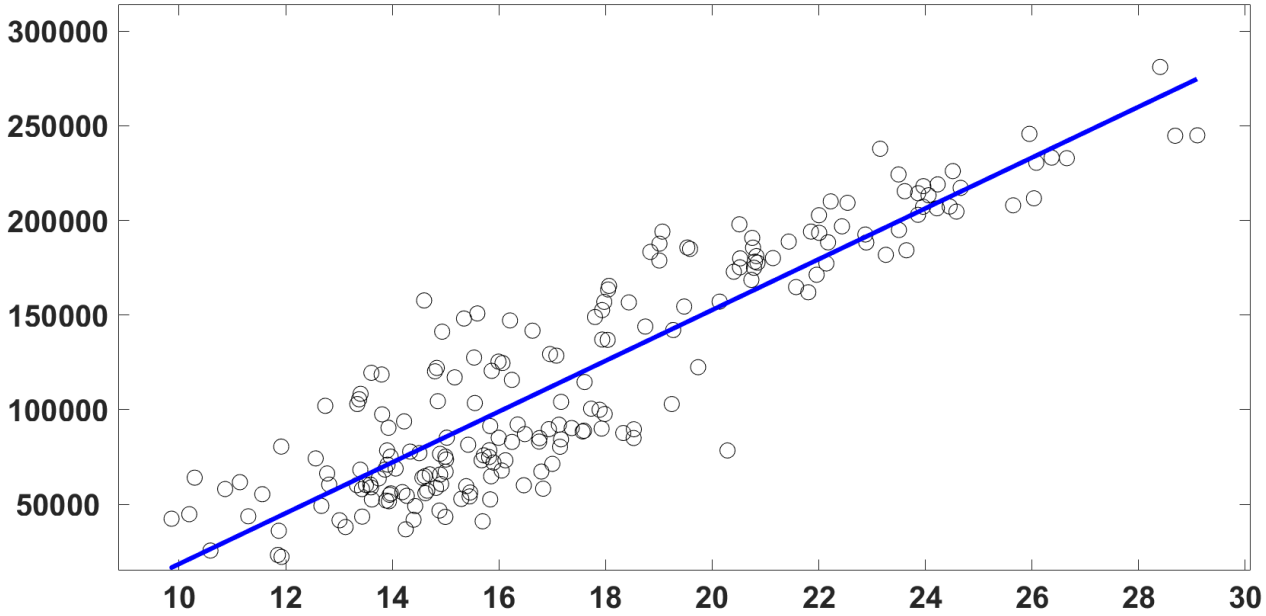
Possible criterion:

$(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} |\epsilon_t|$

Possible criterion:

$(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \epsilon_t^2$



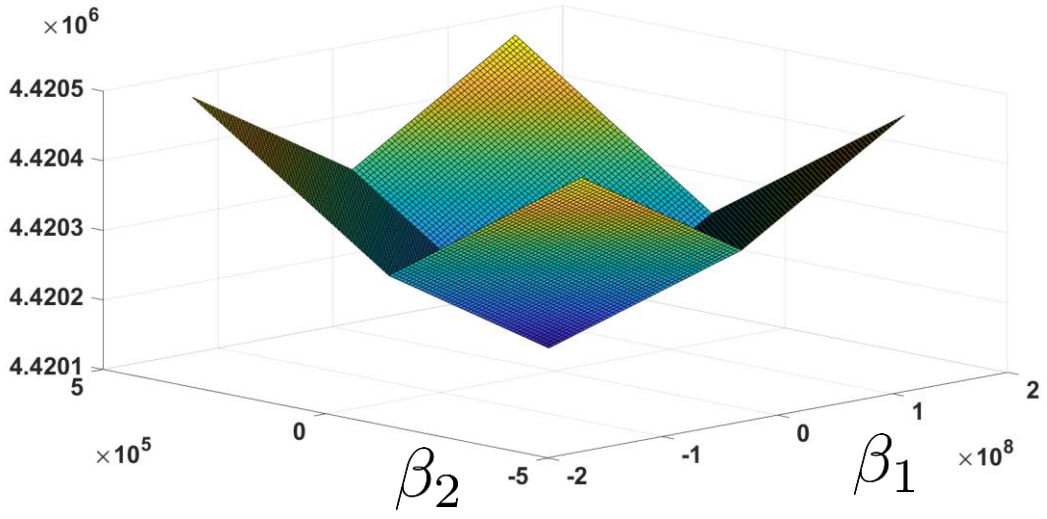


Choosing the line

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Possible criterion:

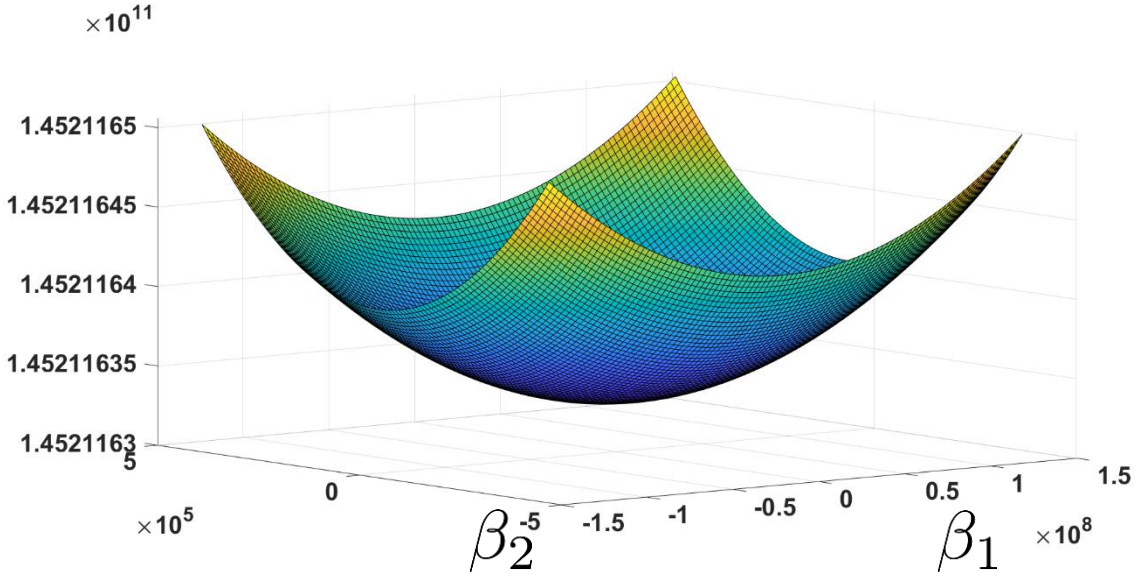
$(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} |\epsilon_t|$



- Difficult to find the optimum.
- No analytical form of the estimator.
- Statistical properties are less studied.

Possible criterion:

$(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \epsilon_t^2$



- Convex function.
- Analytical formula for the estimator.
- Statistical properties are well-known.

Analytical formula

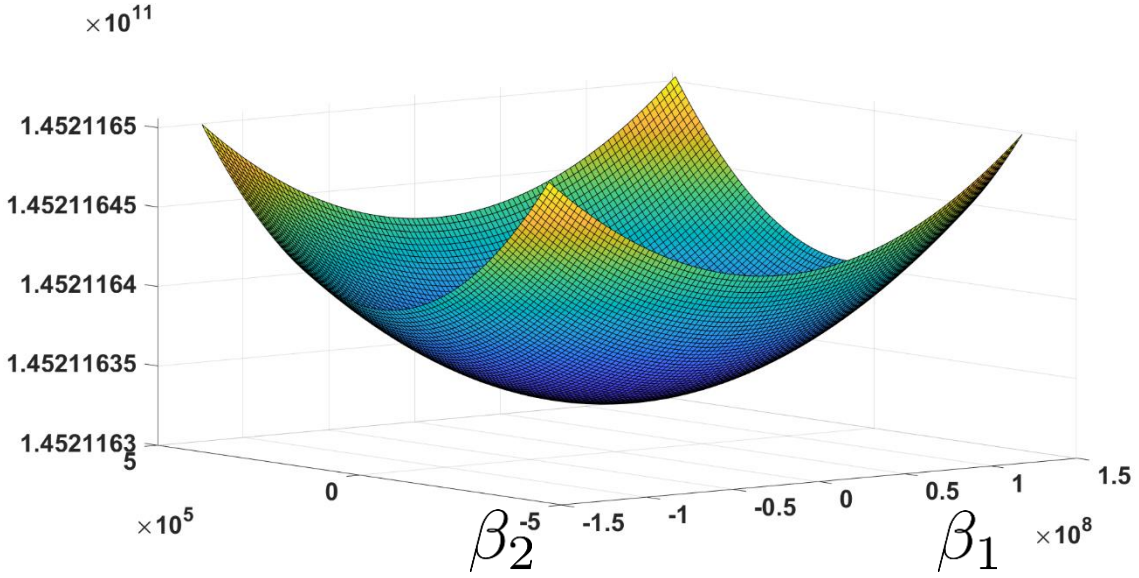
Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Criterion: $(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \epsilon_t^2$

Standard criterion:

$$(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \epsilon_t^2$$

Sum of squared residuals (SSR)
Residual sum of squares (RSS)



Mathematical development:

$$(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \epsilon_t^2$$

$$= \text{argmin}_{\hat{\beta}_1,\hat{\beta}_2} \sum_{t=1}^{T} (y_t - \beta_1 - \beta_2 x_t)^2$$

Ordinary least squares (OLS) estimators

$$\hat{\beta}_2 = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}.$$

OLS computation

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Criterion: $(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \epsilon_t^2$

Capital asset pricing model:

Year, t	Excess returns on asset Thor	Excess return on the market index
1	17.8	13.7
2	39.0	23.2
3	12.8	6.9
4	24.2	16.8
5	17.2	12.3

Analytical formulas:

$$\hat{\beta}_2 = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}.$$

Thor (y_t)	Market (x_t)	$(y_t - \bar{y})$	$(x_t - \bar{x})$	$(y_t - \bar{y})(x_t - \bar{x})$	$(x_t - \bar{x})^2$
17.8	13.7	-4.4	-0.88	3.872	0.7744
39	23.2	16.8	8.62	144.816	74.3044
12.8	6.9	-9.4	-7.68	72.192	58.9824
24.2	16.8	2	2.22	4.44	4.9284
17.2	12.3	-5	-2.28	11.4	5.1984

$$\hat{\beta}_2 = \frac{236.72}{144.188} = 1.6417 (\approx 1.64)$$

$$\hat{\beta}_1 = 22.20 - 1.64\ (14.58) = -1.7366 (\approx -1.74).$$

OLS computation

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Criterion: $(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \epsilon_t^2$

Analytical formulas:

$$\hat{\beta}_2 = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{T}(x_t - \bar{x})^2} = 1.64$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} = -1.74.$$

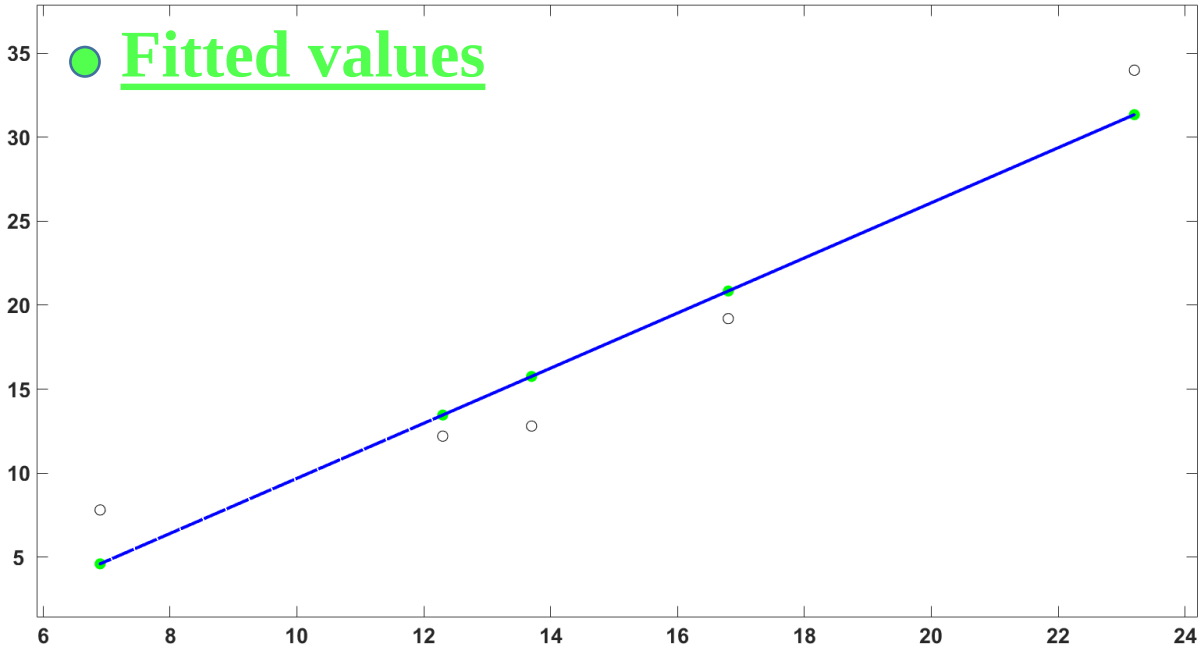
$$\hat{y}_t = \hat{\beta}_1 + \hat{\beta}_2 x_t$$

Year, t	Thor	index
1	17.8	13.7
2	39.0	23.2
3	12.8	6.9
4	24.2	16.8
5	17.2	12.3

Additional terminologies

$\hat{y}_1, \ldots, \hat{y}_T$

Fitted values



OLS interpretation

Capital asset pricing model: $\hat{y}_t = -1.74 + 1.64x_t$

1. Beta of 1.64 → Risky asset
2. « If the excess-market increases by one percent, asset Thor is expected to increase by 1.64, everything else being equal ».

Prediction:

1. If Excess-market equals to 2 → Expected Thor return: 1.54
2. If Excess-market equals to 0 → Expected Thor return: -1.74

Prediction outside the range of the data!

Example: No negative excess market return values in the data

Unlikely to have a good prediction!

Constant regression

What is the best estimator for the constant regression ?

Constant regression: $y_t = \beta_1 + \epsilon_t$

Criterion: $\hat{\beta}_1 = \text{argmin}_{\beta_1} \sum_{t=1}^{T} \epsilon_t^2 = \text{argmin}_{\beta_1} \sum_{t=1}^{T} (y_t - \beta_1)^2$

$\hat{\beta}_1 = \frac{1}{T} \sum_{t=1}^{T} y_t = \bar{y}$

Sample average is the OLS estimator of the constant regression

Additional terminologies

$\sum_{t=1}^{T} (y_t - \bar{y})^2$

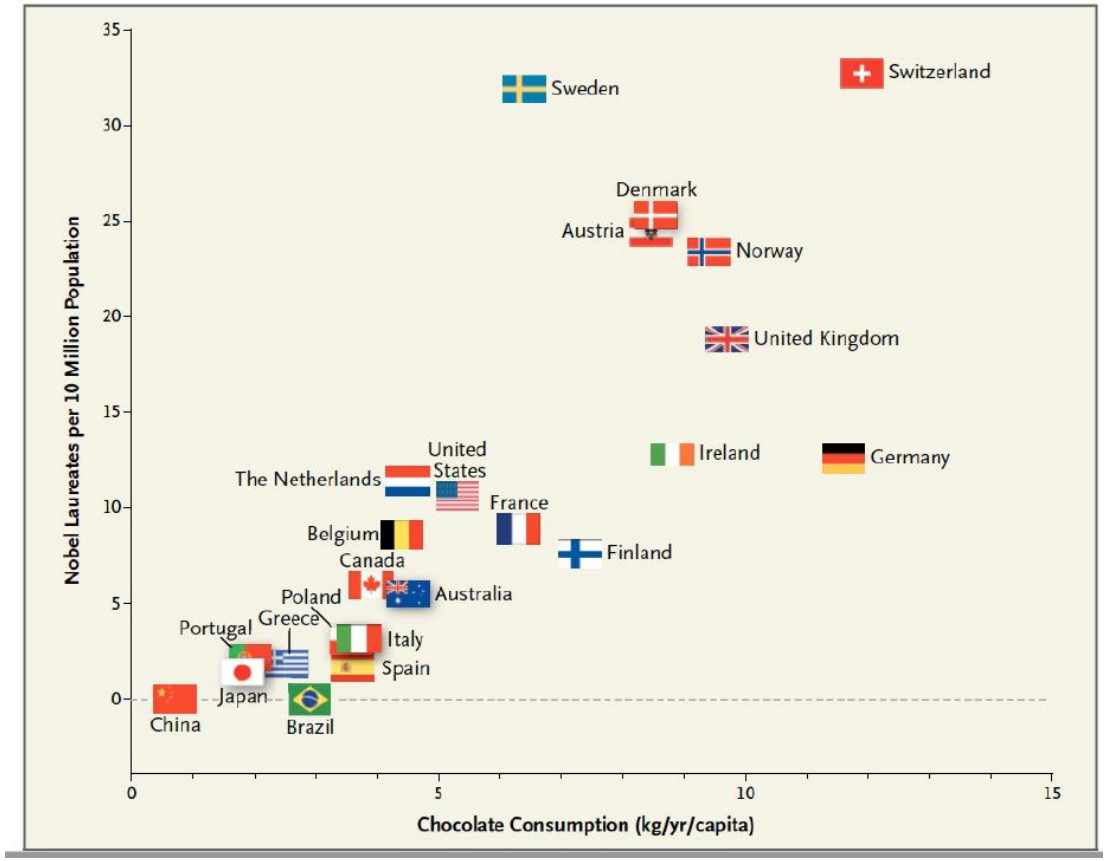
Total sum of squares (TSS)

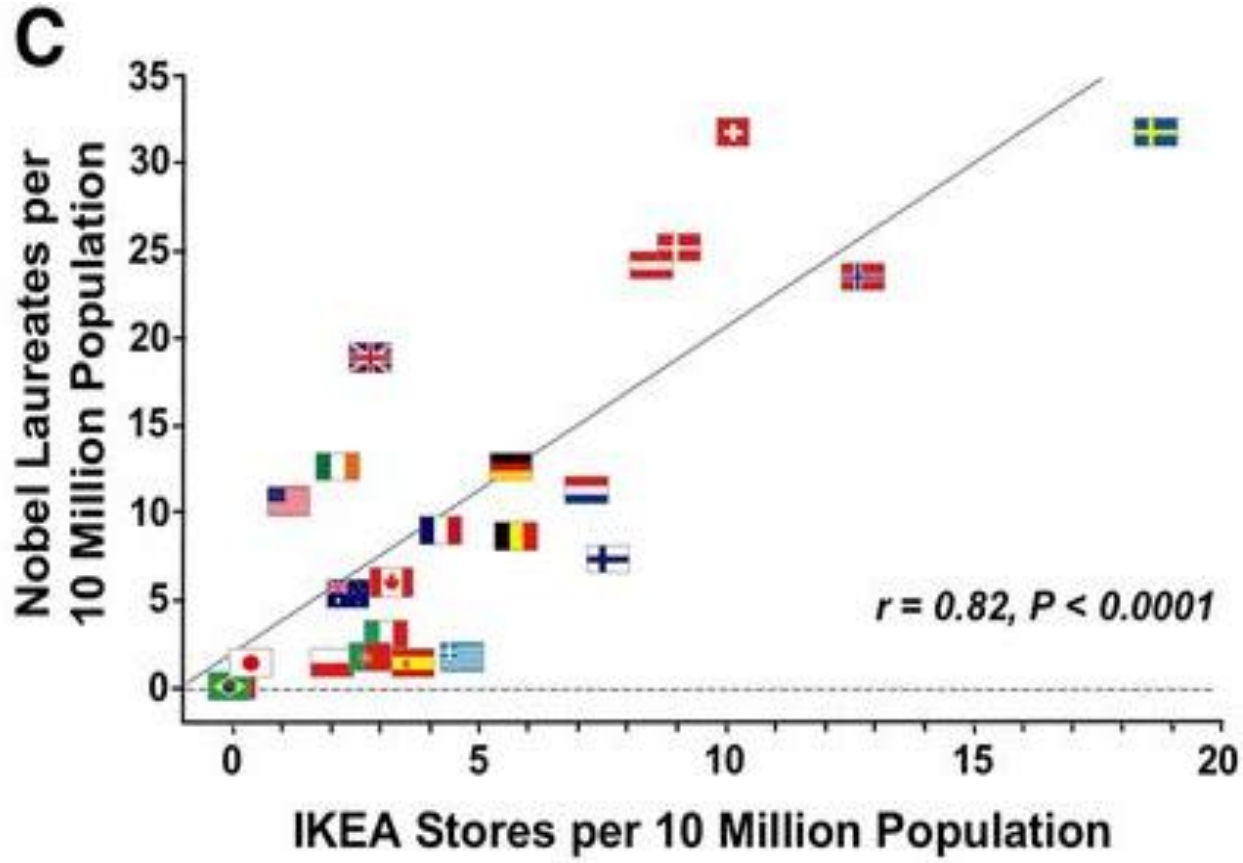
Fitting criterion

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Criterion: $(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \epsilon_t^2$

A measure to see how the regression fits





J Nutr, Volume 143, Issue 6, June 2013, Pages 931–933, https://doi.org/10.3945/jn.113.174813

Which one fits the best ?

Fitting criterion

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Criterion: $(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \epsilon_t^2$

A measure to see how the regression fits

Intuition: Smaller the SSR, the best it is.

Problem: SSR are sensitive to the scale of the dependent variable.

➡ Relative criterion.

We need a criterion independent of the scale of the variables.

Coefficient of determination: $R^2 = 1 - \frac{\sum_{t=1}^{T} \hat{\epsilon}_t^2}{\sum_{t=1}^{T} (y_t - \bar{y})^2}$

- Coefficient of determination between 0 and 1.

By how much do we improve the fit compared to the sample average ?

Alternatively

" $100R^2$ percent of the variation in *y* is reduced by taking into account predictor *x*"

Correlation

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Criterion: $(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \epsilon_t^2$

Relation with the correlation coefficient:

Correlation: measures how two variables are linearly dependent.

Also a good candidate for determining the fit of the regression!

Empirical correlation: $\rho_{\hat{y}y} = \frac{1}{T} \frac{\sum_{t=1}^{T} (\hat{y}_t - \bar{\hat{y}})(y_t - \bar{y})}{\sqrt{\frac{1}{T}\sum_{t=1}^{T} (\hat{y}_t - \bar{\hat{y}})^2 \quad \frac{1}{T}\sum_{t=1}^{T} (y_t - \bar{y})^2}} \in [-1, 1]$

- Simple regression: $\rho_{\hat{y}y} = \rho_{xy}$

Coefficient of determination is equal to the squared empirical correlation

$$R^2 = \rho_{\hat{y}y}^2$$

Fitting criterion

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Criterion: $(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \epsilon_t^2$

Capital asset pricing model:

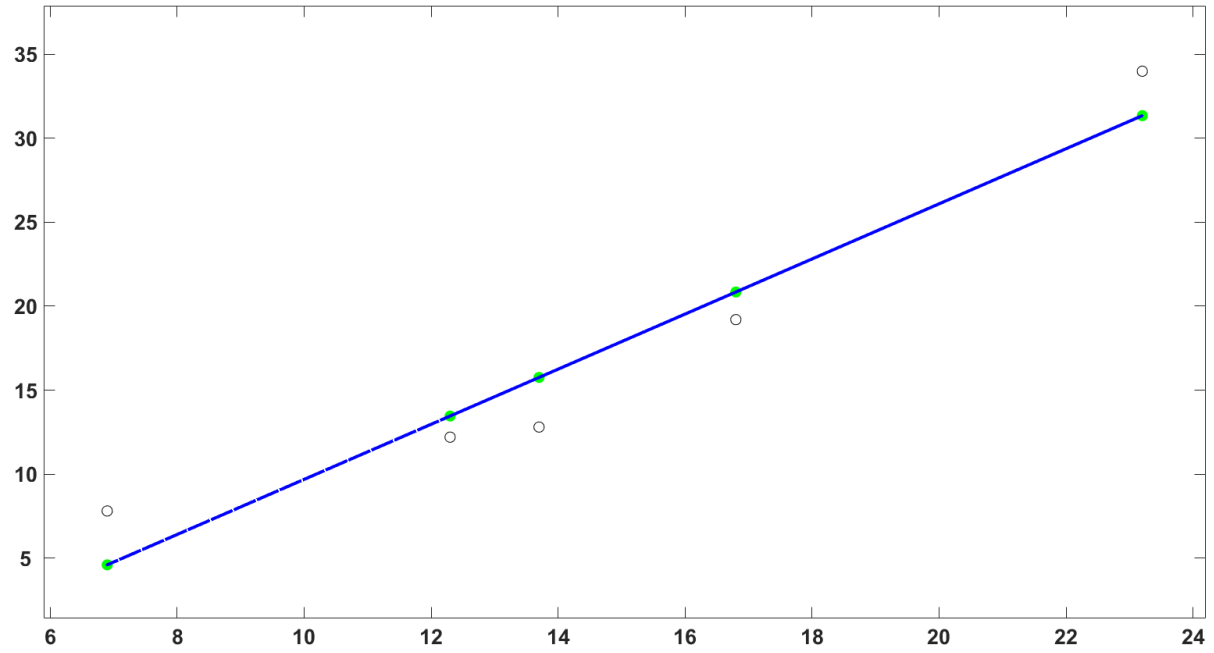
Year, t	Excess returns on asset Thor	Excess return on the market index
1	17.8	13.7
2	39.0	23.2
3	12.8	6.9
4	24.2	16.8
5	17.2	12.3

Fitted values: $\hat{y}_t = -1.74 + 1.64 x_t$

Years, t	$(y_t - \bar{y})$	$\hat{\epsilon}_t = (y_t - \hat{y}_t)$
1	-4.40	-2.96
2	16.80	2.65
3	-9.40	3.21
4	2.00	-1.64
5	-5.00	-1.26

$$R^2 = 1 - \frac{\sum_{t=1}^{T} \hat{\epsilon}_t^2}{\sum_{t=1}^{T} (y_t - \bar{y})^2} = 0.93$$

Very good regression!



