Regression Models

Andrea Monticini
andrea.monticini@gmail.com

Università Cattolica

September 8, 2020

1. Empirical examples
   - The CAPM
   - House pricing
2. Introduction
3. Random Variables
4. Multivariate Distribution
5. Regression Models
6. Simulating Econometric Models
7. Method-of-Moments Estimation
8. Least Squares

Reference: ECTM Ch. 1

The CAPM

Consider the following problem. We want to estimate the beta (β_2) of a stock.

$$y_t = \beta_1 + \beta_2 x_t + u_t \tag{1}$$

where y_t represents the actual return to holding company's stock in month t, x_t, the return on the market portfolio in month t (i.e., the portfolio consisting of all stocks, held in the same proportion as in the market as a whole) and u_t represents other influences on returns.

Application to the CAPM

We want to estimate eq. (1) for return to holding Tesla (r_t)over the period 2013:11 - 2018:09 (monthly returns). In order to have a measure of the market return (r_{mt}) over the same time period, we use the Nasdaq return.

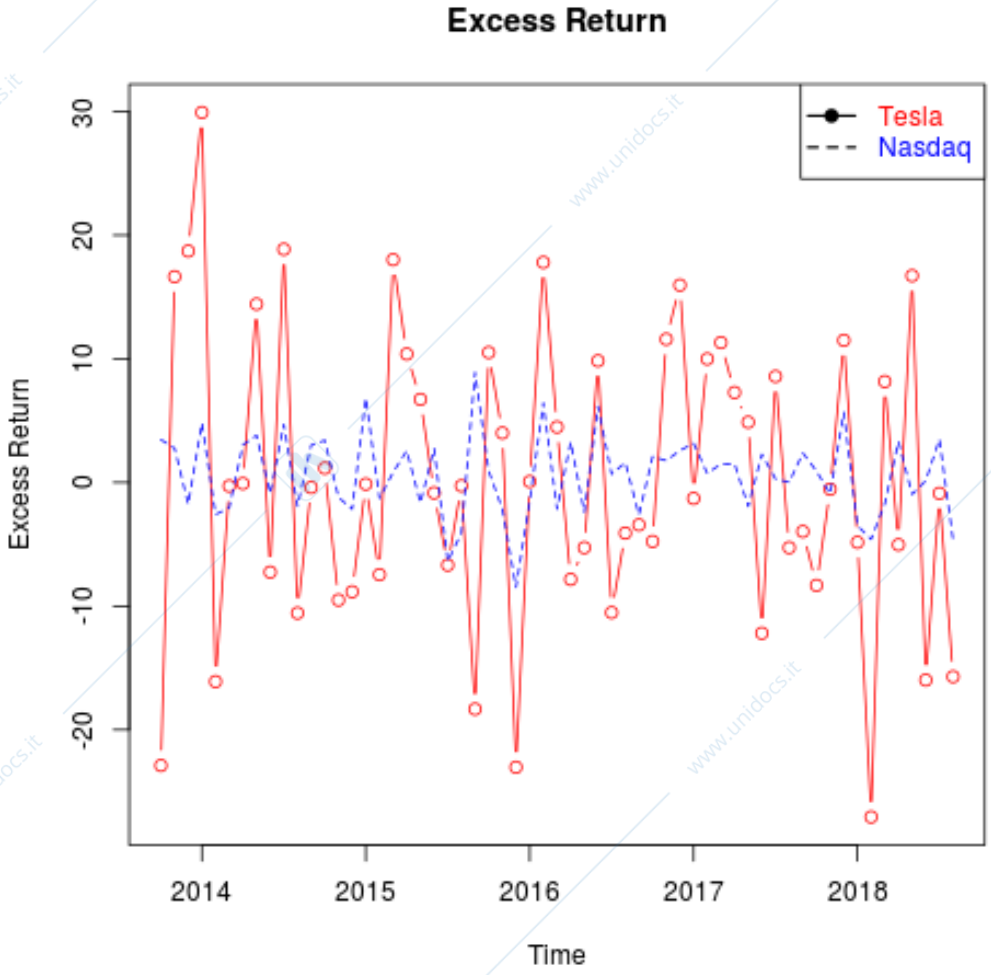
$$r_t = \beta_1 + \beta_2 r_{mt} + u_t$$

To conclude, we use the 3-month Tbill rate as a risk free interest rate.

Where to get the data?

Our dataset is based on Monthly observation downloaded from yahoo finance (we can import the data directly in R as well).