Linearity assumption

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Linear with respect to the parameters:

→ The explanatory variable and the dependent variable can be non linear.

Volatility modelling: $y_t = x_t^{\beta} \epsilon_t \rightarrow \ln y_t^2 = 2\beta \ln x_t + \ln \epsilon_t^2$

Cobb Douglas function: $Y = cK^{\alpha} L^{\beta} \rightarrow \ln(Y) = \ln c + \alpha \ln K + \beta \ln L$

Other example: $y_t = \beta_1 + \frac{\beta_2}{x_t} + \epsilon_t \rightarrow y_t = \beta_1 + \beta_2 \underbrace{z_t}_{=\frac{1}{x_t}} + \epsilon_t$

Non-linear regression: $y_t = \beta_1 + \beta_2 x_t^{\beta_3} + \epsilon_t$

GDP – life expectancy example

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Criterion: $(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \epsilon_t^2$

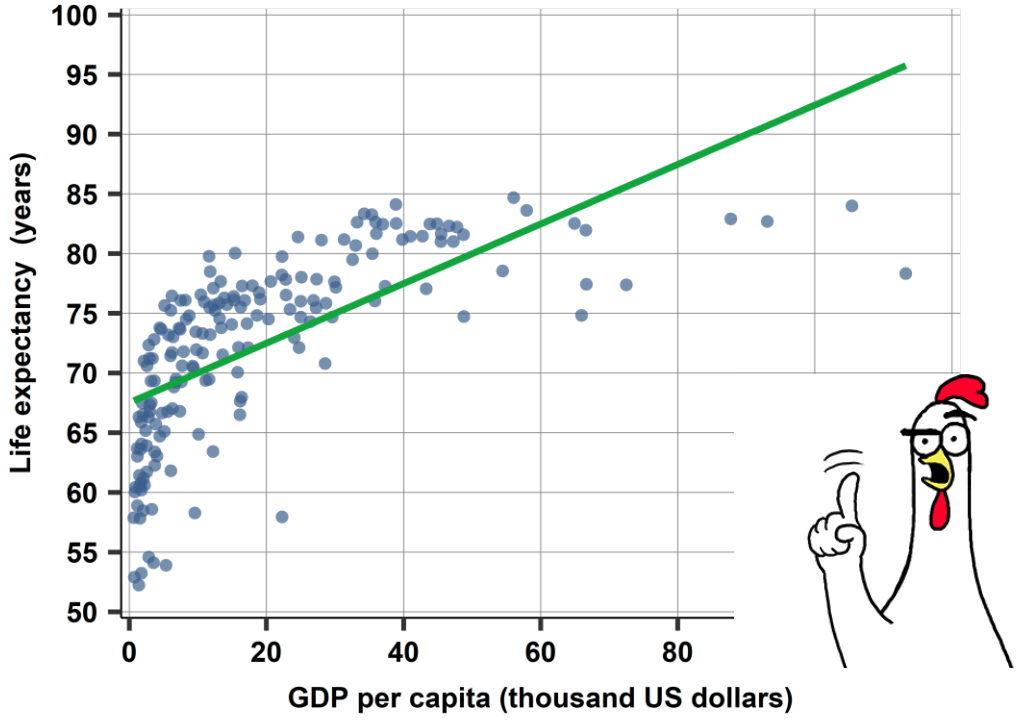
Dependent variable

y_t: Life expectancy.

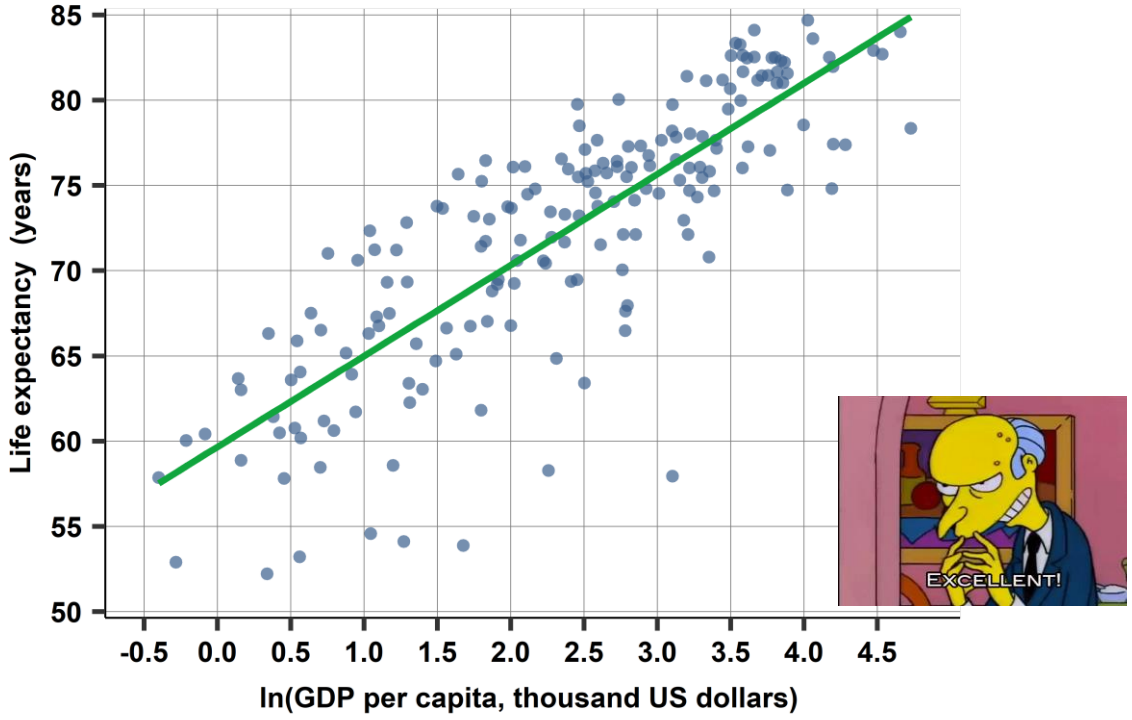
Explanatory variable

x_t : GDP per capita.

$$y_t = \beta_1 + \beta_2 x_t + \epsilon_t$$



$$y_t = \beta_1 + \beta_2 \ln x_t + \epsilon_t$$



Interpretation

Taylor approx: $\ln(x_1) \approx \ln x_0 + \frac{(x_1 - x_0)}{x_0}$

Dependent variable

y_t: Life expectancy.

Explanatory variable

x_t : GDP per capita.

level-level: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t \xrightarrow{x_t + 1} \beta_1 + \beta_2(x_t + 1) + \epsilon_t = y_t + \beta_2$

An additional unit adds the slope to the dependent variable

level-log: $y_t = \beta_1 + \beta_2 \ln x_t + \epsilon_t \xrightarrow{1.01x_t} \beta_1 + \beta_2(\ln x_t + \frac{1.01x_t - x_t}{x_t}) + \epsilon_t = y_t + 0.01\beta_2$

1% additional unit adds 1% of the slope to the dependent variable

log-log: $\ln y_t = \beta_1 + \beta_2 \ln x_t + \epsilon_t \xrightarrow{1.01x_t} \beta_1 + \beta_2(\ln x_t + \frac{1.01x_t - x_t}{x_t}) + \epsilon_t = \ln y_t + 0.01\beta_2$

$\ln y_t + 0.01\beta_2 \approx \ln(y_t[1 + 0.01\beta_2])$

1% additional unit increases the dependent variable by 1% times the slope

Linear regression with one explanatory variable

The statistical model

Linear regression

Capital asset pricing model:

$$y_t = \beta_1 + \beta_2 x_t + \epsilon_t$$

- Drawing a line between the variables

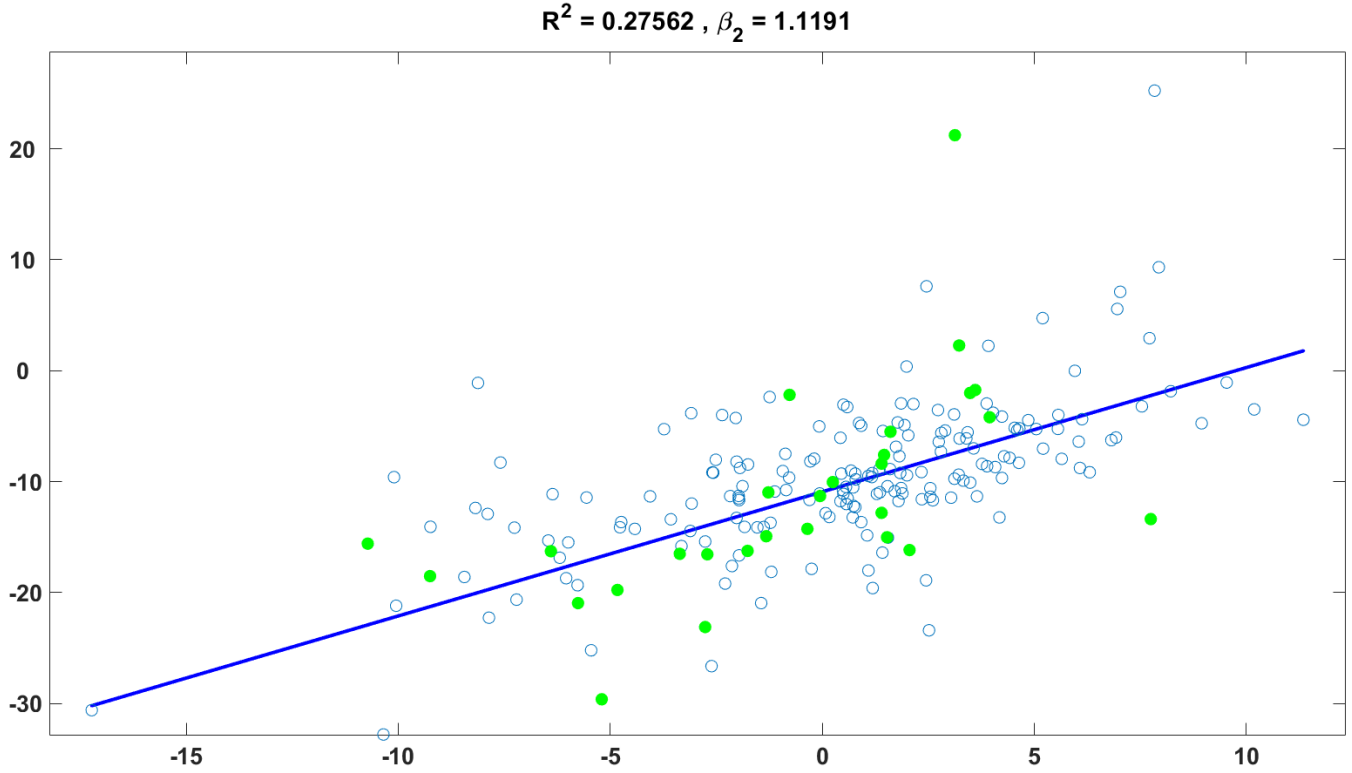
$$y_t = -0.91 + 1.12 x_t + \epsilon_t$$

Dependent variable

y_t: excess returns of a financial asset (IBM).

Explanatory variable

x_t : excess return of the market.



For this sample (T=30):

$$R^2 = 0.276,\ \beta_2 = 1.12$$

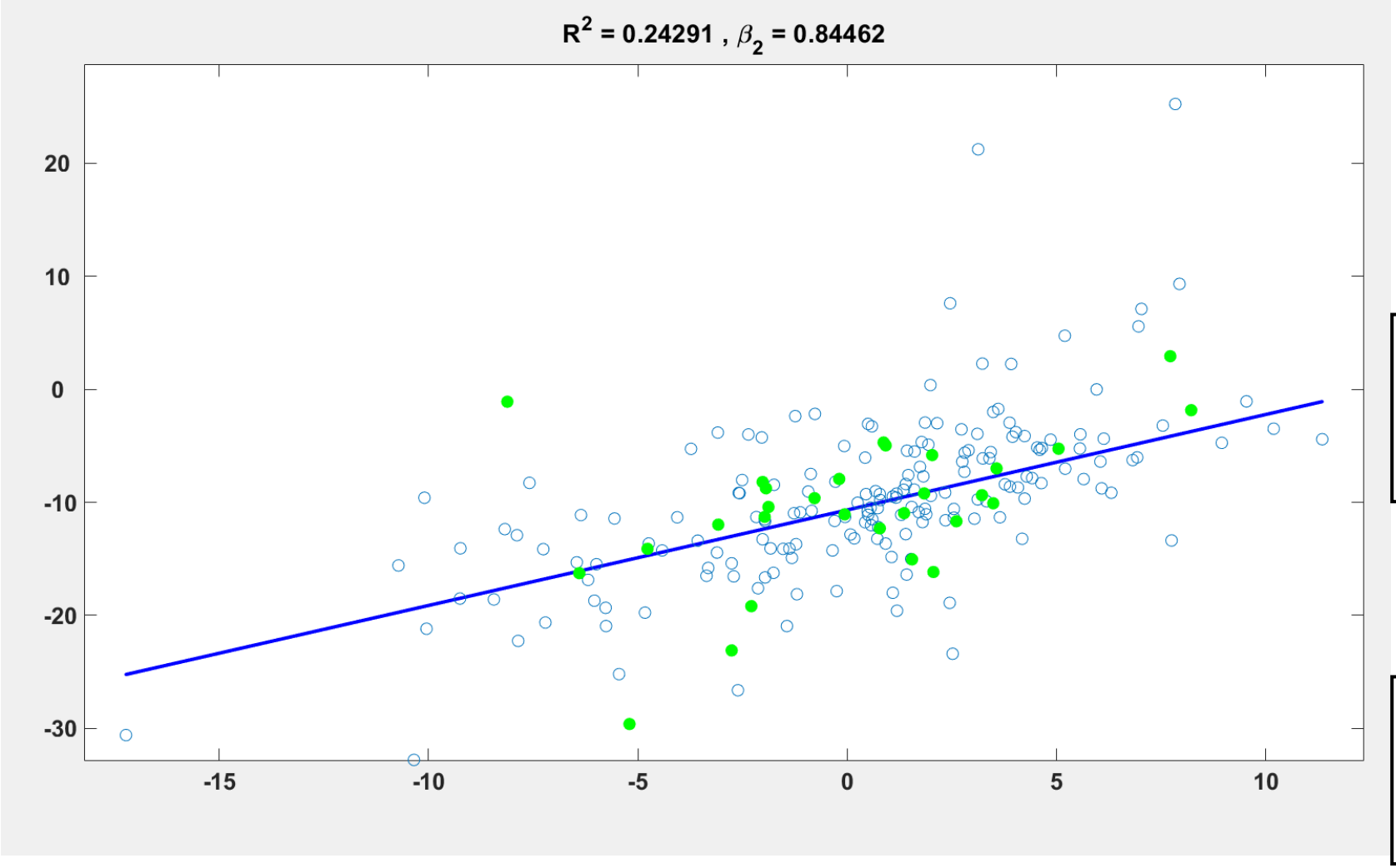
Linear regression

Capital asset pricing model:

$$y_t = \beta_1 + \beta_2 x_t + \epsilon_t$$

Dependent variable IBM

y_t: excess returns of a financial asset (IBM).



Estimates and correlation depends on the sample!

How does it vary from one sample to another?

Need additional assumptions on how data are generated.

A statistical model

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Hypothesis: (conditional on the explanatory variables)

1. **Linear regression:** The variables are linearly related.
2. **No collinearity:** The explanatory variable is not constant.
3. **Strict exogeneity:** Errors have zero expectations conditional on all the explanatory variables.
4. **White noise:** No linear dependence between the error terms.
5. **Homoskedasticity:** The variance of the error term is constant.
6. **Distribution:** Error term is normally distributed.

Why do we need these assumptions ?

Interpreting the parameter values of the regression.
Building intervals where the future value can be.
Building intervals where the true coefficient can be.

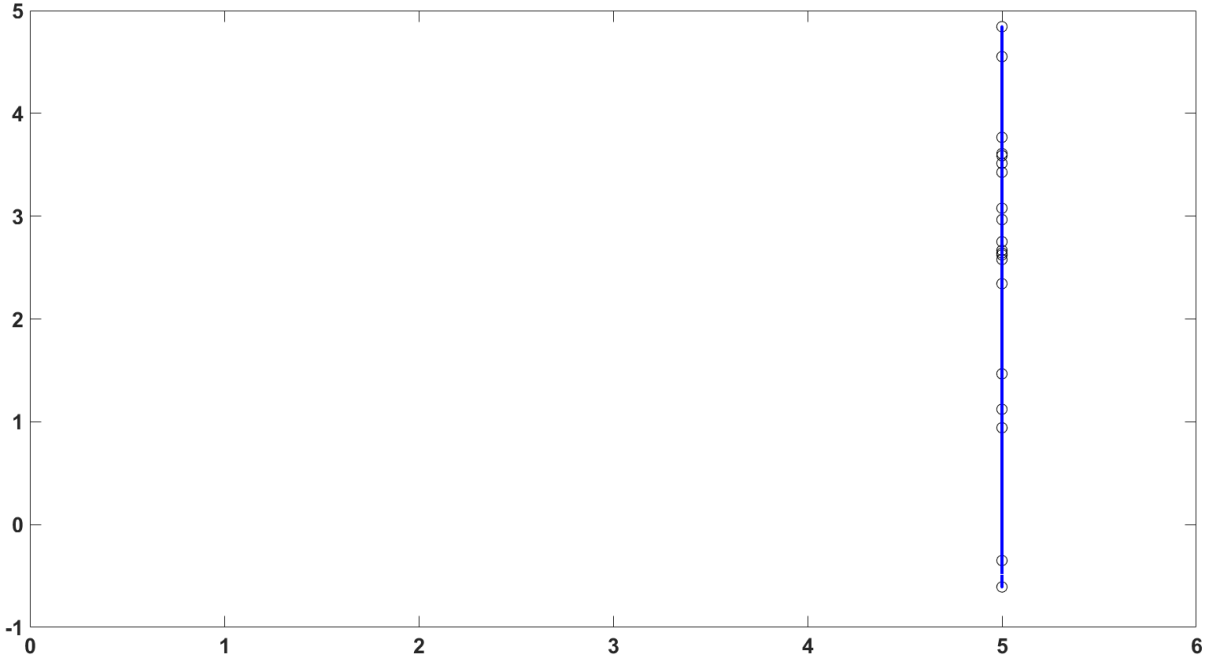
More on assumptions

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

2. No Collinearity: The explanatory variable is not constant.

If the explanatory variable is constant: $y_t = \beta_1 + \beta_2 c + \epsilon_t$

OLS estimate is not defined: $\hat{\beta}_2 = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{T}(x_t - \bar{x})^2} = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{0}$



Parallel line to the y-axis!

Note that the sample size must be larger than 1

More on assumptions

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

3. Strict exogeneity. Errors have zero expectations conditional on all the explanatory variables: $E(\epsilon_t | x_1, \ldots, x_T) = 0 \quad \forall t \in [1, T]$

→ **Errors have zero mean:** $E(\epsilon_t) = 0 \quad \forall t \in [1, T]$

Not a restrictive assumption when the regression includes a constant

If $E(\epsilon_t) = \mu$ →

$$y_t = \underbrace{(\beta_1 + \mu)}_{\gamma} + \beta_2 x_t + \underbrace{(\epsilon_t - \mu)}_{\eta_t}$$

$$= \gamma + \beta_2 x_t + \eta_t$$

→ **No linear dependence between the errors and the explanatory variable:**

$$\mathrm{Cov}(\epsilon_t, x_j) = 0 \; \forall j \in [1, T]$$

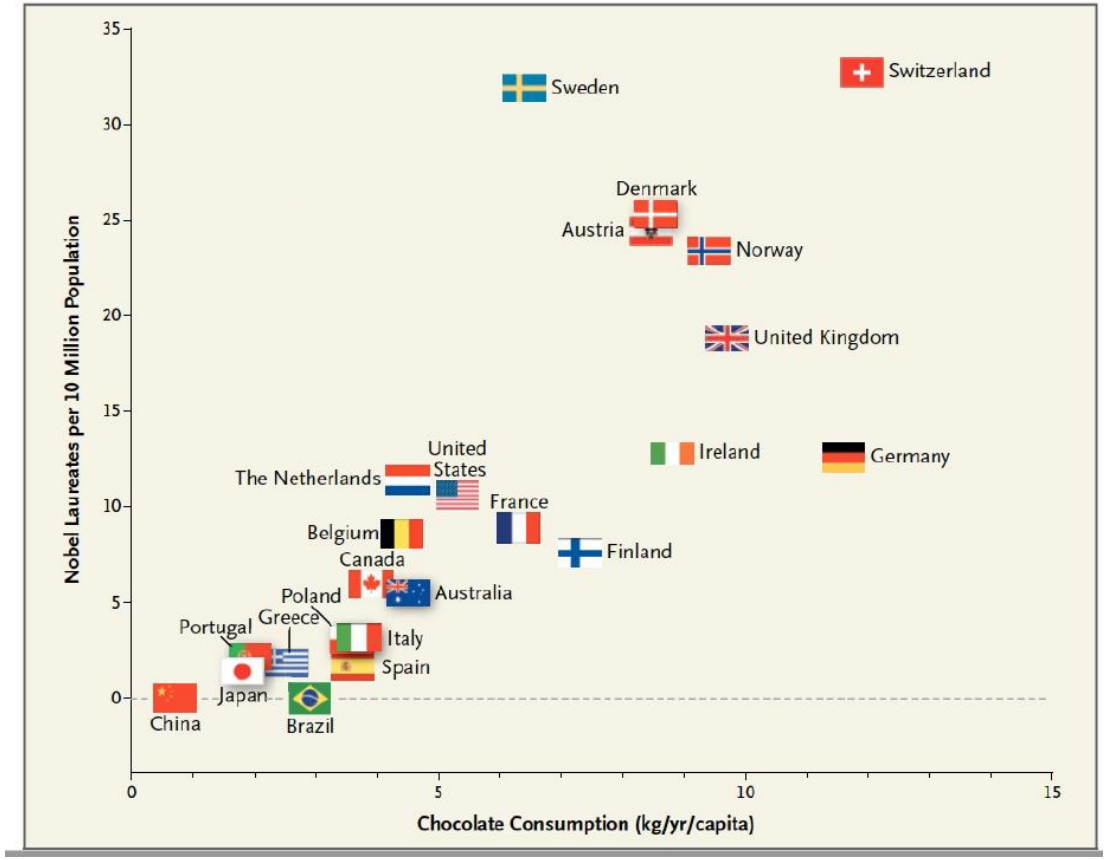
If assumption holds, we can interpret the estimated value $\hat{\beta}_2$

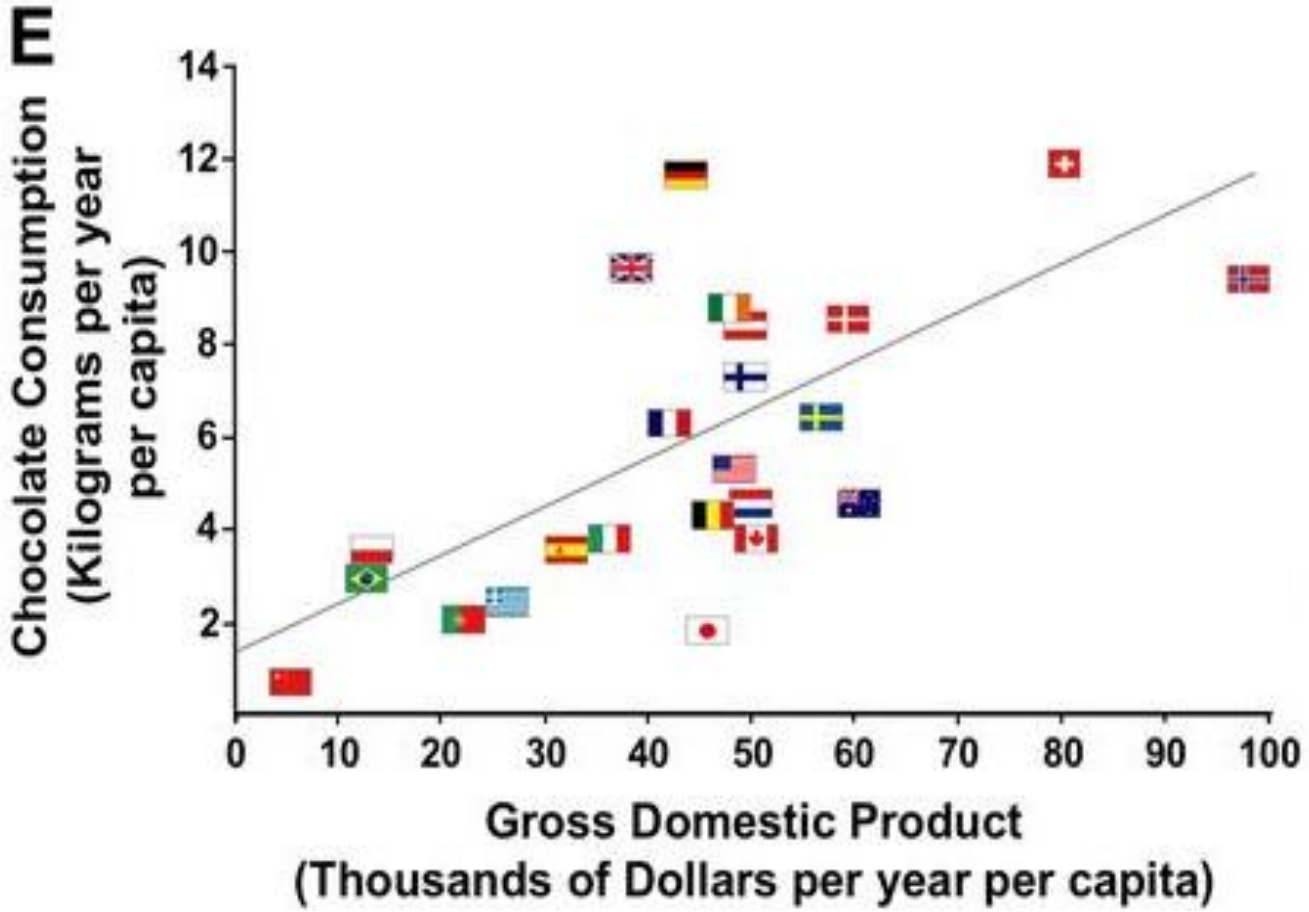
More on assumptions

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

No linear dependence between the errors and the explanatory variable

Omitted variable correlated with the explanatory variable: $y_t = \beta_1 + \beta_2 x_t + \underbrace{\beta_3 z_t + \epsilon_t}_{\eta_t}$





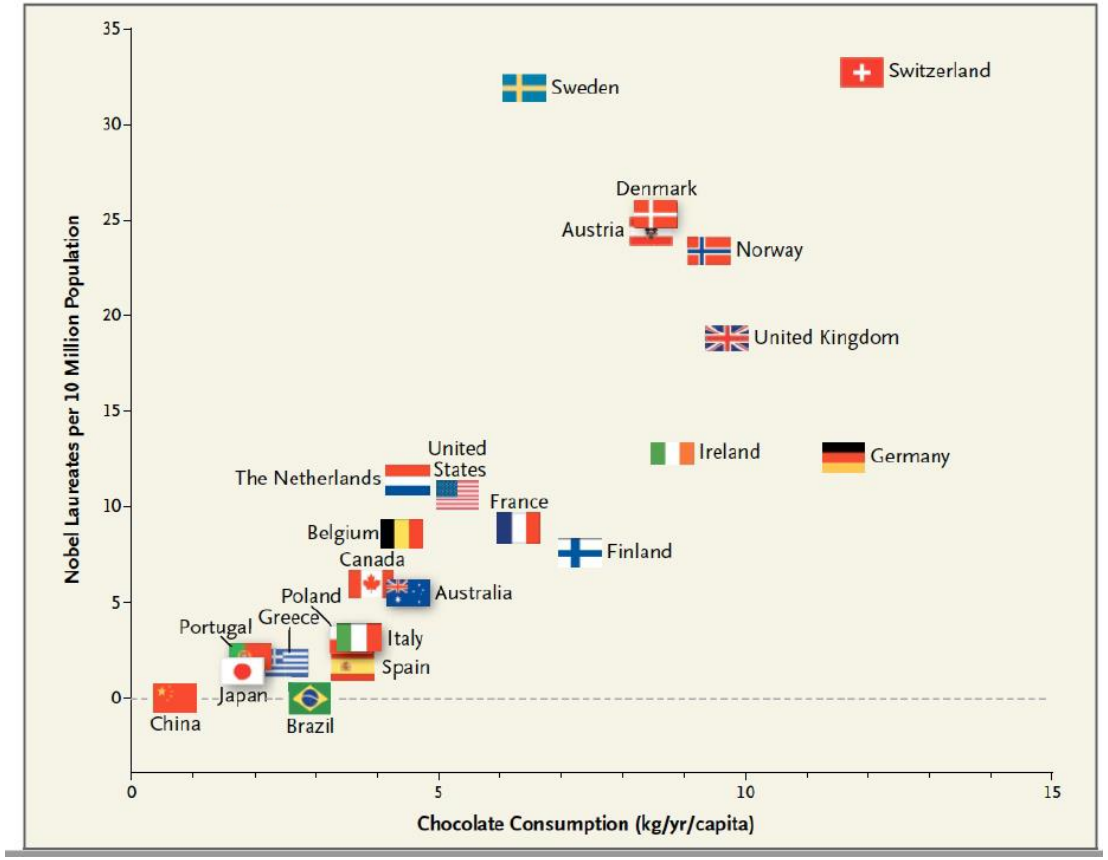
J Nutr, Volume 143, Issue 6, June 2013, Pages 931–933, https://doi.org/10.3945/jn.113.174813

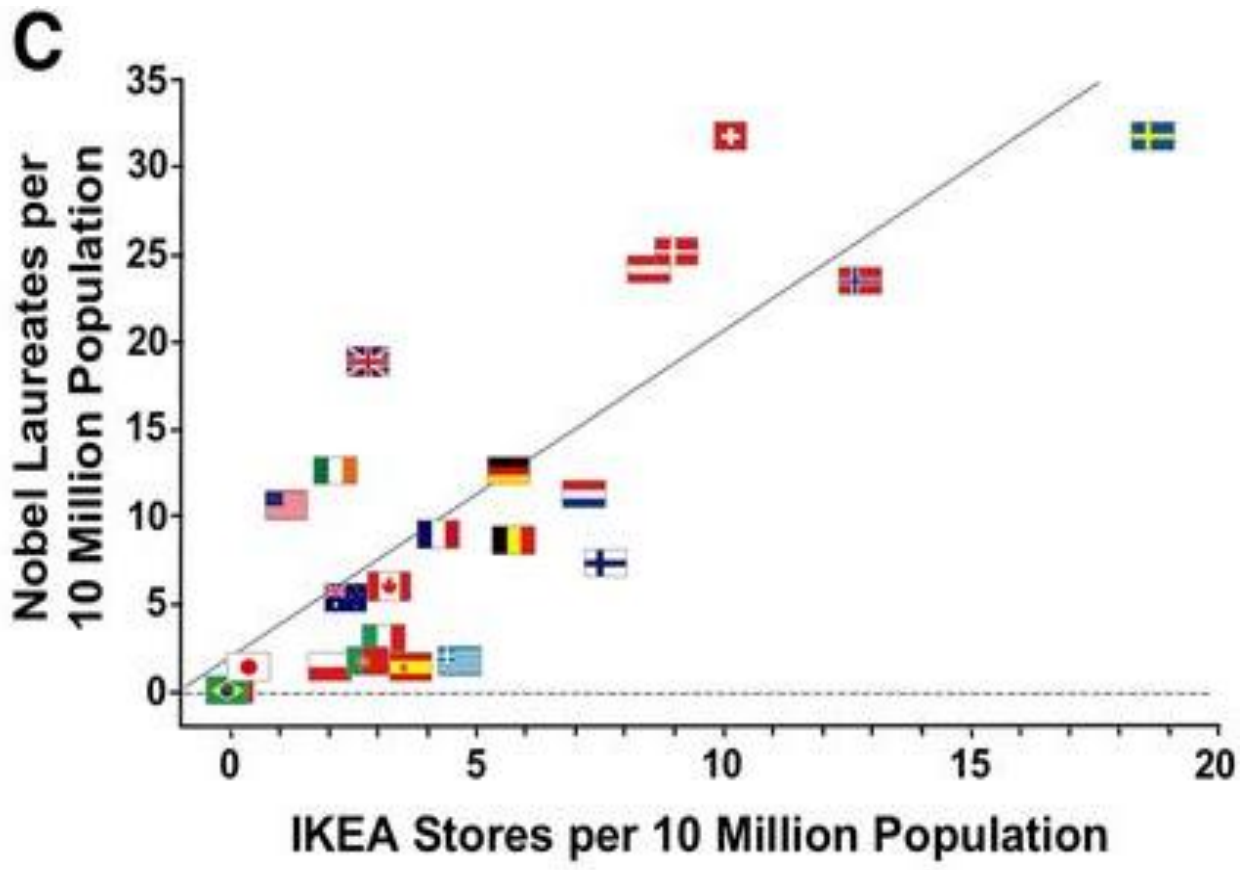
More on assumptions

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

No linear dependence between the errors and the explanatory variable

Omitted variable correlated with the explanatory variable: $y_t = \beta_1 + \beta_2 x_t + \underbrace{\beta_3 z_t + \epsilon_t}_{\eta_t}$



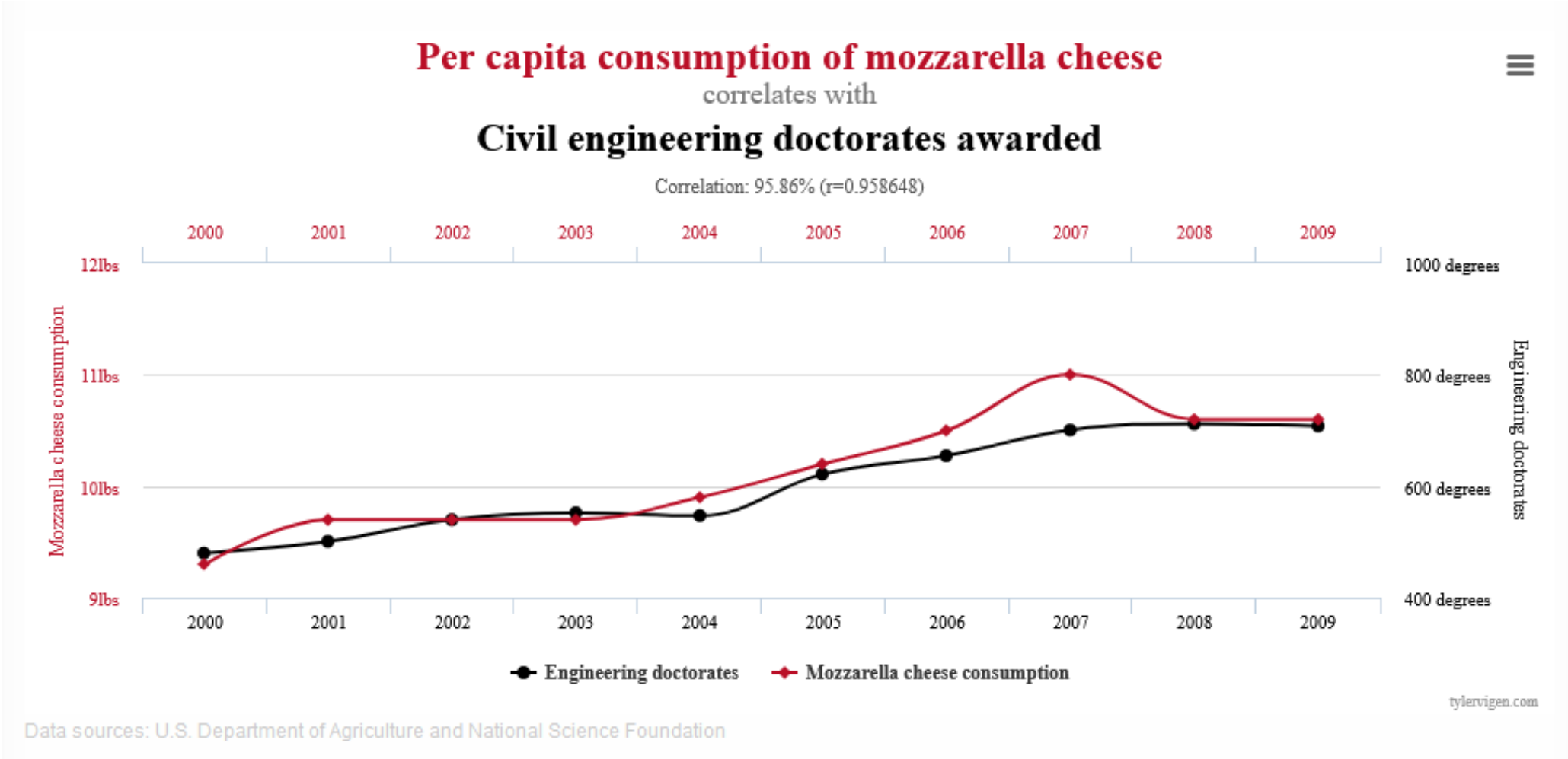


J Nutr, Volume 143, Issue 6, June 2013, Pages 931–933, https://doi.org/10.3945/jn.113.174813

A word of cautious

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

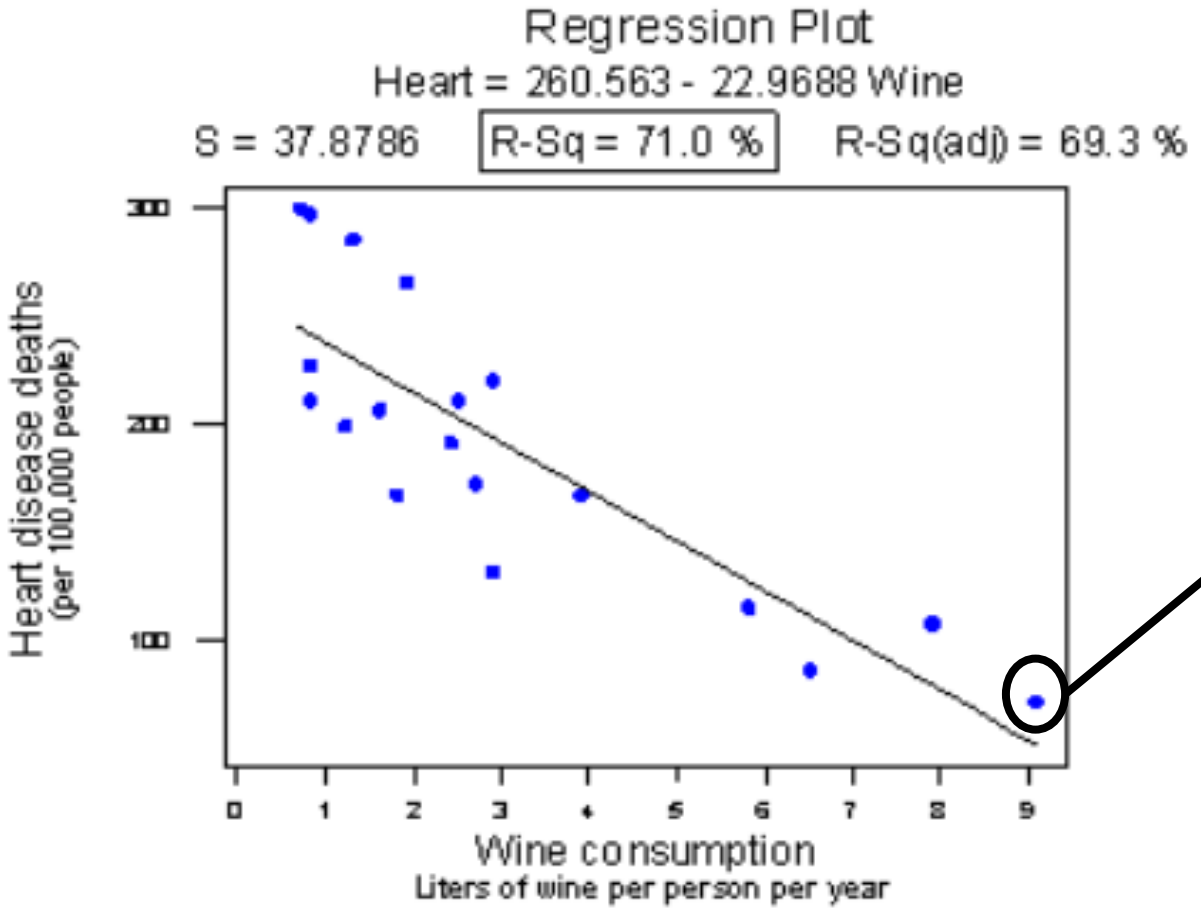
Correlation does not mean causation!



A word of cautious

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Correlation does not mean causation!



France:
9,1 liter per person per year

Pearson correlation of Wine and Heart = -0.843

Source: https://online.stat.psu.edu/stat501/book/export/html/639

More on assumptions

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

4. White noise: No linear dependence between the error terms.

No linear dependence between the error terms : $\text{Cov}(\epsilon_t, \epsilon_j | x_1, \ldots, x_T) = 0 \quad \forall t \neq j$

By Law of iterated expectations (LIE): $E(\text{Cov}(\epsilon_t, \epsilon_j | x_1, \ldots, x_T)) = E(\epsilon_t \epsilon_j) = 0$

Restrictive assumption in time series context

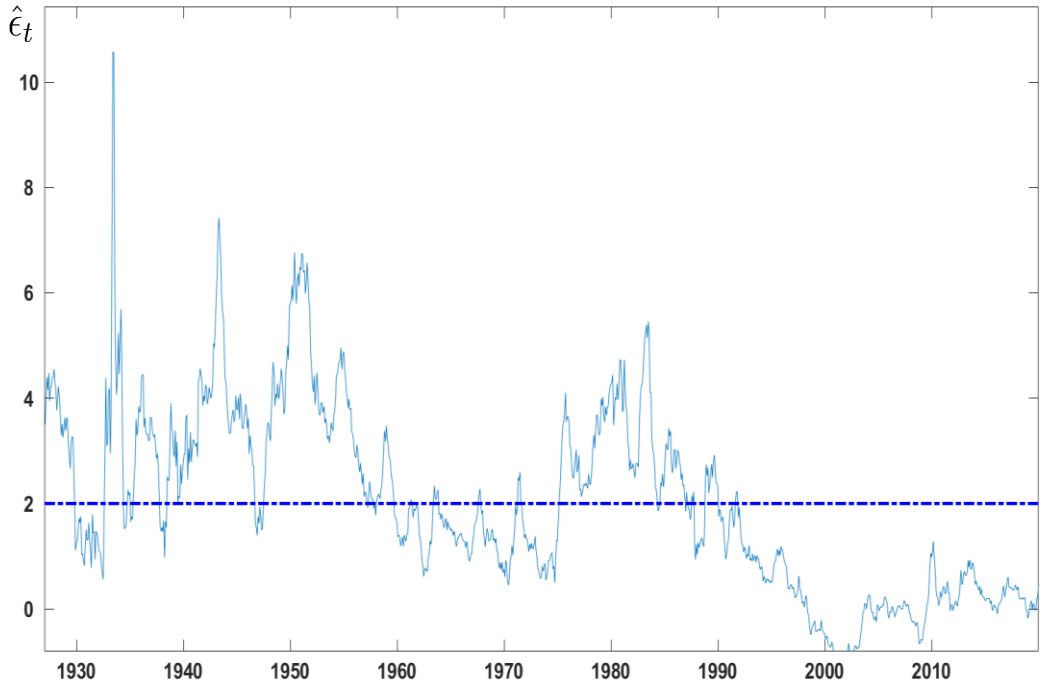
→ An economic crisis → multiple negative shocks.

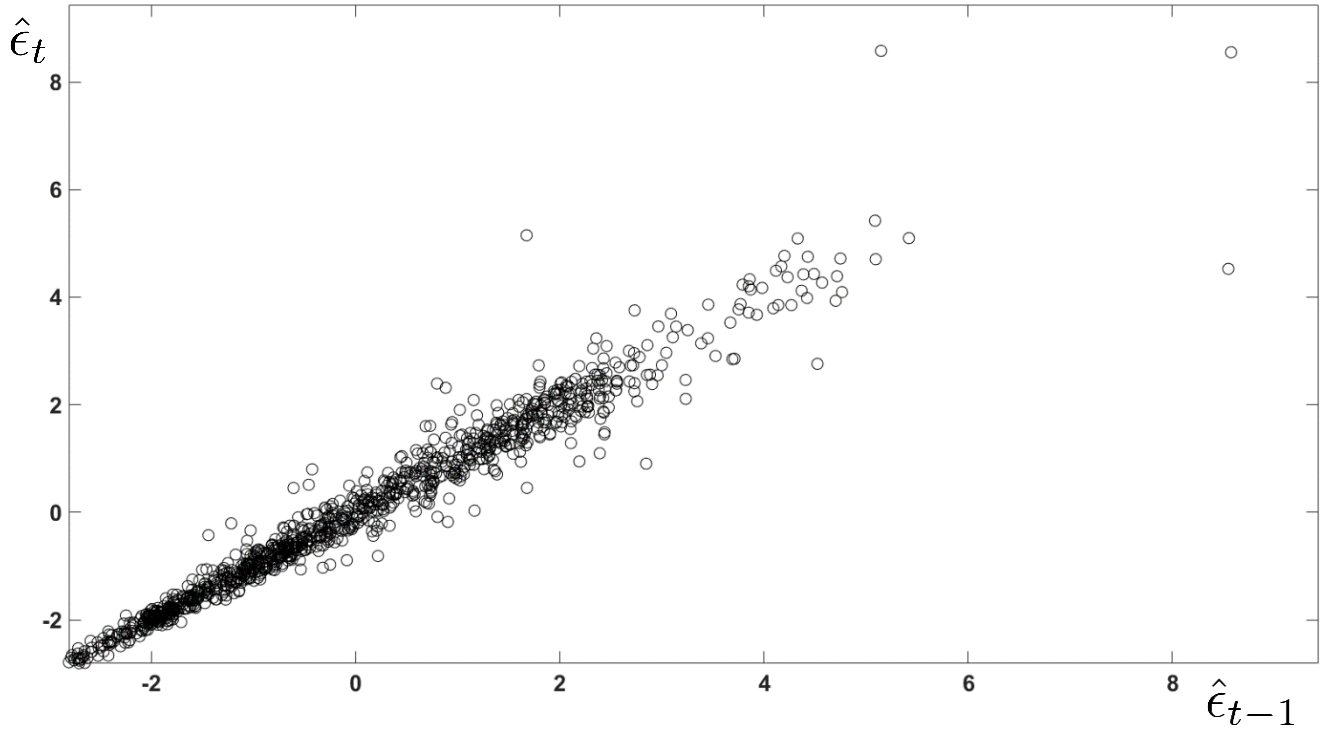
More on assumptions

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

4. White noise: Errors have zero mean and no linear dependence between the error terms.

No linear dependence between the error terms : $\mathrm{Cov}(\epsilon_t, \epsilon_j | x_1, \ldots, x_T) = 0 \quad \forall t \neq j$