- Tesla ▸ link
- Nasdaq ▸ link
- 3 month T-Bill ▸ link



The CAPM

The strict form of the capital asset pricing model (CAPM) predicts that this equation fully "explains" stock returns. Specifically, this means that u_t depends only on random effects particular to that company, and is not predictable by macroeconomic variables. According to the CAPM, when markets operate efficiently in response to complete information, market equilibrium implies that x_t contains all such information, relevant to individual stock returns. This assumption can be tested econometrically.

The CAPM

The parameter β_2 (the "beta") is an indicator of the risk and return associated with the stock. When $\beta_2 = 1$, the expected net return is the same as that of the market portfolio. When $\beta_2 > 1$, the expected return exceeds that of the market portfolio, but there is correspondingly greater risk. When $\beta_2 < 1$ there is lower return, but also less risk. Thus, market equilibrium ensures the existence of a risk-return trade-off. The model also predicts that $\beta_1 = 0$.

We want to know whether:

- $\beta_1 = 0$ or $\beta_1 \neq 0$
- $\beta_2 = 1$ or $\beta_2 > 1$

House pricing

Problem

The data set "hprice.data" contains data on 546 houses sold in Windsor, Canada. We are interested in knowing what determines the house prices. To do that we have data on the lot size of the property (in square feet) x_2, the number of bedrooms x_3, the number of bathrooms x_4, the number of storeys (excluding the basement) x_5.

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + \beta_5 x_{5t} + u_t \tag{2}$$

where y is the sale price of the house in Canadian dollars, and $u_t \sim$ i.i.d.$(0, \sigma^2)$.

Introduction

$$y_t = \beta_1 + \beta_2 x_t + u_t \tag{3}$$

- t is used to index the observations of a sample
- The total number of observations, called sample size, will be denoted by n.
- The subscript t runs from 1 to n
- Each observation comprises an observation on a:
  - dependent variable, y_t
  - explanatory variable (independent variable), x_t
- The relation (3) links the observations on the dependent and the explanatory variables for each observation in terms of
  - two unknown parameters, β_1 and β_2
  - an unobserved error term, u_t
- Of the five quantities in equation (3), y_t and x_t are observed, and three β_1, β_2, u_t are not.

- The variable that appear in an econometric model are **random variables**. To characterize a random variable, we must specify the set of all possible values that the random variable can take on.
- The simplest case is a scalar random variable $(1, 2, 3)$
- We observe realizations of random variables. For a scalar r.v. we observe a single real value.
- If X is any r.v., probabilities can be assigned to subsets of the full set of possibilities of values for X. Such subset are called events, and their probabilities are assigned by a probability distribution, according to few general rules.

Discrete Random Variables

- When X is a discrete random variable can take on a finite number of values, which we denote as $x_1, , x_2, \ldots$. The probability distribution assigns probabilities, that is, numbers between 0 and 1 to each of these values, in such a way that the probabilities sum to 1

$$\sum_{i=1}^{\infty} p(x_i) = 1 \tag{4}$$

  where $p(x_i)$ is the probability assigned to x_i
- In econometrics the most commonly encountered discrete random variables occur in the context of binary data, which can take on the values 0 and 1

Continuous Random Variables

- When X is a continuous random variable can take on any value in some continuous subset of the real line, or possibly the whole real line. The probability distribution can be represented by a cumulative distribution function CDF. This function $F(x)$ is defined on the real line. Its value is $\Pr(X \leq x)$
- The dependent variable in a regression model is normally a continuous r.v.

Probability Distributions

Three rules:

1. All probabilities lie between 0 and 1;
2. The null set is assigned probability 0, and the full set of possibilities is assigned probability 1;
3. The probability assigned to an event that is the union of two disjoint events is the sum of the probabilities assigned to those disjoint events.

- $F(x)$ tends to 0 as $x \to -\infty$
- $F(x)$ tends to 1 as $x \to +\infty$
- $F(x)$ is a weakly increasing function of x
- For a continuous r.v., the CDF assigns probabilities to every interval on the real line. However, if we try to assign a probability to a single point, the result is always zero.