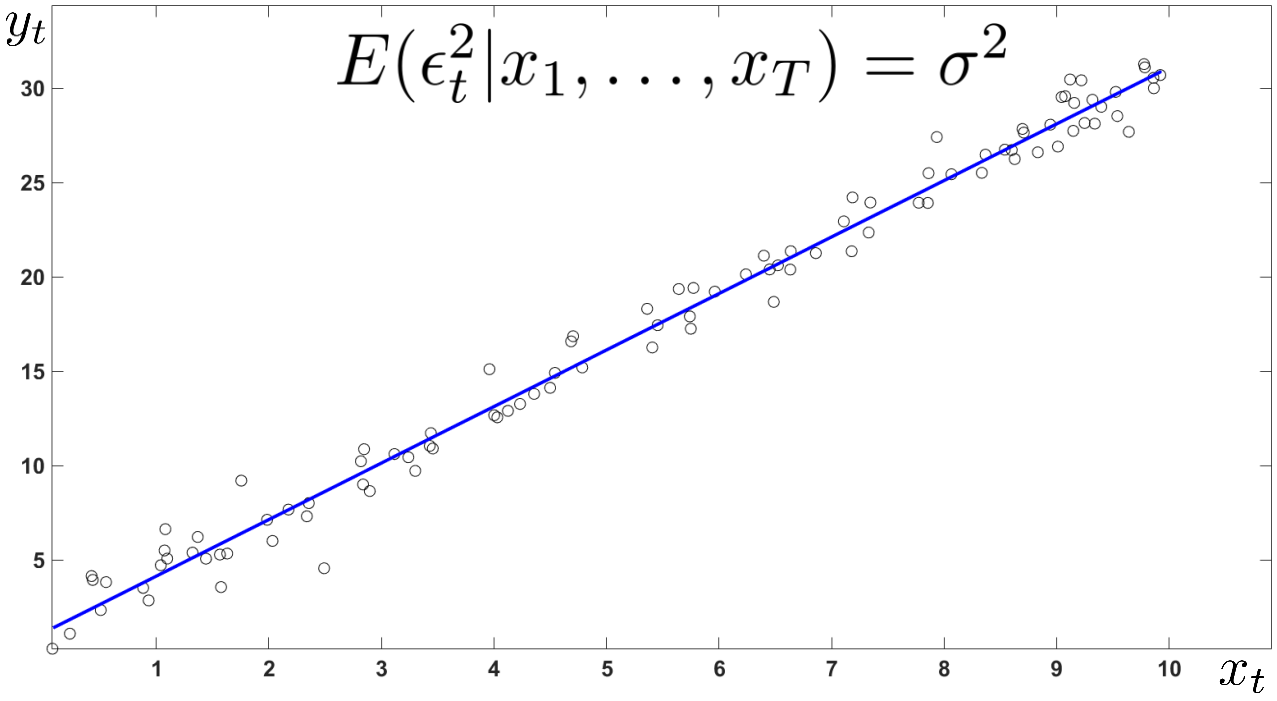
More on assumptions

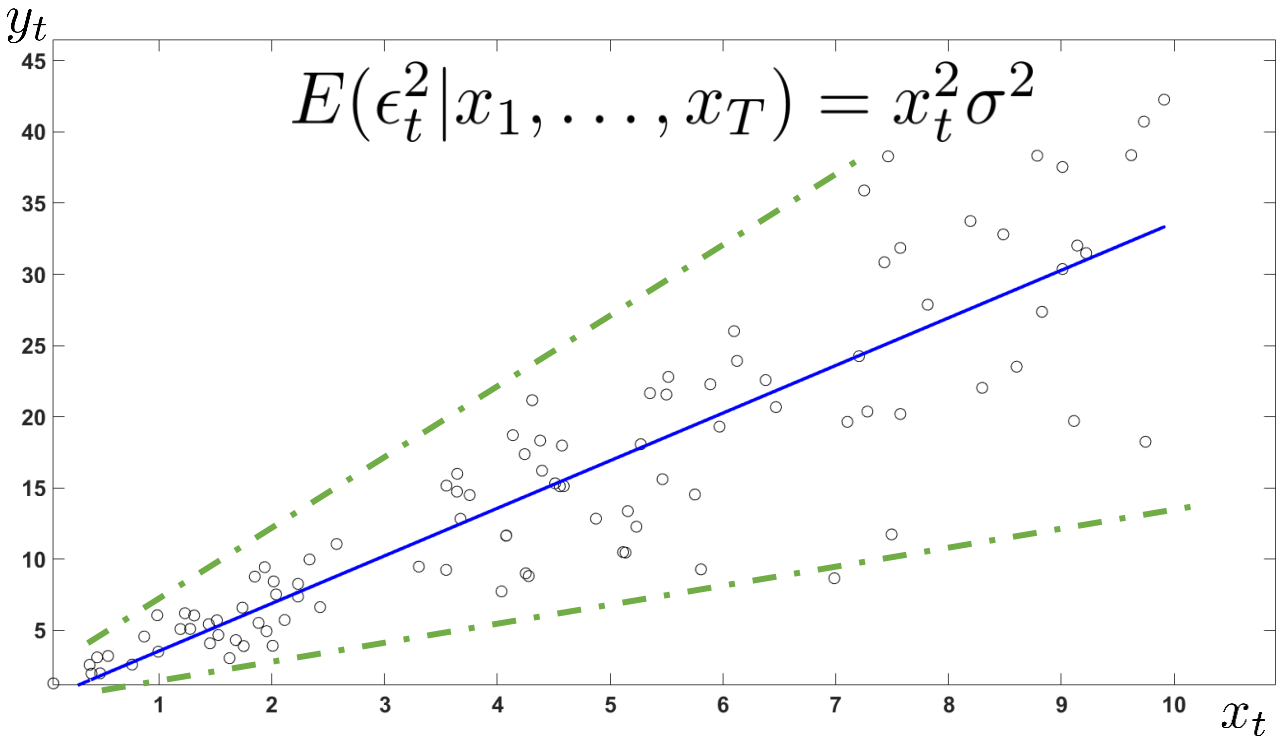
Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

5. Homoskedasticity. The variance of the error term is constant:

$$E(\epsilon_t^2 | x_1, \ldots, x_T) = \sigma^2 \quad \forall t \in [1, T]$$

Unconditional variance is constant: $E(\epsilon_t^2) = \sigma^2 \quad \forall t \in [1, T]$ (LIE)





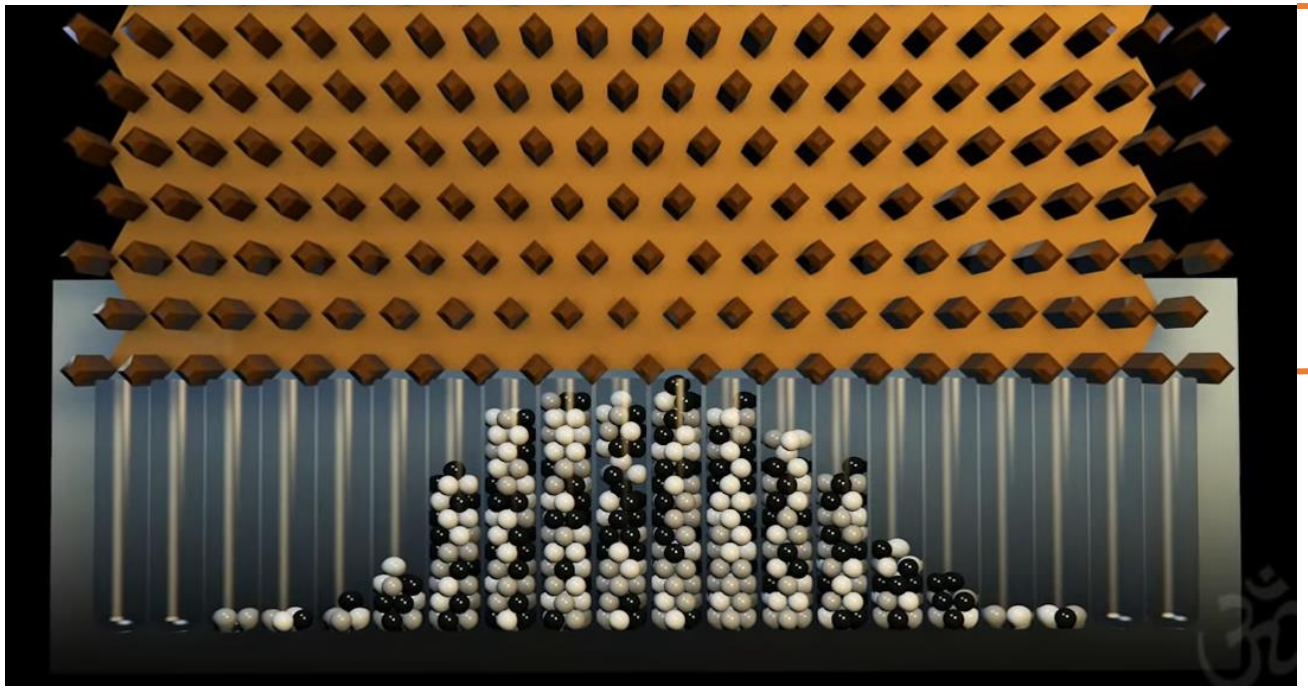
More on assumptions

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

6. Error distribution. Error terms are normally distributed: $\epsilon_t \sim N(0, \sigma^2) \quad \forall t \in [1, T]$

Restrictive assumption, motivated by the CLT: $\epsilon_t = \eta_{t,1} + \eta_{t,2} + \ldots + \eta_{t,N}$

The error term is a function of multiple shocks.



Each layer can be understood as a shock

Properties of our estimator

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Criterion: $(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \hat{\epsilon}_t^2$

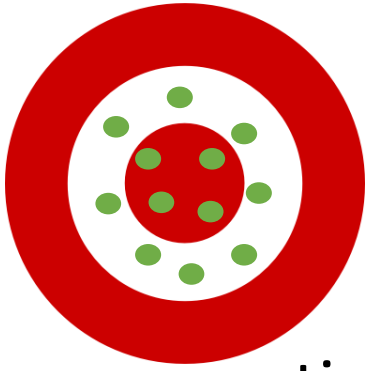
Ordinary least squares (OLS) estimator

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}, \quad \hat{\beta}_2 = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

Assuming hypotheses 1 to 5 (No need of the Normal distribution)

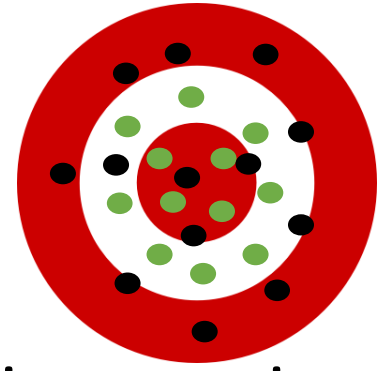
→ OLS estimator is **unbiased**, **efficient** and **consistent**

Unbiasedness



On average, estimates = true coefficient

Efficiency



Variance estimator < variance other estimators

estimator = coefficient.

Properties of our estimator

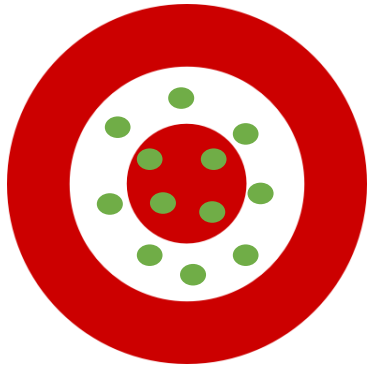
Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Criterion: $(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \hat{\epsilon}_t^2$

Ordinary least squares (OLS) estimator

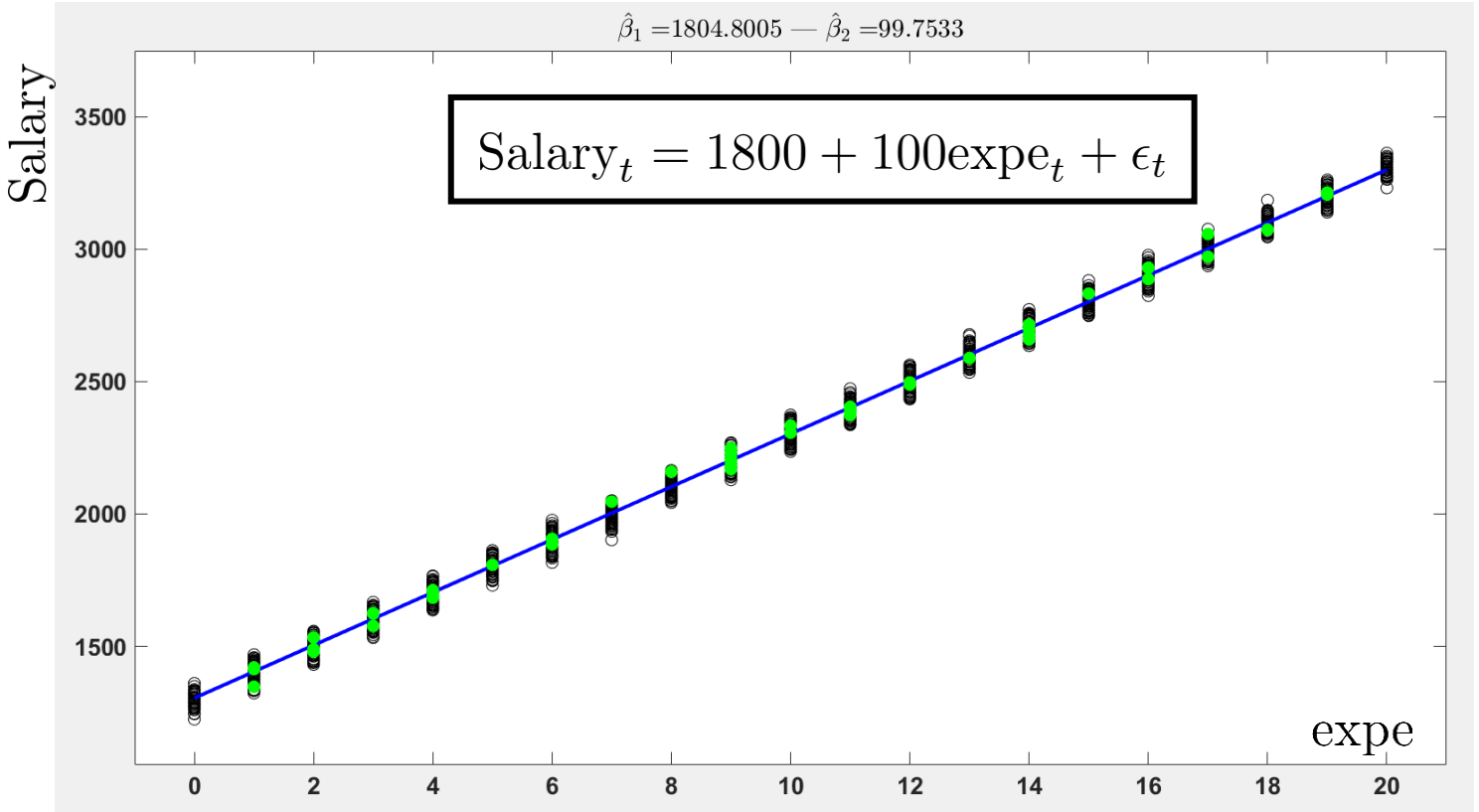
$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}, \quad \hat{\beta}_2 = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

Unbiasedness



$$E(\hat{\beta}_1 | x_1, \ldots, x_T) = \beta_1$$

$$E(\hat{\beta}_2 | x_1, \ldots, x_T) = \beta_2$$



$\frac{1}{1000}\sum_{i=1}^{1000} \hat{\beta}_1^{(i)} = 1800.1$

$\frac{1}{1000}\sum_{i=1}^{1000} \hat{\beta}_2^{(i)} = 100.01$

Properties of our estimator

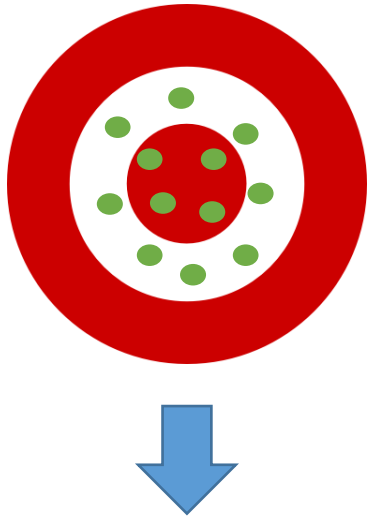
Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Criterion: $(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \hat{\epsilon}_t^2$

Ordinary least squares (OLS) estimator

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}, \quad \hat{\beta}_2 = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

Unbiasedness



$$E(\hat{\beta}_1|x_1, \ldots, x_T) = \beta_1$$

$$E(\hat{\beta}_2|x_1, \ldots, x_T) = \beta_2$$

Proof:

$$\begin{aligned}
E(\hat{\beta}_2|x_1, ..., x_T) &= E\left(\frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}|x_1, ..., x_T\right), \\
&= \frac{\sum_{t=1}^{T}(x_t - \bar{x})E((y_t - \bar{y})|x_1, ..., x_T)}{\sum_{t=1}^{T}(x_t - \bar{x})^2}, \\
&= \frac{\sum_{t=1}^{T}(x_t - \bar{x})(\beta_1 + \beta_2 x_t - \beta_1 - \beta_2 \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}, \\
&= \frac{\sum_{t=1}^{T}(x_t - \bar{x})\beta_2(x_t - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}, \\
&= \beta_2 \frac{\sum_{t=1}^{T}(x_t - \bar{x})^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}, \\
&= \beta_2
\end{aligned}$$

Properties of our estimator

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Criterion: $(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \hat{\epsilon}_t^2$

Ordinary least squares (OLS) estimator

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}, \quad \hat{\beta}_2 = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

Another unbiased estimator:

$$\tilde{\beta}_1 = \bar{y}^{(10)} - \tilde{\beta}_2 \bar{x}^{(10)},$$

$$\tilde{\beta}_2 = \frac{\sum_{t=1}^{10}(x_t - \bar{x}^{(10)})(y_t - \bar{y}^{(10)})}{\sum_{t=1}^{10}(x_t - \bar{x}^{(10)})^2}$$

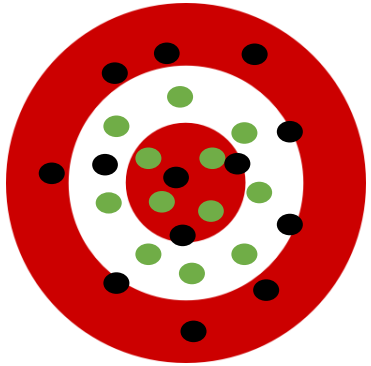
with

$$\bar{y}^{(10)} = \frac{1}{10}\sum_{t=1}^{10} y_t$$

$$\bar{x}^{(10)} = \frac{1}{10}\sum_{t=1}^{10} x_t$$

Estimator is linear w.r.t y

Efficiency w.r.t. other linear unbiased estimators



$$V(\hat{\beta}_1) \leq V(\tilde{\beta}_1)$$

$$V(\hat{\beta}_2) \leq V(\tilde{\beta}_2)$$

Properties of our estimator

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Criterion: $(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \hat{\epsilon}_t^2$

OLS estimator: best linear unbiased estimator (BLUE)

Best = efficient estimator among the **linear unbiased** estimator.

Another estimator:

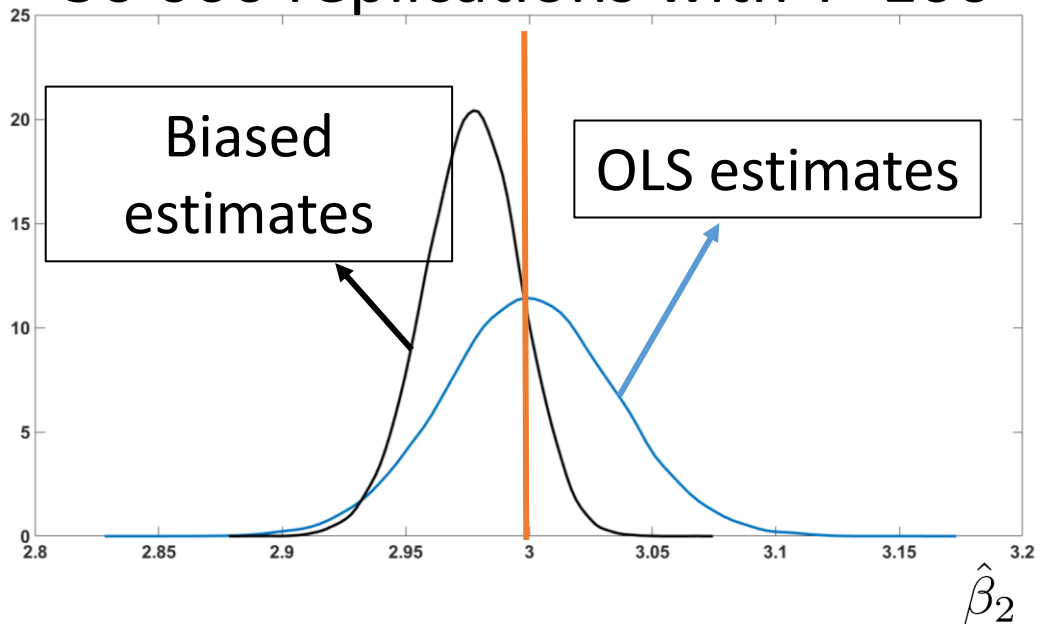
$$\tilde{\beta}_1 = 0,$$
$$\tilde{\beta}_2 = 2$$

$$\text{Var}(\tilde{\beta}_1 | x_1, \ldots, x_T) = 0$$

Biased and not consistent

$$y_t = \underbrace{\beta_1}_{=1} + \underbrace{\beta_2}_{=3} \underbrace{x_t}_{\sim U[0,10]} + \underbrace{\epsilon_t}_{\sim N(0,1)}$$

50 000 replications with T=100



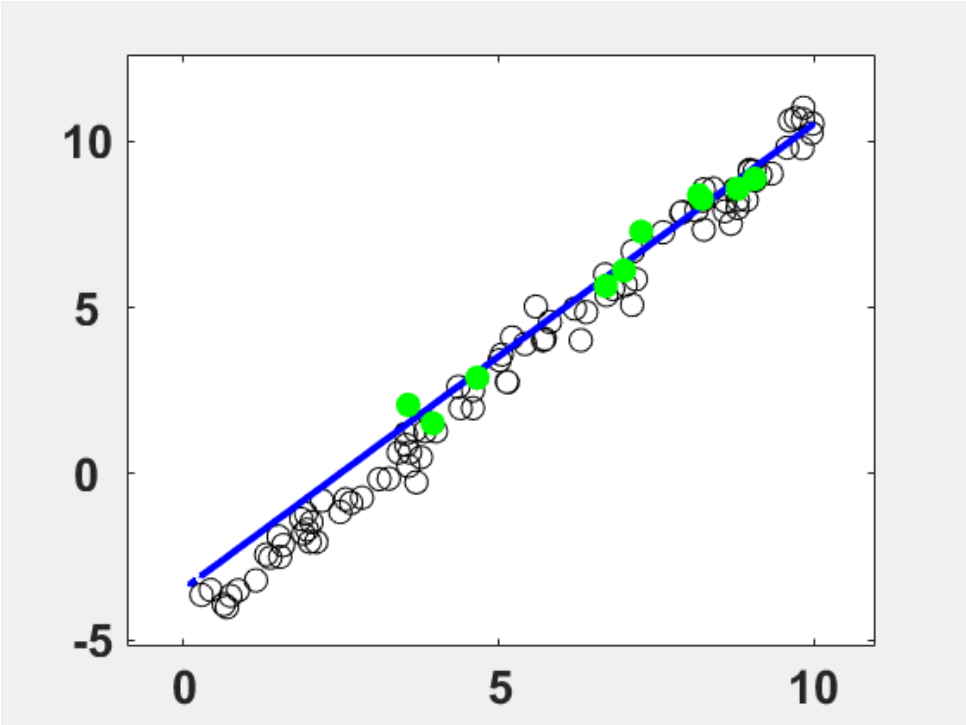
OLS Variance

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

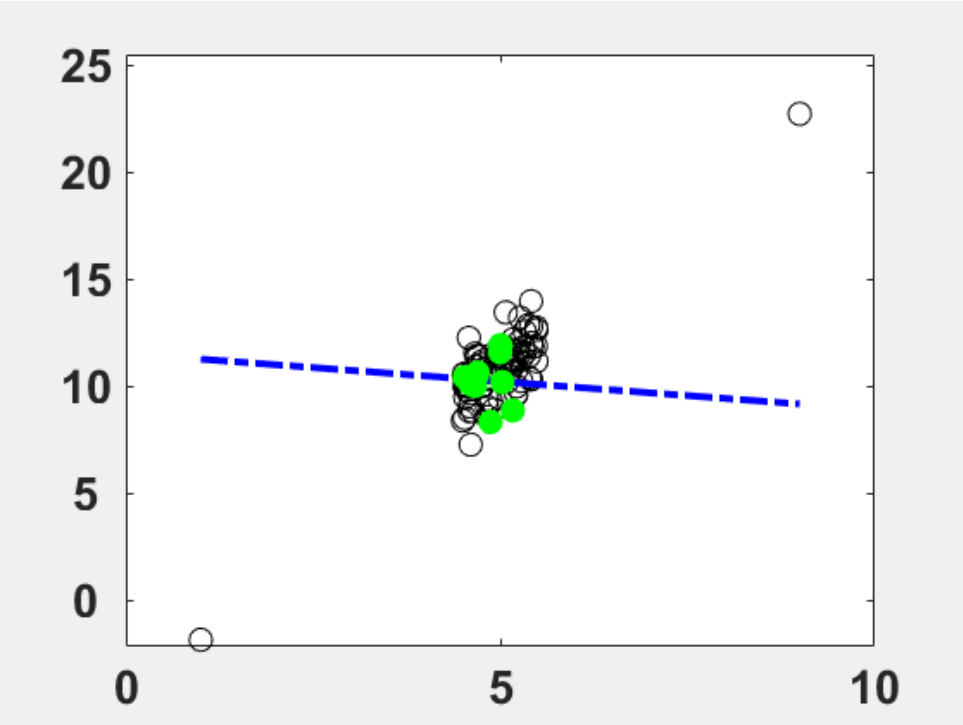
Criterion: $(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \hat{\epsilon}_t^2$

OLS estimator: $E(\hat{\beta}_1|x_1,\ldots,x_T) = \beta_1, E(\hat{\beta}_2|x_1,\ldots,x_T) = \beta_2$

Variance of the OLS estimator: $V(\hat{\beta}_1|x_1,\ldots,x_T) =?, V(\hat{\beta}_2|x_1,\ldots,x_T) =?$



Variance depends on how the exp. var is spread



Variance of the intercept depends on how the exp. var are far from zero.

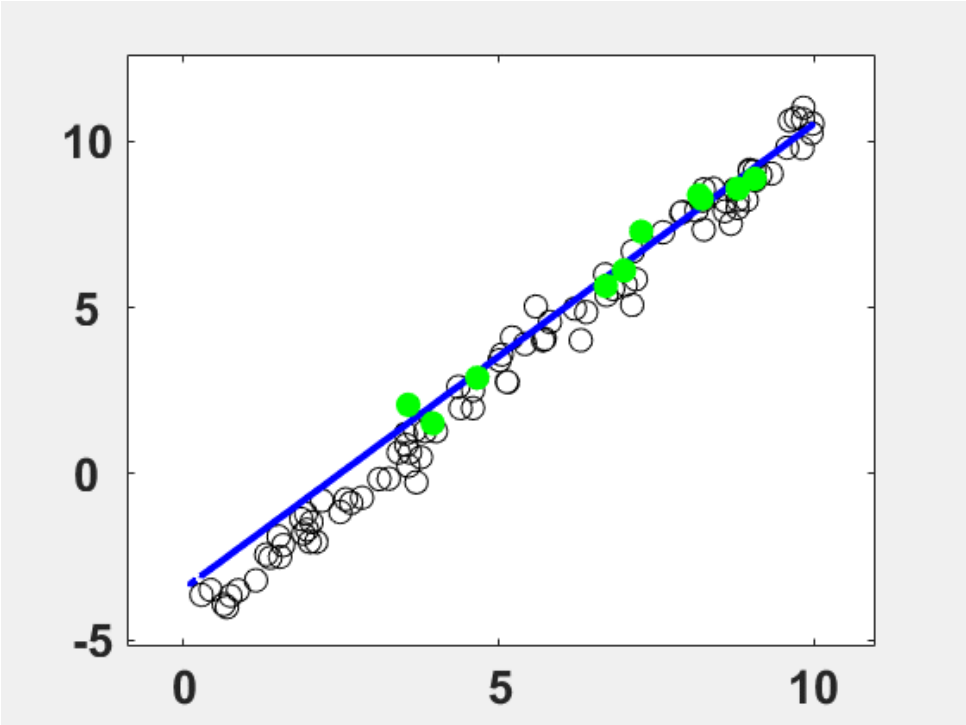
OLS Variance

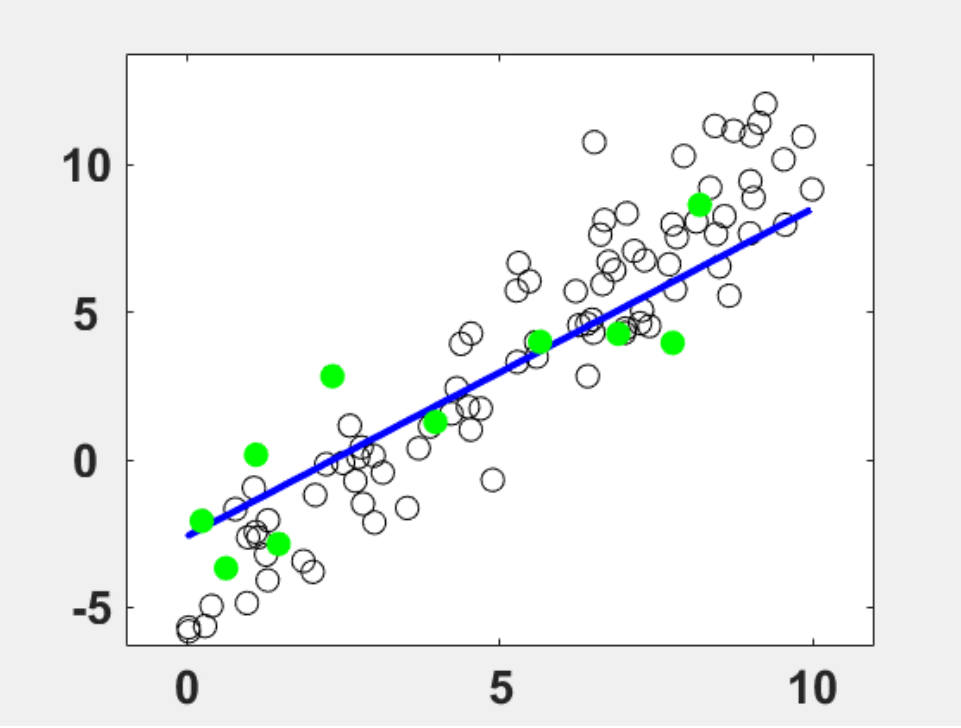
Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Criterion: $(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \hat{\epsilon}_t^2$

OLS estimator: $E(\hat{\beta}_1|x_1,\ldots,x_T) = \beta_1, E(\hat{\beta}_2|x_1,\ldots,x_T) = \beta_2$

Variance of the OLS estimator: $V(\hat{\beta}_1|x_1,\ldots,x_T) =?, V(\hat{\beta}_2|x_1,\ldots,x_T) =?$





Variance depends on how large can the errors be.

OLS Variance

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Criterion: $(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \hat{\epsilon}_t^2$

Variance of the OLS estimator:

$$V(\hat{\beta}_1 | x_1, \ldots, x_T) = \sigma^2 \frac{\sum_{t=1}^{T} x_t^2}{T \sum_{t=1}^{T} (x_t - \bar{x})^2}$$

Variance of the intercept depends on how the exp. var are far from zero.

$$V(\hat{\beta}_2 | x_1, \ldots, x_T) = \sigma^2 \frac{1}{\sum_{t=1}^{T} (x_t - \bar{x})^2}$$

Variance depends on how the exp. var is spread

(For a proof, see p 143 of Brooks)

Variance of the error

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Criterion: $(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \hat{\epsilon}_t^2$

Estimating the variance of the error.

Unbiased estimator of the variance: $\hat{\sigma}^2 = \frac{\sum_{t=1}^{T} \hat{\epsilon}_t^2}{T-2}$

Intuition:

Variance: $\sigma^2 = \text{Var}(\epsilon_t^2 | x_1, \ldots, x_T) = E(\epsilon_t^2 | x_1, \ldots, x_T) - \underbrace{E(\epsilon_t | x_1, \ldots, x_T)^2}_{=0}$

Law of large number: $E(\epsilon_t^2 | x_1, \ldots, x_T) \approx \frac{1}{T} \sum_{t=1}^{T} \epsilon_t^2$

Approximation of the error term: $\frac{1}{T} \sum_{t=1}^{T} \epsilon_t^2 \approx \frac{1}{T} \sum_{t=1}^{T} \hat{\epsilon}_t^2$

Estimated error term fixed by two equations: $\frac{1}{T} \sum_{t=1}^{T} \epsilon_t^2 \approx \frac{1}{T-2} \sum_{t=1}^{T} \hat{\epsilon}_t^2$

Linear regression with one explanatory variable

Statistical tests

Statistical inference

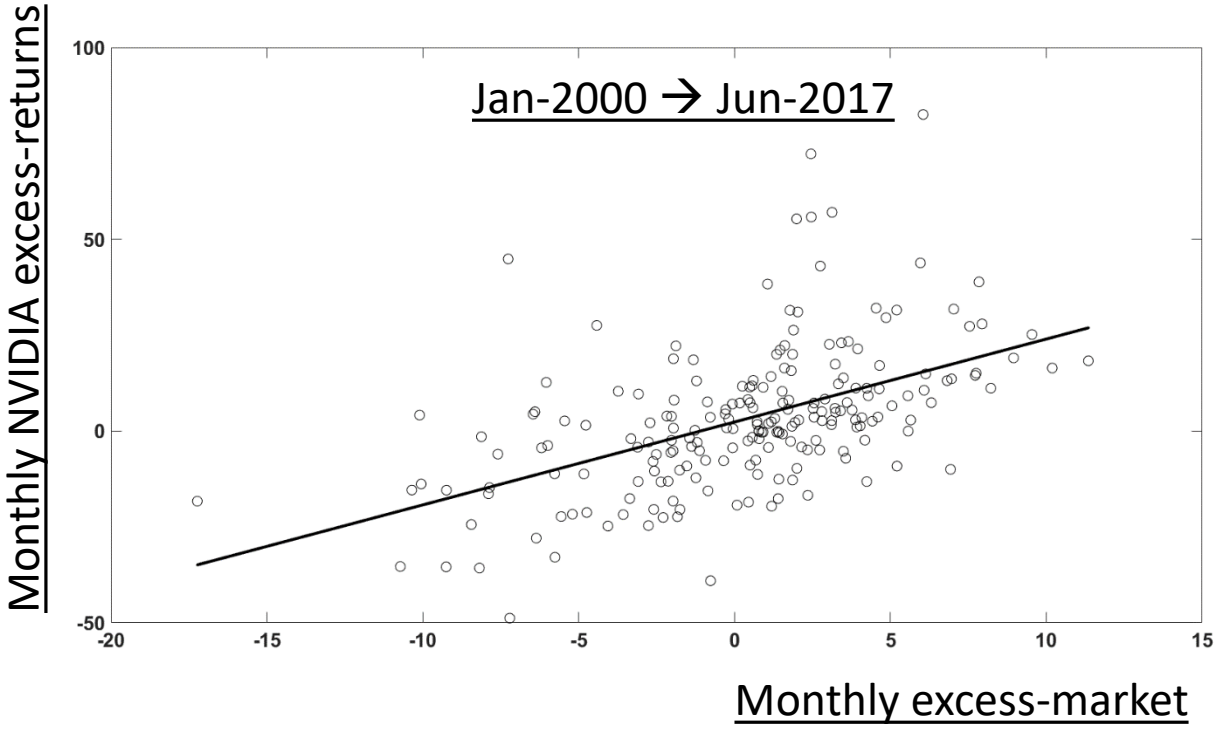
Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

Criterion: $(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} \sum_{t=1}^{T} \hat{\epsilon}_t^2$

Hypothesis testing: ”Can I trust my estimates ?”

Capital asset pricing model:

$$y_t = 2.35 + 2.16 x_t$$



Beta of 2.16 → Risky asset

How risky is it ?

Is it likely that the true coefficient equal 1 ?

Statistical inference

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t$

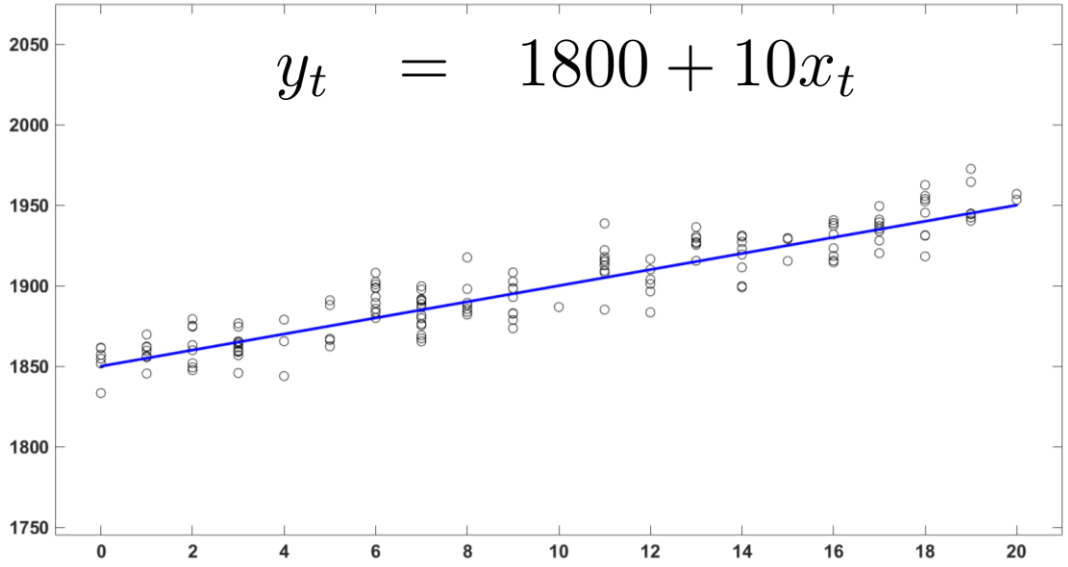
Salary-Experience example: Company of 150 employees.

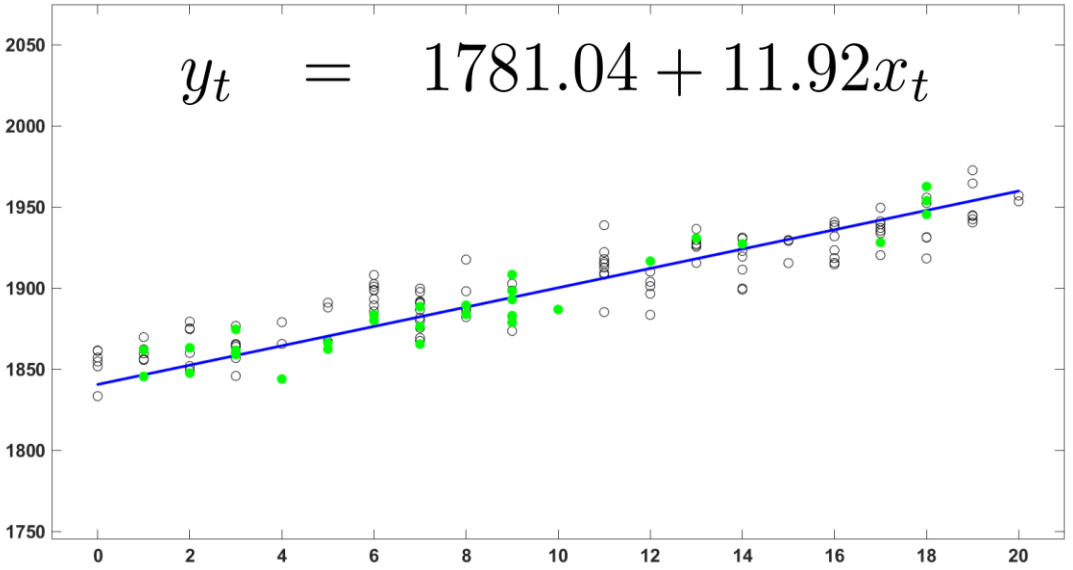
Dependent variable: gross salary per month.

Explanatory variable: Experience per year.

True linear regression: $y_t = 1800 + 10x_t + \epsilon_t \quad \text{with } \epsilon_t \sim N(0, \underbrace{625}_{25^2})$

- Expected wage of 1800 when Xp = 0
- Expected wage increases by 120 per year
- Error term: Initial wage depends on economic state and on negotiations.

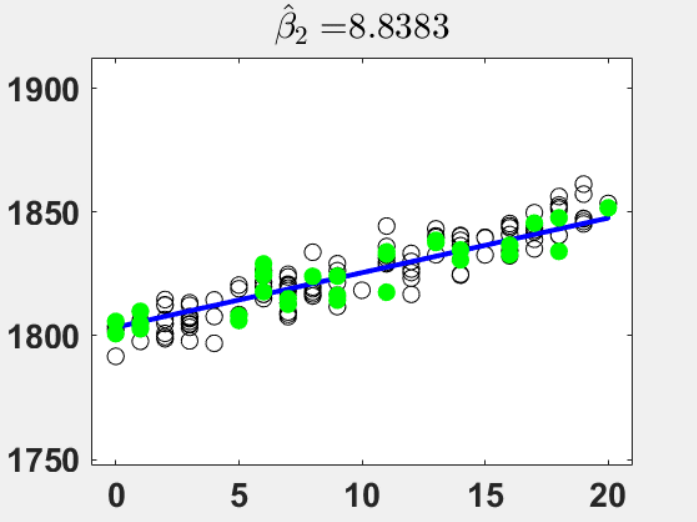




Statistical inference

Linear regression: $y_t = 1800 + 10x_t + \epsilon_t,$
$\epsilon_t \sim N(0, 625)$

Estimates depend on the selected employees:



Hypothesis tests:

Two-sided test: $H_0 : \beta_2 = 10$ vs $H_1 : \beta_2 \neq 10$

One-sided test: $H_0 : \beta_2 = 10$ vs $H_1 : \beta_2 > 10$ or $H_0 : \beta_2 = 10$ vs $H_1 : \beta_2 < 10$

Information to perform the test:

Variance of the error: $\hat{\sigma}^2 = \frac{\sum_{t=1}^{T} \hat{\epsilon}_t^2}{T-2} = 427.35$

Variance of the estimator: $V(\hat{\beta}_2 | x_1, \ldots, x_T) = \frac{\hat{\sigma}^2}{\sum_{t=1}^{T} (x_t - \bar{x})^2} = 0.56$

Statistical inference

Linear regression: $y_t = 1800 + 10x_t + \epsilon_t,$
$\epsilon_t \sim N(0, 625)$

Assumption 6: **Normality of the error term.**

$$y_t = \beta_1 + \beta_2 x_t + \epsilon_t \Rightarrow y_t|x_t \sim N(\beta_1 + \beta_2 x_t, \sigma^2)$$

Linear combinations of normal distributions: $\sum_{t=1}^{T} \omega_t y_t \sim N(\cdot, \cdot)$

Expectation: $E(\sum_{t=1}^{T} \omega_t y_t | x_1, \ldots, x_T) = \sum_{t=1}^{T} \omega_t (\beta_1 + \beta_2 x_t)$

Variance: $V(\sum_{t=1}^{T} \omega_t y_t | x_1, \ldots, x_T) = \sigma^2 \sum_{t=1}^{T} \omega_t^2$ (because of no linear dependence of the errors)

OLS estimators:

$$\begin{aligned}
\hat{\beta}_2 &= \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{T}(x_t - \bar{x})^2} \\
&= \sum_{t=1}^{T} \underbrace{\frac{(x_t - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}}_{\omega_t} y_t, \\
&= \sum_{t=1}^{T} \omega_t y_t.
\end{aligned}$$

$$\Rightarrow \quad \begin{aligned}
\hat{\beta}_2 | x_1, \ldots, x_T &\sim N(E(\hat{\beta}_2 | x_1, \ldots, x_T), V(\hat{\beta}_2 | x_1, \ldots, x_T)), \\
&= N(\beta_2, \frac{\sigma^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2})
\end{aligned}$$

Statistical inference

Linear regression: $y_t = 1800 + 10x_t + \epsilon_t,$
$\epsilon_t \sim N(0, 625)$

Assumption 6: **Normality of the error term.**

The estimators are normally distributed!

$$\hat{\beta}_1|x_1,\ldots,x_T \sim N(\beta_1, \sigma^2 \frac{\sum_{t=1}^T x_t^2}{T\sum_{t=1}^T (x_t-\bar{x})^2})$$

$$\hat{\beta}_2|x_1,\ldots,x_T \sim N(\beta_2, \frac{\sigma^2}{\sum_{t=1}^T (x_t-\bar{x})^2})$$

$$Z_{\beta_1} = \frac{\hat{\beta}_1-\beta_1}{\sqrt{\sigma^2 \frac{\sum_{t=1}^T x_t^2}{T\sum_{t=1}^T (x_t-\bar{x})^2}}} \sim N(0,1)$$

$$Z_{\beta_2} = \frac{\hat{\beta}_2-\beta_2}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^T (x_t-\bar{x})^2}}} \sim N(0,1)$$

Every estimate Z_{β_2} is a realization of $N(0,1)$ when evaluated at the true β_2

Statistical inference

Linear regression: $y_t = 1800 + 10x_t + \epsilon_t,$
$\epsilon_t \sim N(0, 625)$

With Assumption 6: OLS estimators are normally distributed!

Statistical test: Test if $\beta_2 = 10$

$$Z_{10} = \frac{\hat{\beta}_2 - 10}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}}} \sim N(0,1)$$

➡ Hypothesis is true:

Estimates will be realizations of a N(0,1)!

Test if $\beta_2 = 15$

$$Z_{15} = \frac{\hat{\beta}_2 - 15}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}}} \sim N(0,1)$$

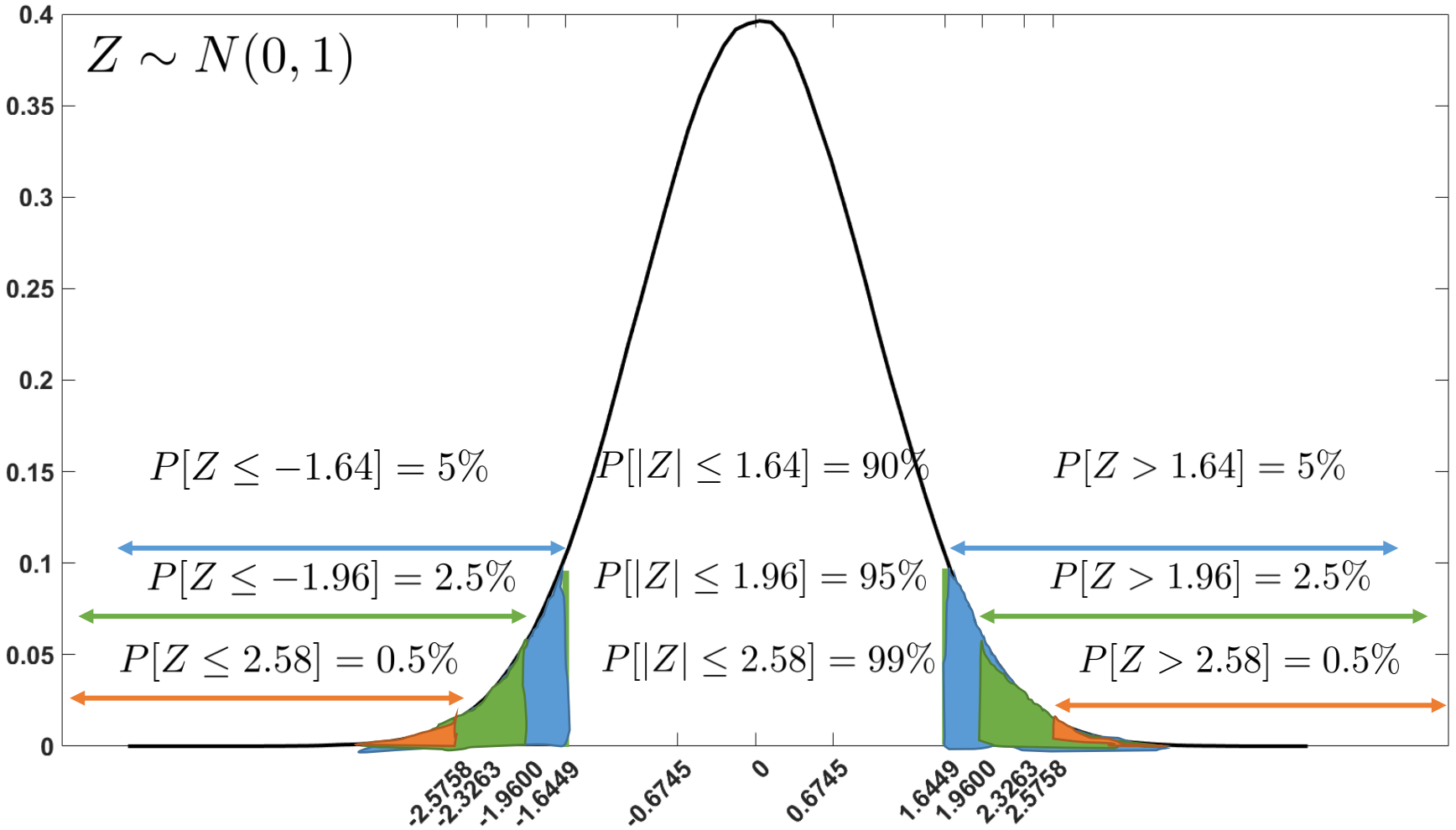
➡ Hypothesis is false:

Estimates are not realizations of a N(0,1)!

Statistical inference

pdf: $f(z) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{z^2}{2}\}$

Standard Normal distribution:



Fluctuation interval:

90%: [-1.64;1.64]

95%: [-1.96;1.96]

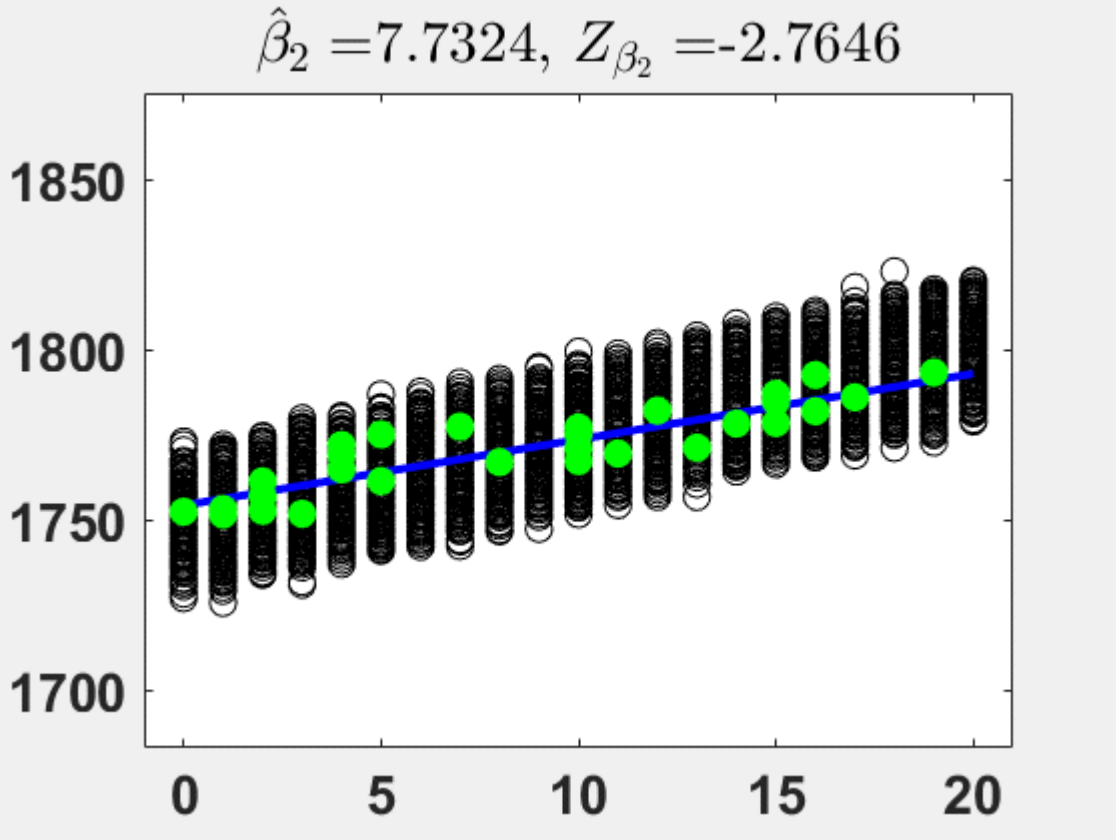
99%: [-2.58;2.58]

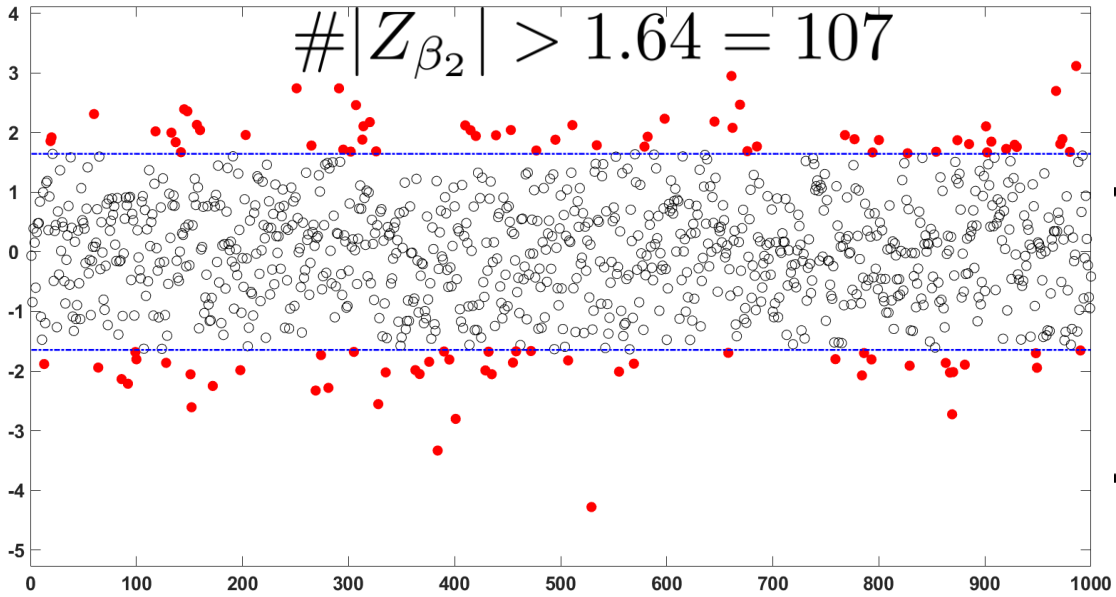
Statistical inference

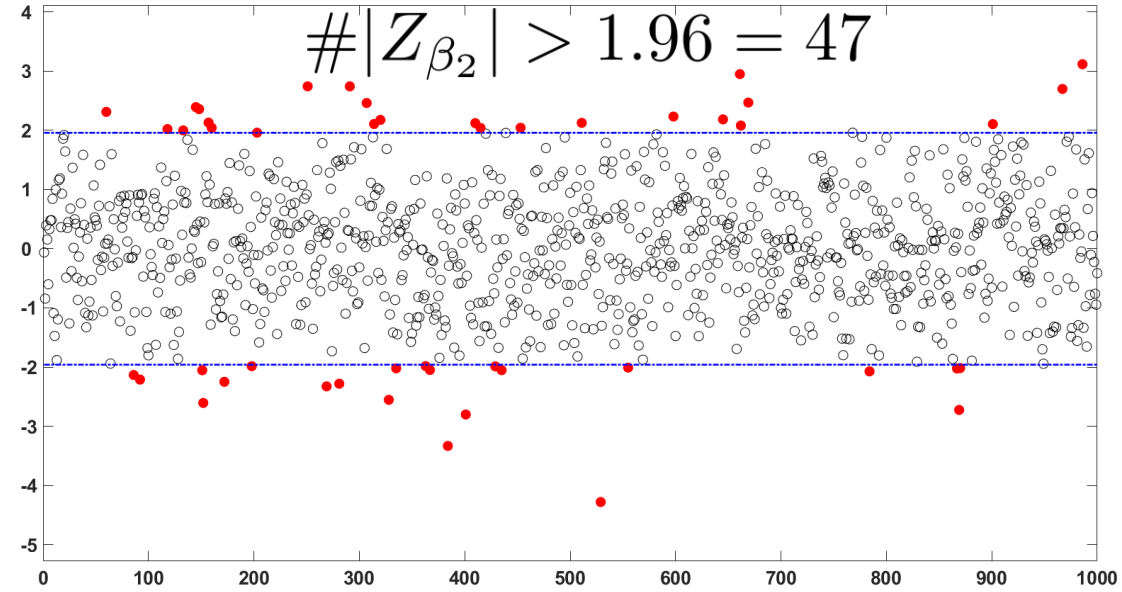
Linear regression: $y_t = 1800 + 10x_t + \epsilon_t,$
$\epsilon_t \sim N(0, 625)$

Example: Test if $\beta_2 = 10 : Z_{10} = \dfrac{\hat{\beta}_2 - 10}{\sqrt{\dfrac{\sigma^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}}}$

$$H_0 : \beta_2 = 10 \text{ vs } H_1 : \beta_2 \neq 10$$





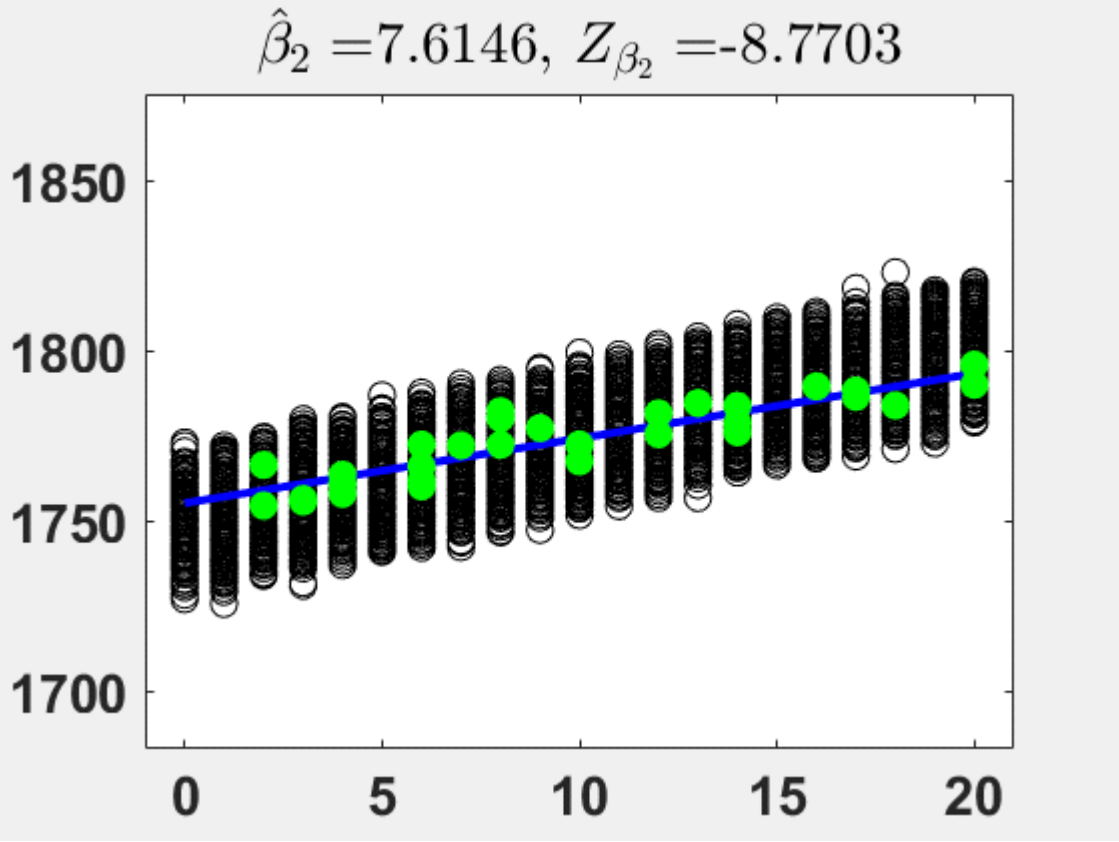


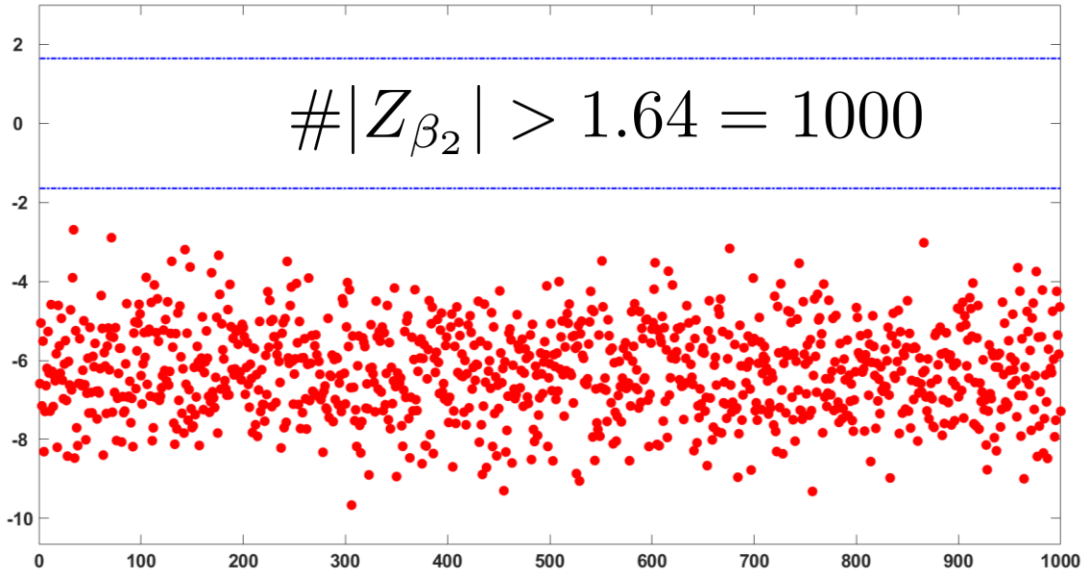
Statistical inference

Linear regression: $y_t = 1800 + 10x_t + \epsilon_t,$
$\epsilon_t \sim N(0, 625)$

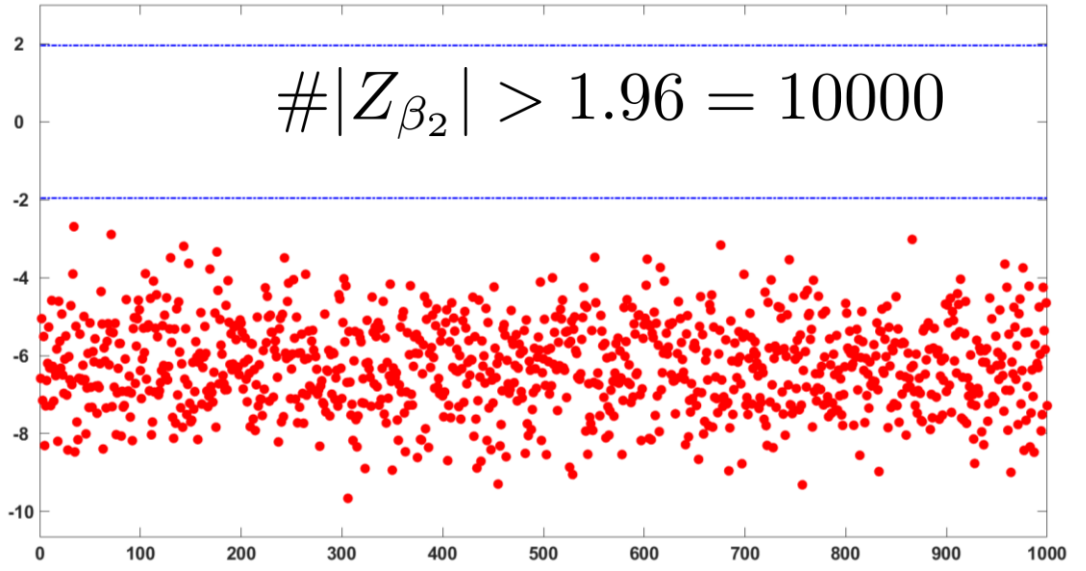
Example: Test if $\beta_2 = 15 : Z_{15} = \frac{\hat{\beta}_2 - 15}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}}}$

$$H_0 : \beta_2 = 15 \text{ vs } H_1 : \beta_2 \neq 15$$





90%: [-1.64;1.64]



95%: [-1.96;1.96]

Statistical inference

Linear regression: $y_t = 1800 + 10x_t + \epsilon_t,$
$\epsilon_t \sim N(0, 625)$

Two sided statistical test: $H_0 : \beta_2 = 10$ vs $H_1 : \beta_2 \neq 10$

1. Compute the **Z-statistics**: $Z_{\beta_2} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}}}$
2. Choose a significant level: $1 - \alpha$ → 95% implies $\alpha = 0.05$
3. Find the Normal criterion for the significant level: $P[Z > Z_{crit}] = \frac{\alpha}{2}$

→ 95% significant test gives $\alpha = 0.05$ and implies $Z_{crit} = 1.96$

4. Reject the Null if $|Z_{\beta_2}| > Z_{crit}$

If $|Z_{\beta_2}| \leq Z_{crit}$, we do not accept the Null hypothesis.

Statistical inference

Linear regression: $y_t = 1800 + 10x_t + \epsilon_t,$
$\epsilon_t \sim N(0, 625)$

One-sided statistical test: $H_0 : \beta_2 = 10$ vs $H_1 : \beta_2 > 10$

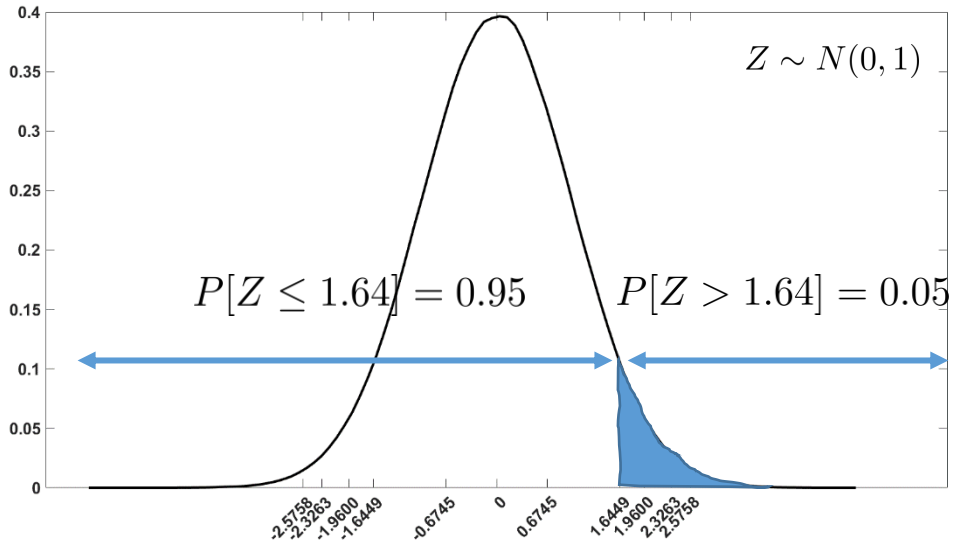
1. Compute the **Z-statistics**: $Z_{\beta_2} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}}}$
2. Choose a significant level: $1 - \alpha$ ⇒ 95% implies $\alpha = 0.05$
3. Find the Normal criterion for the significant level: $P[Z > Z_{crit}] = \alpha$

⇒ 95% significant test gives $\alpha = 0.05$ and implies $Z_{crit} = 1.64$

4. Reject the Null if $|Z_{\beta_2}| > Z_{crit}$

Why this change ?

⇒ Only interested in one tail of the distribution

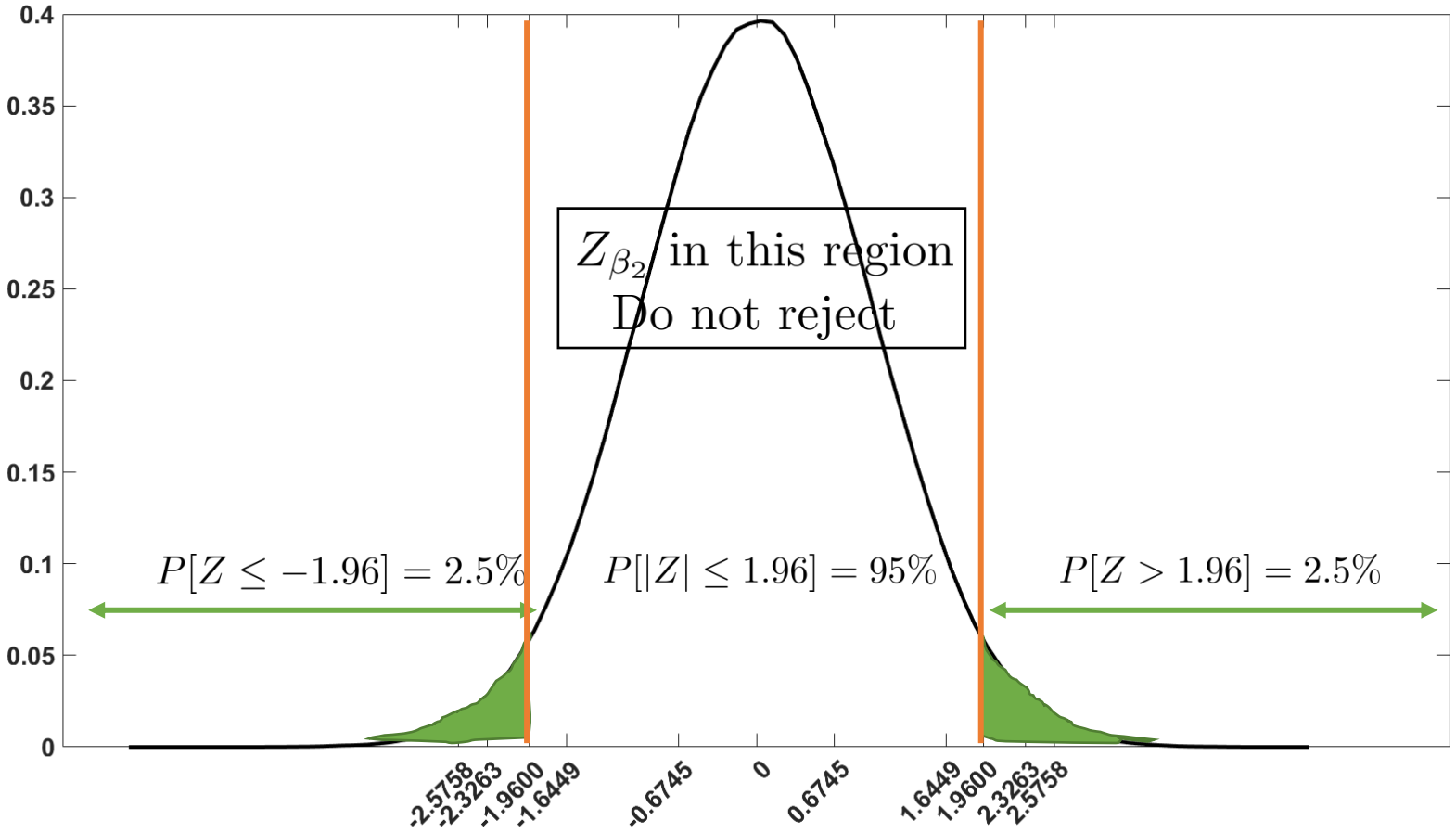


Confidence interval

Z-statistics: $Z_{\beta_2} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}}}$

Two sided statistical test:

We do not reject the Null if $-Z_{crit} \leq Z_{\beta_2} \leq Z_{crit}$ → $-Z_{crit} \leq \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} \leq Z_{crit}$



Confidence interval (CI)

$$[\hat{\beta}_2 - Z_{crit} SE(\hat{\beta}_2); \hat{\beta}_2 + Z_{crit} SE(\hat{\beta}_2)]$$

If β_2 lies in the CI
Do not reject

Example: CI

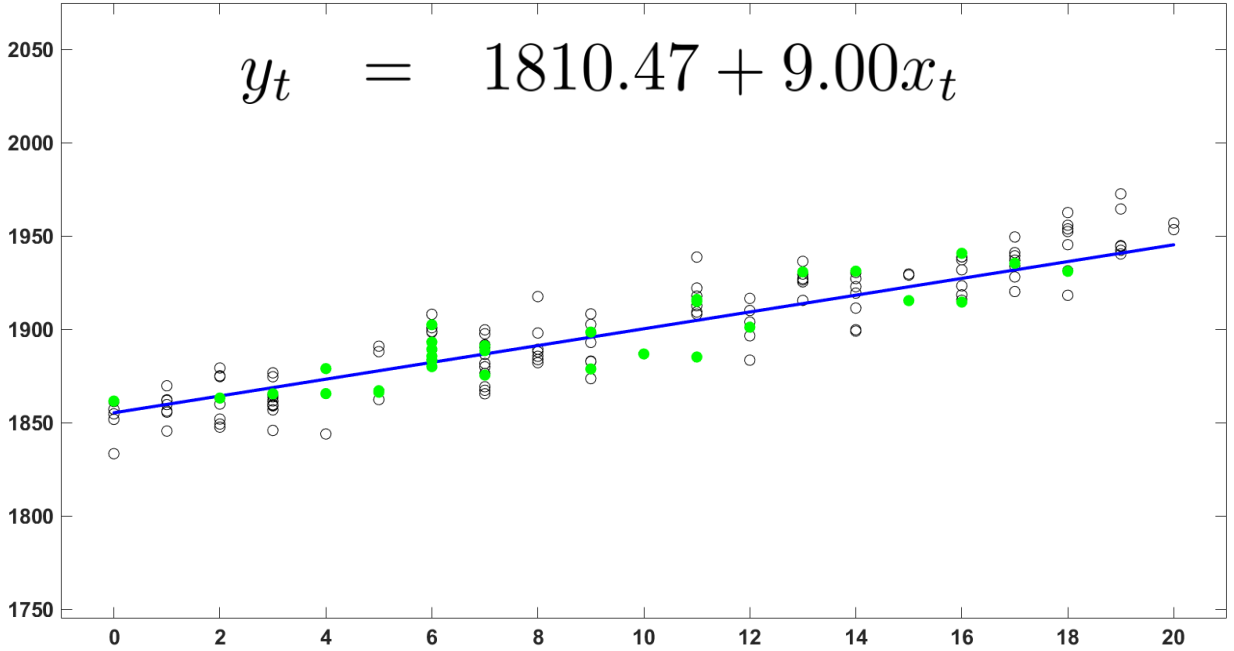
Z-statistics: $Z_{\beta_2} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}}}$

Salary-Experience:

Exact regression: $y_t = 1800 + 10x_t + \epsilon_t$

Dependent variable: gross salary per month.

Explanatory variable: Experience per year.



$\sigma^2 = 425$

$V(\hat{\beta}_2|x_1, \ldots, x_T) = \frac{\sigma^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2} = 0.69$

Two-sided test: $H_0 : \beta_2 = 10$ vs $H_1 : \beta_2 \neq 10$

$$Z_{\beta_2} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}}} = \frac{9.00 - 10}{\sqrt{0.69}} = -1.02$$

Choose a significant level: $\alpha = 0.05$

$\alpha = 0.05$ and implies $Z_{crit} = 1.96$

We do not reject the hypothesis

Confidence interval (CI)

$[\hat{\beta}_2 - Z_{crit}SE(\hat{\beta}_2); \hat{\beta}_2 + Z_{crit}SE(\hat{\beta}_2)] = [7.4; 10.60]$

We do not reject any value lying in the CI

Another sample

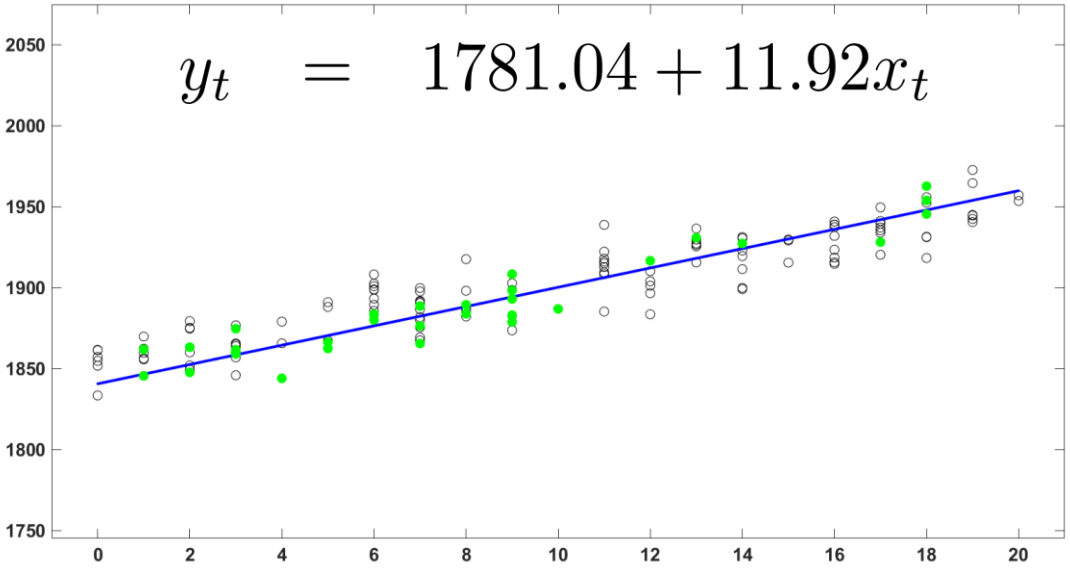
Z-statistics: $Z_{\beta_2} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}}}$

Salary-Experience:

Exact regression: $y_t = 1800 + 10x_t + \epsilon_t$

Dependent variable: gross salary per month.

Explanatory variable: Experience per year.



$\sigma^2 = 425$

$V(\hat{\beta}_2 | x_1, \ldots, x_T) = \frac{\sigma^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2} = 0.56$

Two-sided test: $H_0 : \beta_2 = 10$ vs $H_1 : \beta_2 \neq 10$

$$Z_{\beta_2} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}}} = \frac{11.92 - 10}{\sqrt{0.56}} = 2.57$$

Choose a significant level: $\alpha = 0.05$

$\alpha = 0.05$ and implies $Z_{crit} = 1.96$

We reject the hypothesis!

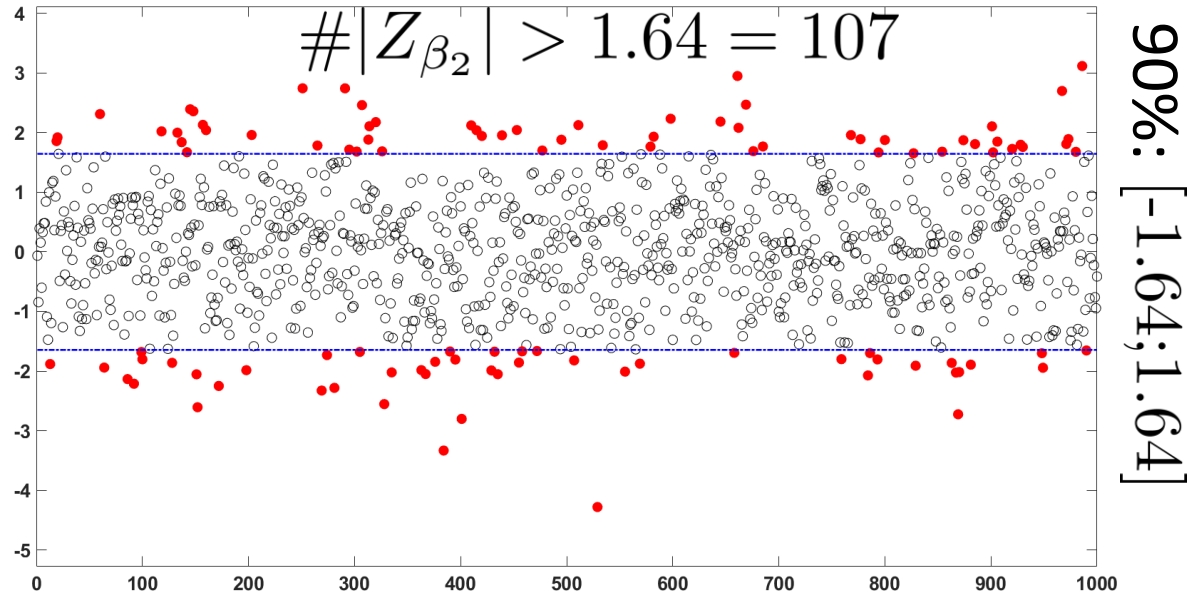
Confidence interval (CI)

$[\hat{\beta}_2 - Z_{crit} SE(\hat{\beta}_2); \hat{\beta}_2 + Z_{crit} SE(\hat{\beta}_2)] = [10.45; 13.39]$

We do not reject any value lying in the CI

Statistical inference

Linear regression: $y_t = 1800 + 10x_t + \epsilon_t,$
$\epsilon_t \sim N(0, 625)$



Important take away

When the hypothesis is correct:

➡ Probability of rejecting the hypothesis= α

Terminology:

Under the null, the probability of a false negative is equal to α

Error of type one

➡ Probability that CI does not contain the true coefficient is equal to α

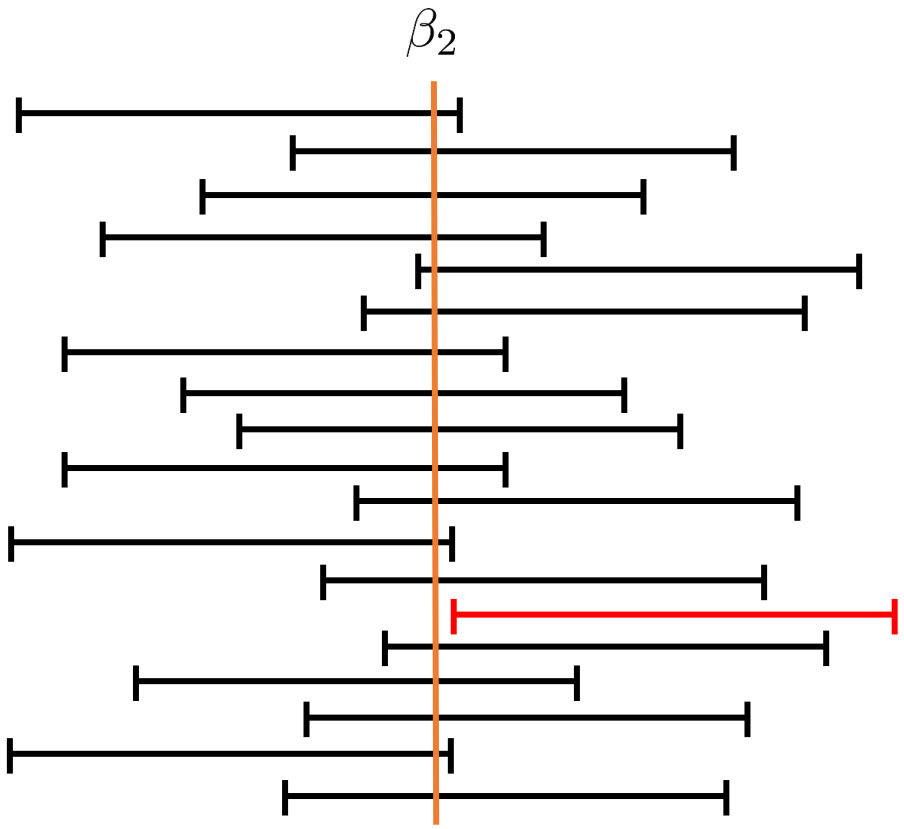
Statistical inference

Linear regression: $y_t = 1800 + 10x_t + \epsilon_t,$
$\epsilon_t \sim N(0, 625)$

Important take away

When the hypothesis is correct:

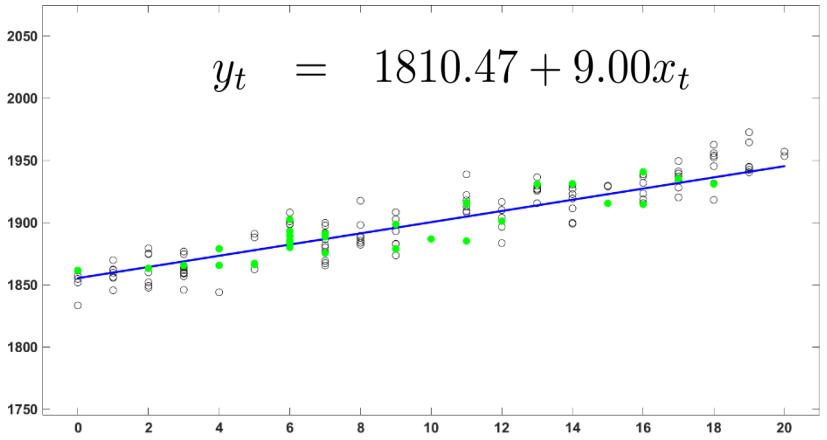
➡ Probability that CI does not contain the true coefficient is equal to α



P-value

Salary-Experience:

$$V(\hat{\beta}_2|x_1,\ldots,x_T) = \frac{\sigma^2}{\sum_{t=1}^{T}(x_t-\bar{x})^2} = 0.69$$



Z-statistics: $Z_{\beta_2} = \frac{\hat{\beta}_2-\beta_2}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^{T}(x_t-\bar{x})^2}}}$

Two-sided test: $H_0 : \beta_2 = 10$ vs $H_1 : \beta_2 \neq 10$

$$Z_{\beta_2} = \frac{\hat{\beta}_2-\beta_2}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^{T}(x_t-\bar{x})^2}}} = \frac{9.00-10}{\sqrt{0.69}} = -1.02$$

Choose a significant level: $P[Z > Z_{crit}] = \frac{\alpha}{2}$

$\alpha = 0.01$ ➡ $Z_{crit} = 2.58$ ✓

$\alpha = 0.05$ ➡ $Z_{crit} = 1.96$ ✓

$\alpha = 0.1$ ➡ $Z_{crit} = 1.64$ ✓

$\alpha = 0.2$ ➡ $Z_{crit} = 1.28$ ✓

$\alpha = 0.4$ ➡ $Z_{crit} = 0.84$ ✗

$P[Z > 1.02] = \frac{\text{p-value}}{2}$

P-value: $P[Z > |Z_\beta|] = \frac{\text{p-value}}{2}$

$$\text{p-value} = 2(1 - P[Z \leq |Z_\beta|])$$

P-value is the level at which the hypothesis is rejected

Example:

$$\text{p-value} = 2(1 - P[Z \leq 1.02]) = 0.31$$

We do not reject the null for any significant level smaller than $\alpha = 0.31$

Statistical inference

Linear regression: $y_t = 1800 + 10x_t + \epsilon_t,$
$\epsilon_t \sim N(0, 625)$

We have a test statistics: $Z_{\beta_2} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}}}$

Small problem: Do not know the value of σ^2

Instead: We use the unbiased estimate $\hat{\sigma}^2 = \frac{\sum_{t=1}^{T} \hat{\epsilon}_t^2}{T-2}$

When the estimate is too small → We will reject more the null hypothesis!

To take into account the variability of the variance estimate

→ We use a student distribution instead of a Normal.

Test statistics: $t_{\beta_2} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}}} \sim t(T-2)$

Statistical inference

Linear regression: $y_t = 1800 + 10x_t + \epsilon_t,$
$\epsilon_t \sim N(0, 625)$

Student distribution: $t(\mathrm{df})$

df: degree of freedom = Number of observations – Number of coefficients = T - 2.

Intuition: When df increases, the student distribution converges to the Normal distribution.

Df	P[t>1.64]	P[t>1.96]
5	0.081	0.054
15	0.061	0.034
25	0.057	0.031
100	0.052	0.026
Normal	0.05	0.025

Statistical inference

Linear regression: $y_t = 1800 + 10x_t + \epsilon_t,$
$\epsilon_t \sim N(0, 625)$

Statistical test: $H_0 : \beta_2 = \beta^*$ vs $H_1 : \beta_2 \neq \beta^*$

1. Compute the standard error of the estimator: $SE(\hat{\beta}_2) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}}$
2. Compute the **t-statistics**: $t = \frac{\hat{\beta}_2 - \beta^*}{SE(\hat{\beta}_2)}$
3. Choose a significant level: $1 - \alpha$

t-test:

- Find the criterion
$$P[t_{T-2} > t_{crit}] = \frac{\alpha}{2}$$
- Reject the Null if
$$|t| > t_{crit}$$

Confidence Interval

- Find the criterion
$$P[t_{T-2} > t_{crit}] = \frac{\alpha}{2}$$
- Build the Confidence Interval
$$[\hat{\beta}_2 - t_{crit}SE(\hat{\beta}_2); \hat{\beta}_2 + t_{crit}SE(\hat{\beta}_2)]$$
- Reject the Null if
$$\beta^* \notin [\hat{\beta}_2 - t_{crit}SE(\hat{\beta}_2); \hat{\beta}_2 + t_{crit}SE(\hat{\beta}_2)]$$

P-value

- Find the p-value
$$p_{val} = 2(1 - P[t_{T-2} \leq |t|])$$
- Reject the Null if
$$p_{val} < \alpha$$

Significant level

Usual significant test: 95% ⇒ $t_{crit} = 1.96$ in large sample

Hot research topic

The review of Financial Studies (2015)

... and the Cross-Section of Expected Returns FREE

Campbell R. Harvey, Yan Liu, Heqing Zhu

The Review of Financial Studies, Volume 29, Issue 1, January 2016, Pages 5–68,
https://doi.org/10.1093/rfs/hhv059
Published: 09 October 2015

PDF Split View Cite Permissions Share ▾

Abstract

Hundreds of papers and factors attempt to explain the cross-section of expected returns. Given this extensive data mining, it does not make sense to use the usual criteria for establishing significance. Which hurdle should be used for current research? Our paper introduces a new multiple testing framework and provides historical cutoffs from the first empirical tests in 1967 to today. A new factor needs to clear a much higher hurdle, with a *t*-statistic greater than 3.0. We argue that most claimed research findings in financial economics are likely false.

Journal of Finance (2017)

Presidential Address: The Scientific Outlook in Financial Economics

CAMPBELL R. HARVEY

First published: 08 July 2017 | https://doi.org/10.1111/jofi.12530 | Citations: 123

SECTIONS PDF TOOLS SHARE

Abstract

ABSTRACT

Given the competition for top journal space, there is an incentive to produce "significant" results. With the combination of unreported tests, lack of adjustment for multiple tests, and direct and indirect *p*-hacking, many of the results being published will fail to hold up in the future. In addition, there are basic issues with the interpretation of statistical significance. Increasing thresholds may be necessary, but still may not be sufficient: if the effect being studied is rare, even $t > 3$ will produce a large number of false positives. Here I explore the meaning and limitations of a *p*-value. I offer a simple alternative (the minimum Bayes factor). I present guidelines for a robust, transparent research culture in financial economics. Finally, I offer some thoughts on the importance of risk-taking (from the perspective of authors and editors) to advance our field.

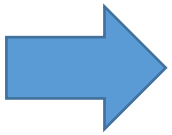
P-hacking

Let us find a new financial factor:

1. We consider financial returns of an asset (our dependent variable)
2. We randomly generate many explanatory variables
3. For each explanatory variable, we consider the simple regression:

$$y_t = \beta_1 + \beta_2 x_t + \epsilon_t$$

- For each explanatory variable, compute the t-stat with Null hypothesis: $H_0 : \beta_2 = 0$



Take the factor with the highest t-stat and publish a paper…

Linear regression with one explanatory variable

Relaxing the Normality assumption

Second Pillar of statistics

Law of large numbers (LLN)

One realization is unpredictable, the average of the realizations is predictable

Central limit theorem (CLT)

At which pace the average of the realizations converges

Alternatively

How many realizations are needed to have a reliable average

Let $\{Z_t\}_{t=1}^{T}$ be independent identically distributed (i.i.d.) with $E(Z_t) = \mu$ and $\mathrm{V}(Z_t) = \sigma^2$.

Then $\lim_{T\to\infty} \sqrt{T}[\frac{1}{T}\sum_{t=1}^{T} \frac{Z_t-\mu}{\sigma}] \to_d N(0,1)$

Proof

Simply put: For large T, $\frac{1}{T}\sum_{t=1}^{T} Z_t \sim N(\mu, \frac{\sigma^2}{T})$

A sample average is a realization of a Normal distribution with variance shrinking at rate T.

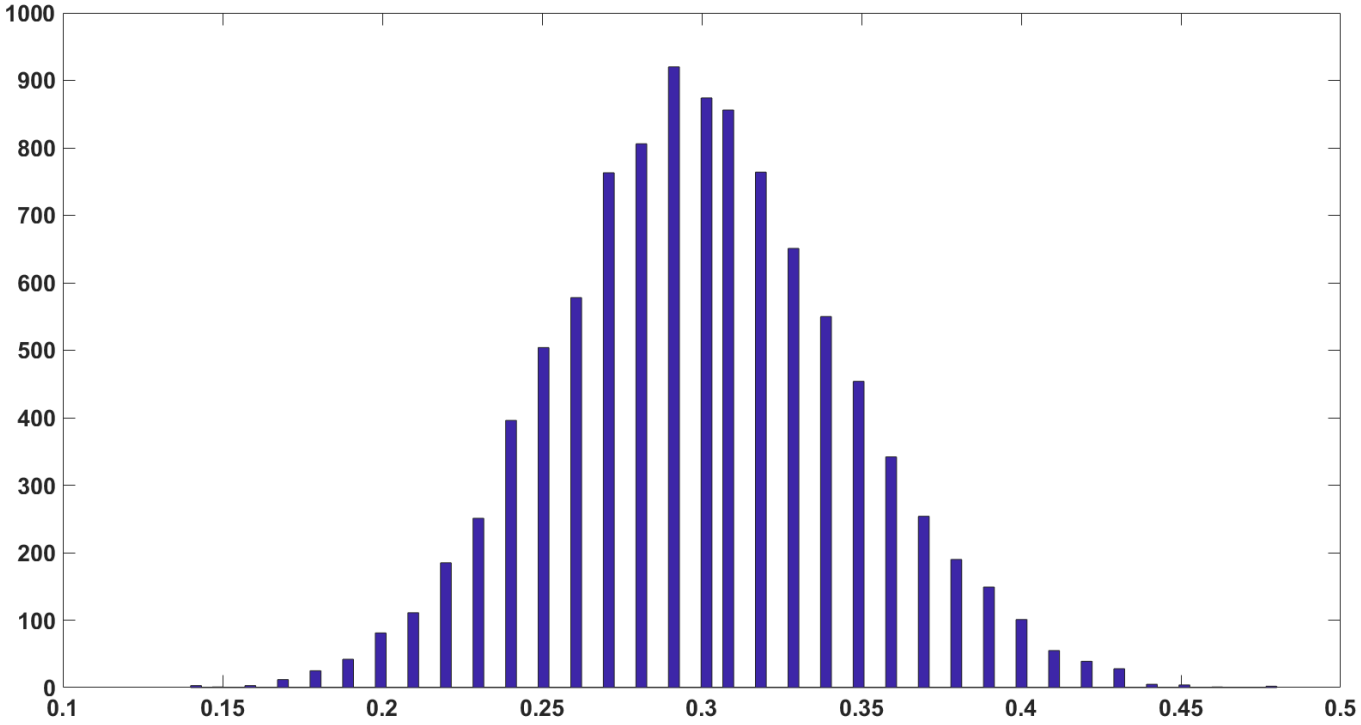
Second Pillar of statistics

Illustration:

Poll: “Do you believe that God can create a stone that He cannot carry ? ” **30%** ✓

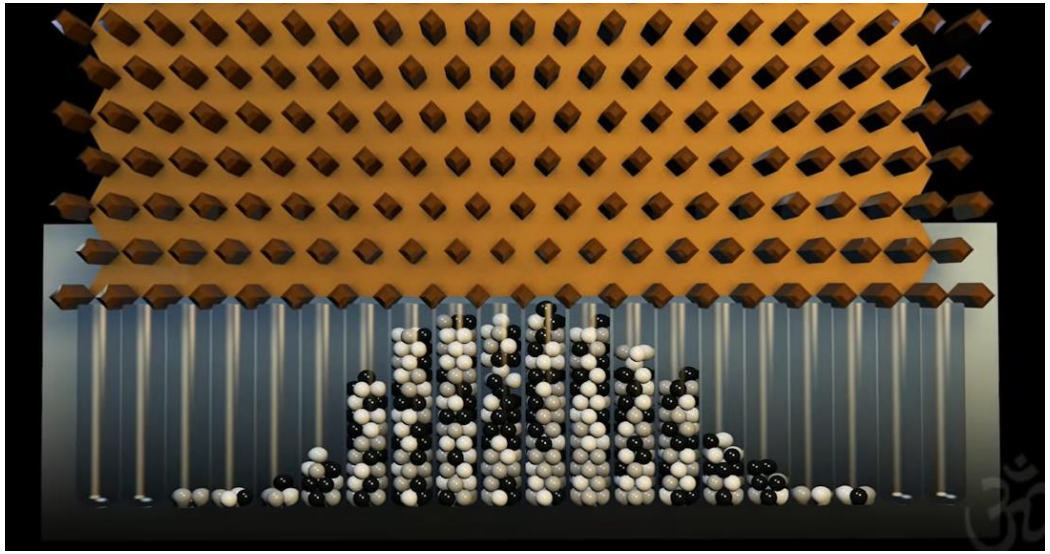
Poll of 100 people: 32%, 25%, 35%, 20%, 28%, 27%, …

➡ Realizations of a Normal distribution!



https://www.youtube.com/watch?v=GQru3xCLDX0

Each ball is a poll



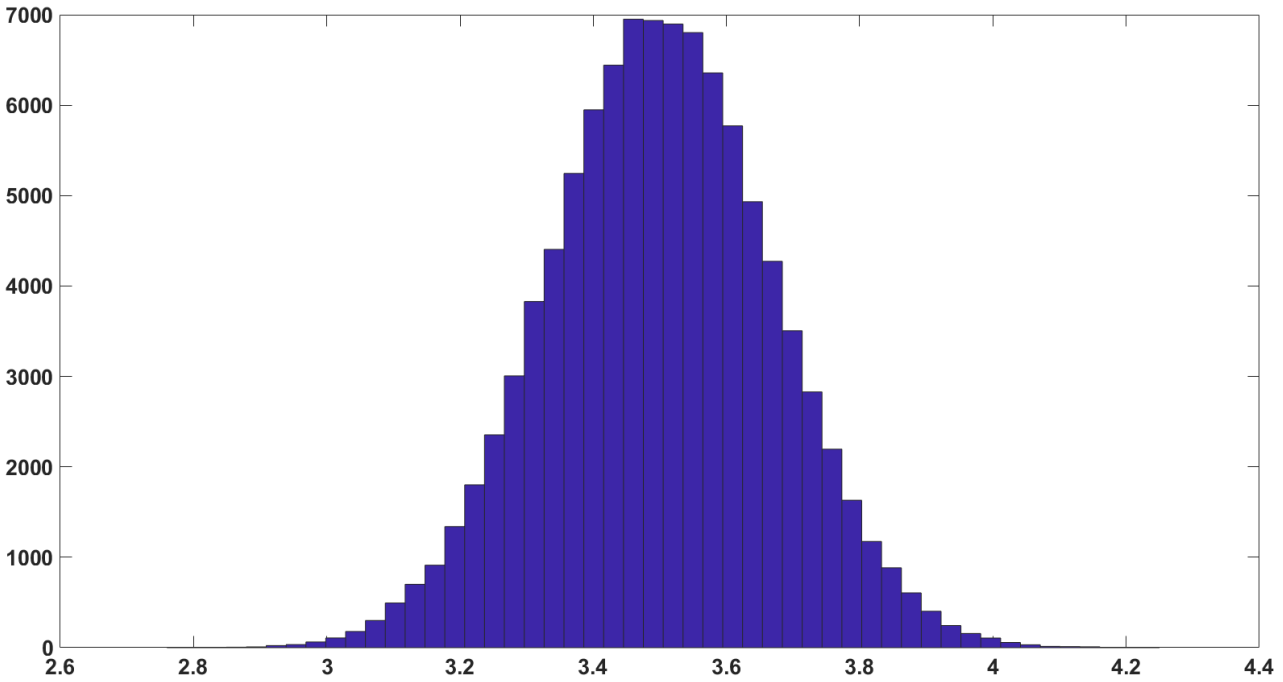
Second Pillar of statistics

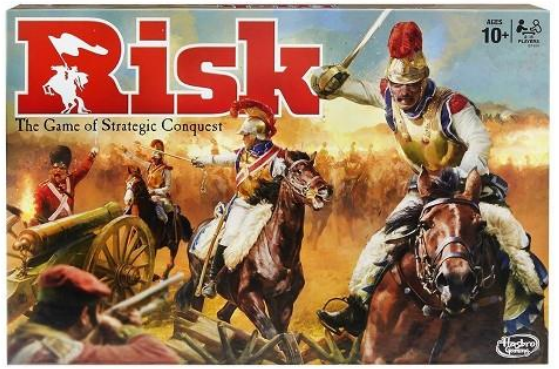
$E(X) = 3.5,\ \sigma^2 = V(X) = 2.92$

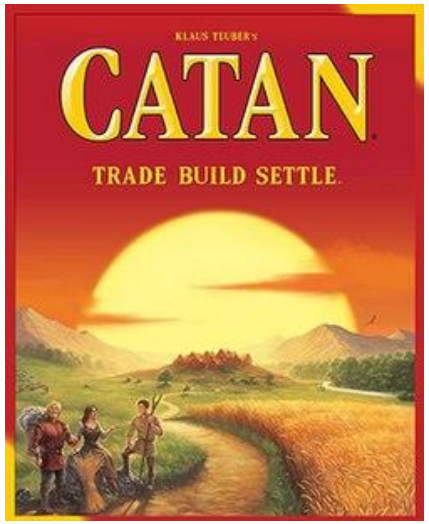
Illustration: Discrete Uniform distribution $P[X = 1] = P[X = 2] = \ldots = P[X = 6] = \frac{1}{6}$

Rolling 100 times a die (average): 3.36, 3.4, 3.6, 3.52, 3.33, …

⇒ Realizations of a Normal distribution!







$T = 100$ ⇒ $\frac{\sigma^2}{T} = 0.0292$ ⇒ $SE = \frac{\sigma}{\sqrt{T}} = 0.17$

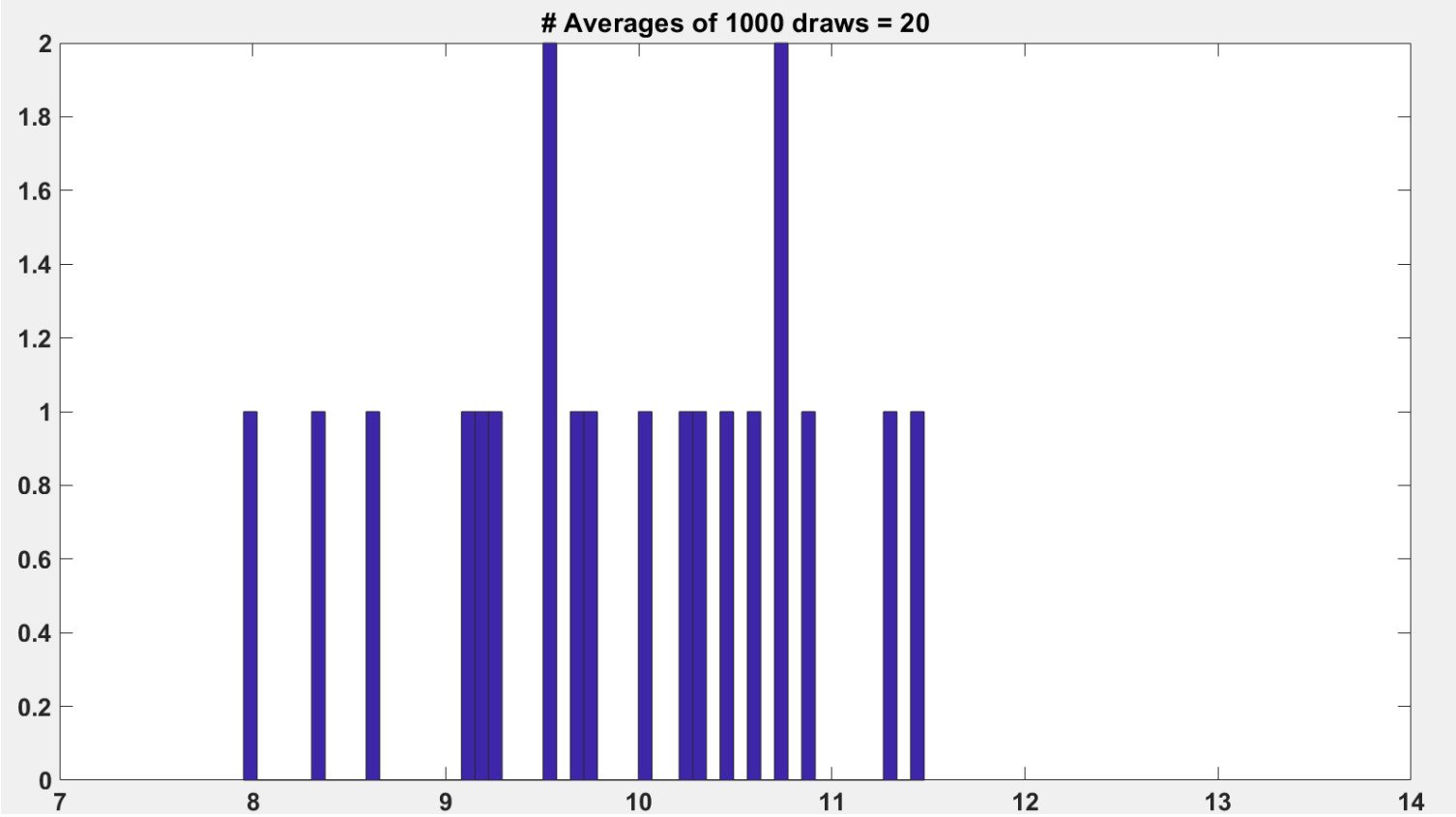
95% Interval: [3.5-1.96 (0.17) 3.5+1.96 (0.17)] = [3.17 3.83]

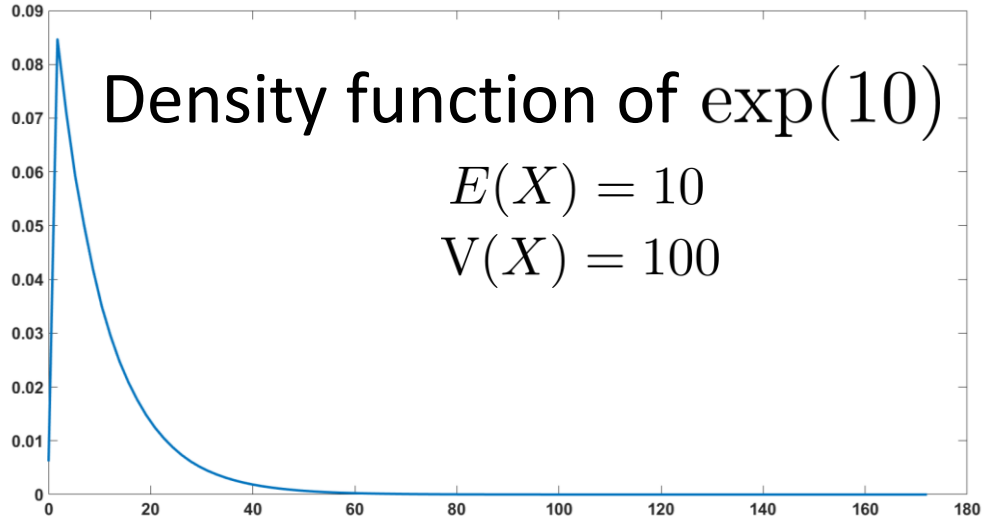
Second Pillar of statistics

Illustration: Waiting time for a bus ➡ Exponential distribution: $\exp(\lambda)$

Waiting 1000 times at the bus station (mn): 11.58, 9.46, 8.56, 7.75, …

➡ Averages are realizations of a Normal distribution!





Density function of $\exp(10)$

$E(X) = 10$

$\mathrm{V}(X) = 100$

Far away from a Normal distribution

CLT: Application

$$\hat{\beta}_2 = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

Find an estimator which depends on an average of the error terms

Your estimator will be normally distributed when T is large!

Most important statistical trick!

Linear regression: $y_t = \beta_1 + \beta_2 x_t + \epsilon_t, \text{ with } \epsilon_t \sim \text{iid}(0, \sigma^2)$

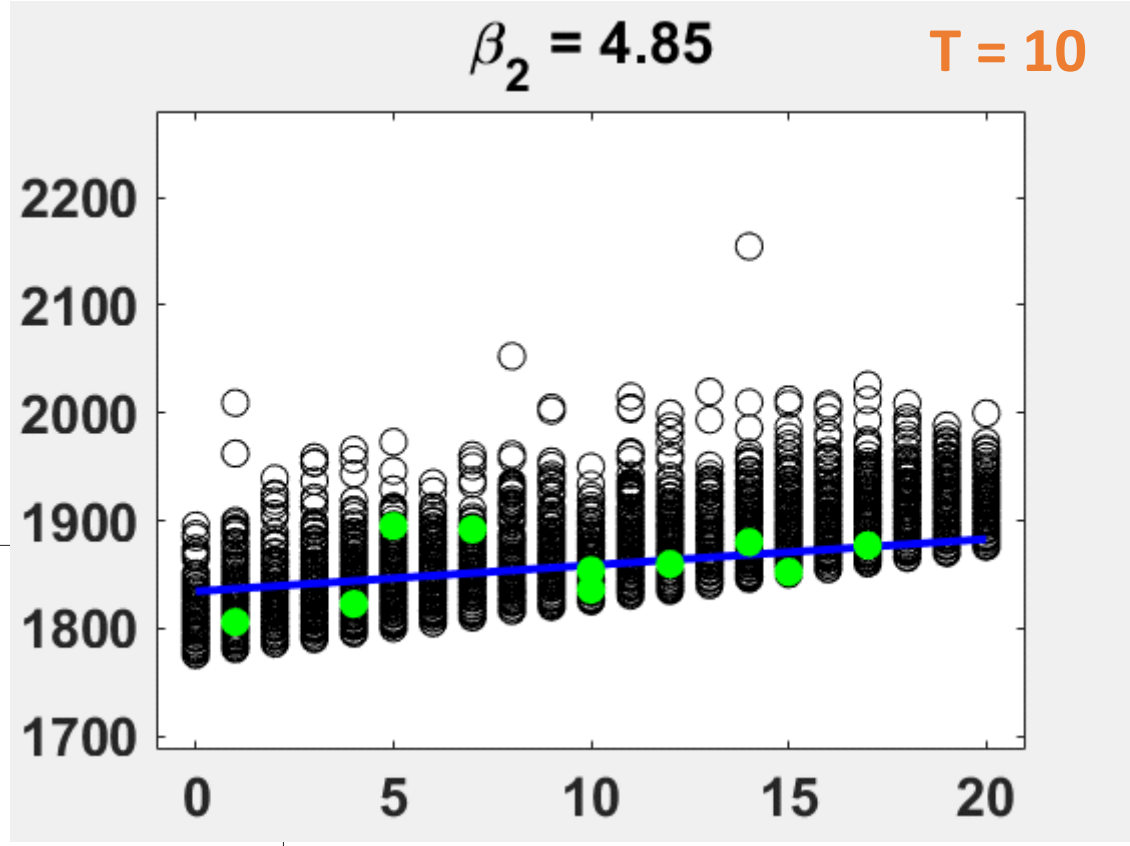
Ordinary least squares (OLS) estimator:

$$\begin{aligned}
\hat{\beta}_2 &= \beta_2 + \sum_{t=1}^{T} \frac{(x_t - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}[\epsilon_t - \frac{1}{T}\sum_{t=1}^{T}\epsilon_t] \\
&= \beta_2 + \sum_{t=1}^{T} \frac{(x_t - \bar{x})\epsilon_t}{\sum_{t=1}^{T}(x_t - \bar{x})^2}, \\
&= \beta_2 + \frac{1}{\frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2} \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})\epsilon_t
\end{aligned}$$

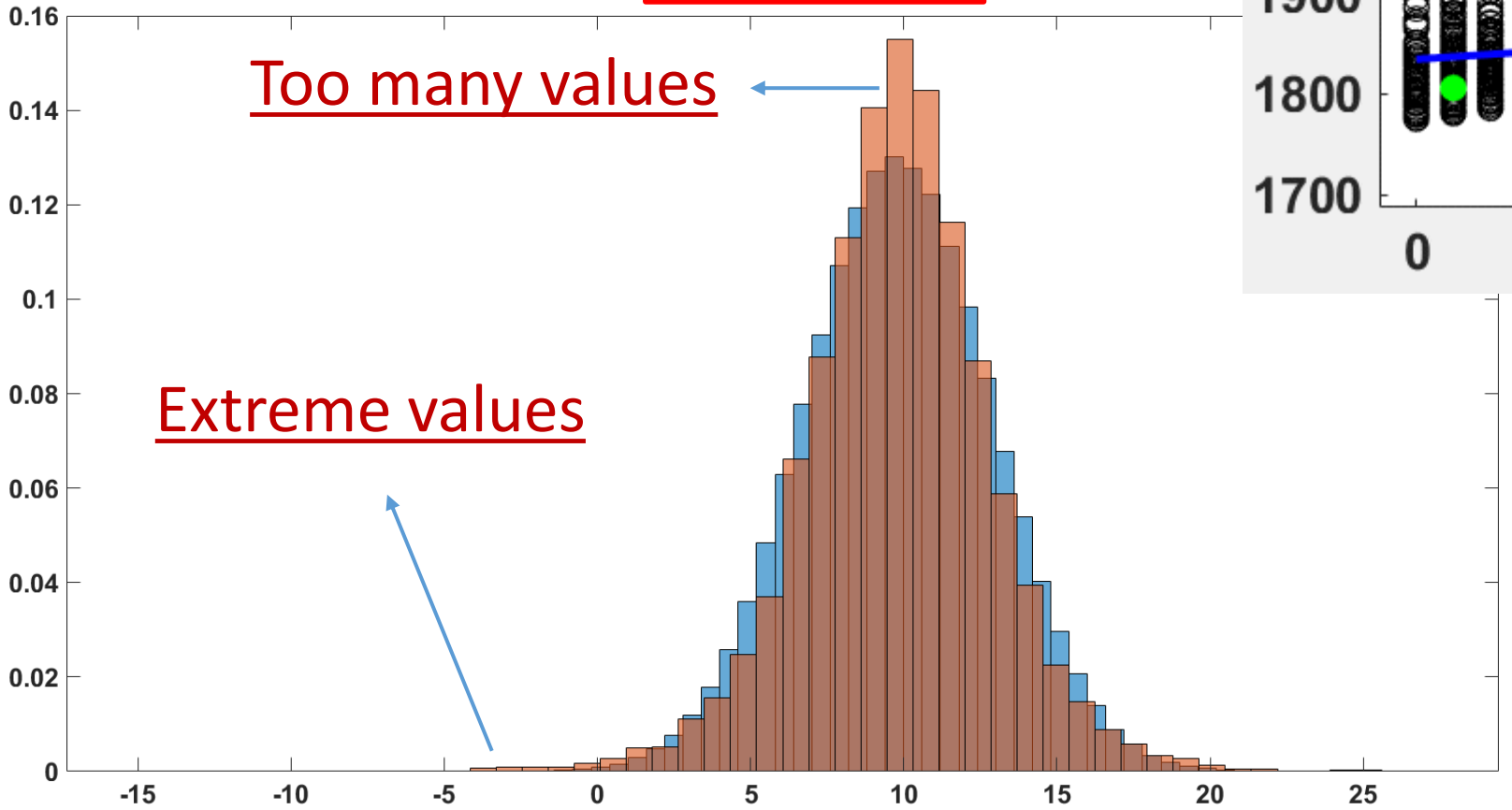
CLT: Application

Exact regression:

$$y_t = 1800 + 10x_t + \epsilon_t \text{ with } \epsilon_t \sim \exp(50) - 50$$



T = 10



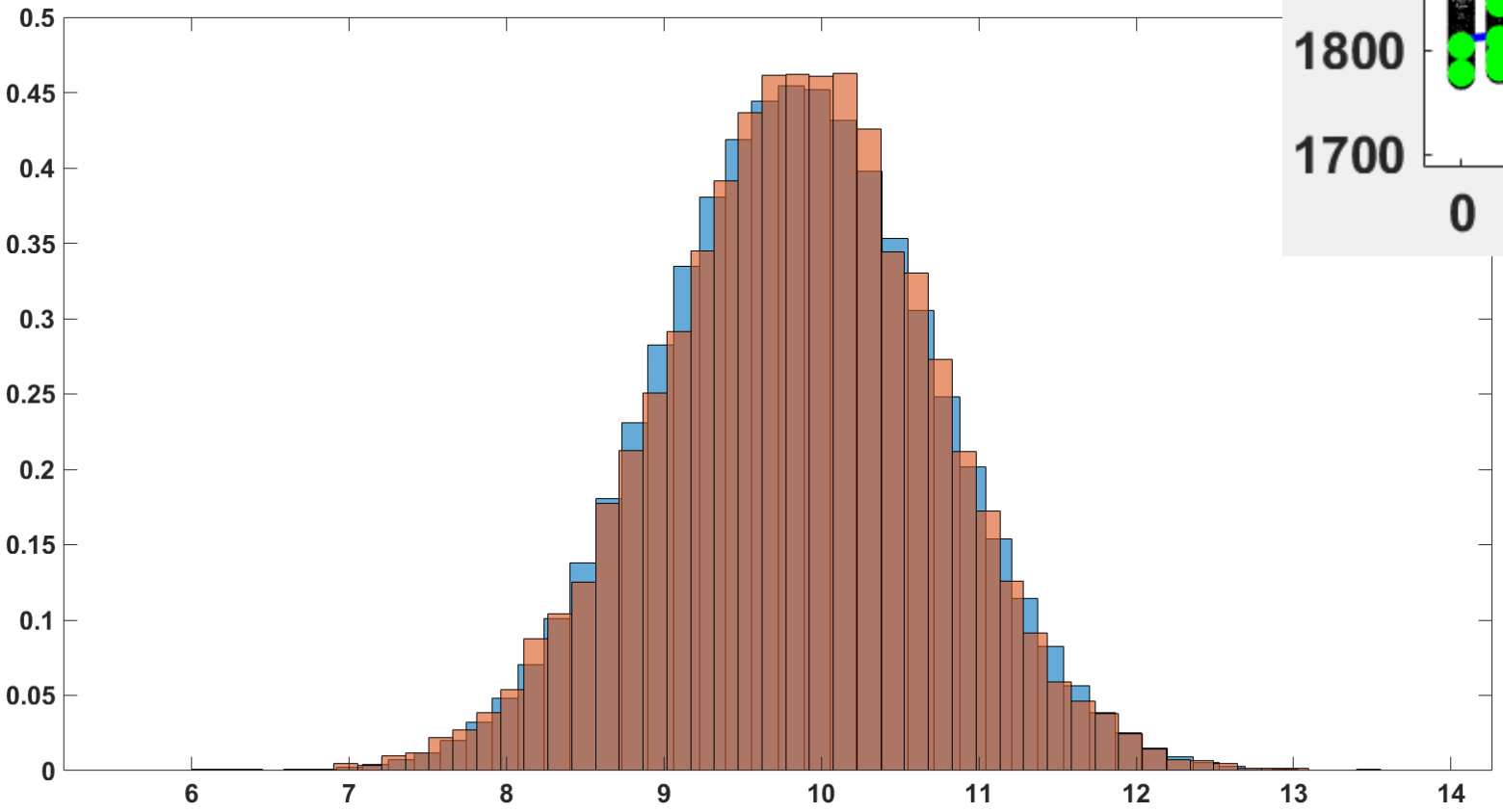
Unbiased estimator but not normally distributed.

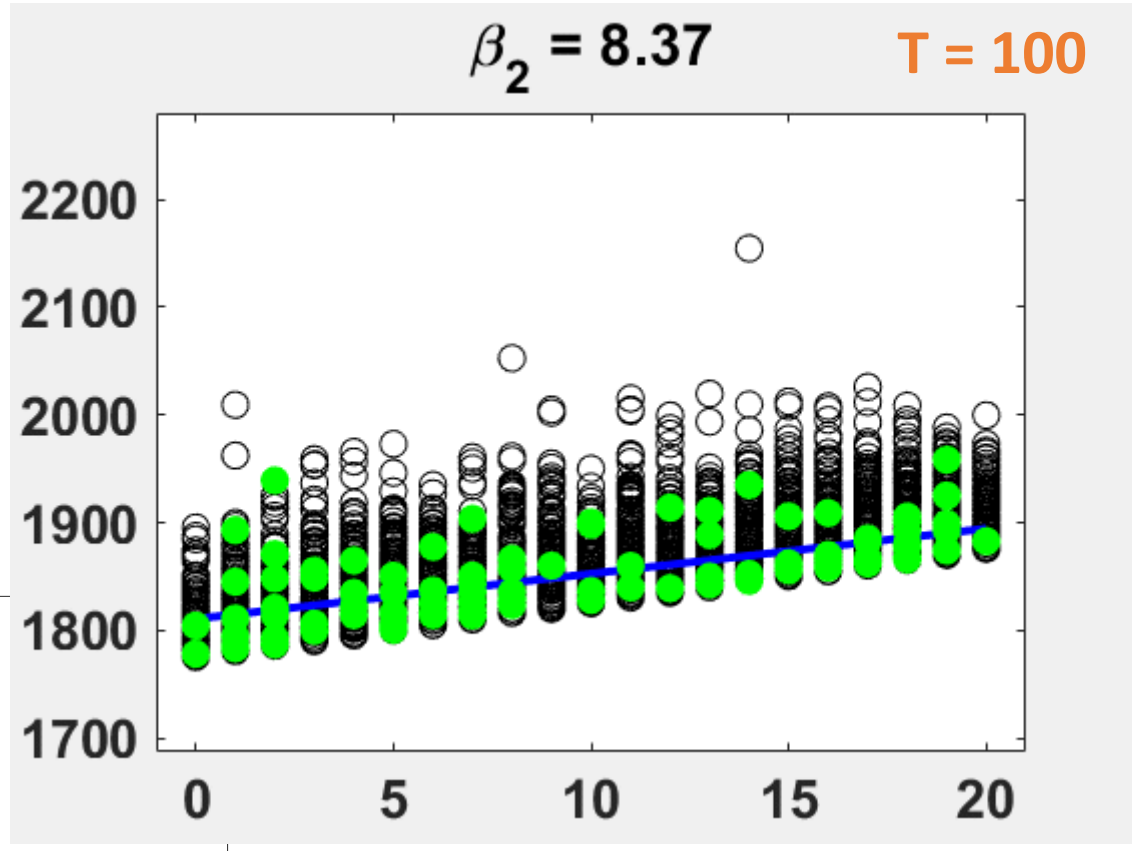
CLT: Application

Exact regression:

$$y_t = 1800 + 10x_t + \epsilon_t \text{ with } \epsilon_t \sim \exp(50) - 50$$

T = 100





Unbiased estimator and normally distributed.