Probability Density Function

- The density function (f) is the derivative (if it exists) of the CDF

$$f(x) = F'(x)$$

- Because $\mathrm{F}(-\infty) = 0$ and $\mathrm{F}(\infty) = 1$, every PDF must be normalized to integrate to unity

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d} = \int_{-\infty}^{\infty} F'(x)\,\mathrm{d} = F(\infty) - F(-\infty) = 1 \tag{5}$$

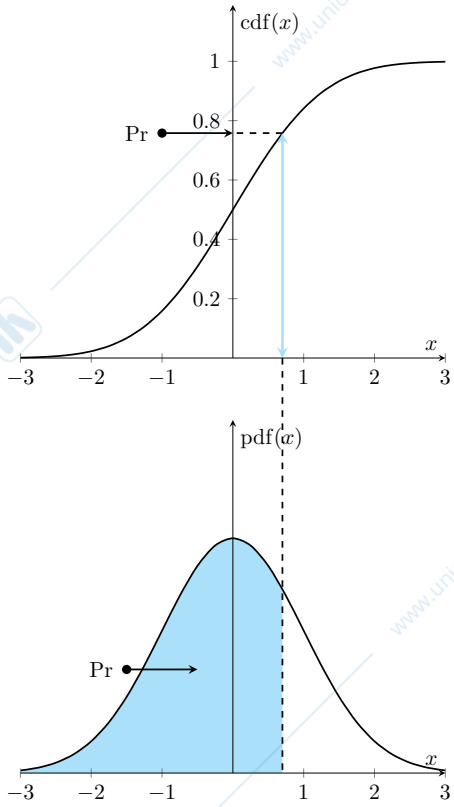
- Probabilities can be computed in terms of the PDF as well as the CDF.

$$\Pr(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(x)\,\mathrm{d}x \tag{6}$$

- The PDF of the standard normal distribution is

$$\phi(x) = (2\pi)^{(-1/2)} \exp\left(-\frac{1}{2}x^2\right) \tag{7}$$

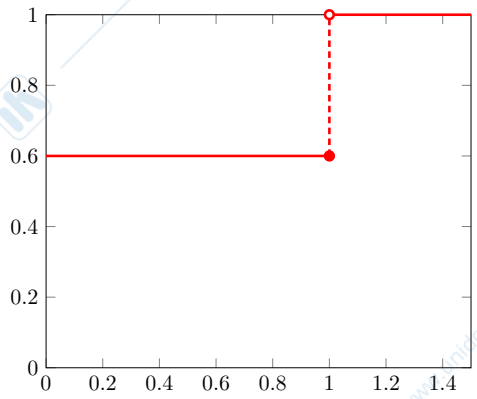
Standard Normal



Consider a binary r.v. X that can take on ony two values, 0 and 1, and let the probability that $X = 0$ be p. It follows that the probability that $X = 1$ is $1 - p$. The CDF in this case is

$$F(x) = \begin{cases} 0 \text{ for} & x < 0 \\ p \text{ for} & 0 \leq x < 1 \\ 1 \text{ for} & x \geq 1 \end{cases}$$

The CDF of a binary r.v. $p = 0.6$



Moments of Random Variables

The expectation of a random variable is called the **mean** or better the **population mean**. A common nation is μ. For a discrete r.v. we have

$$\mathbb{E}(X) = \sum_{i=1}^{m} p(x_i)x_i \tag{8}$$

while for a continuous r.v., we have

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x \tag{9}$$

The expectation of a r.v. is often referred to as its first moment. The higher moments, if they exist, are the expectation of the r.v. raised to a power.

$$m_k(X) = \int_{-\infty}^{\infty} x^k f(x)\,\mathrm{d}x \tag{10}$$

It is often more useful to work with central moments

$$\mu_k = \mathbb{E}\left(X - \mathbb{E}(X)\right)^k = \int_{-\infty}^{\infty} (x-\mu)^k f(x)\,\mathrm{d}x$$

where $\mu \equiv \mathbb{E}(X)$

For a discrete X, the k^{th} central moment is

$$\mu_k \equiv \mathbb{E}\left(X - \mathbb{E}(X)\right)^k = \sum_{i=1}^{m} p(x_i)\,(x_i - \mu)^k$$

The most important central moment is the second. It is called the variance of the random variable ans is written as $\mathrm{Var}(X)$

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x$$

Moments of a population distribution: mean, variance, standard deviation

- **Mean** = expected value (expectation) of Y ($\mathbb{E}(Y)$). It is the long-run average value of Y over repeated realizations of Y
- **Variance** = measure of the squared spread of the distribution
- **Standard Deviation** = $\sqrt{\text{variance}}$. Estimates of the standard deviations are often referred to as **standard errors**.

- **Skewness** $= \frac{\mathbb{E}[Y-\mu_Y]^3}{\sigma_Y^3}$ measure of asymmetry of a distribution
  - skewness $= 0$: distribution is symmetric skewness $> (<) 0$: distribution has long right (left) tail
- **Kurtosis** $= \frac{\mathbb{E}[Y-\mu_Y]^4}{\sigma_Y^4}$ measure of mass in tails (measure of probability of large values)
  - kurtosis $= 3$: normal distribution
  - kurtosis > 3: heavy tails ("leptokurtotic")

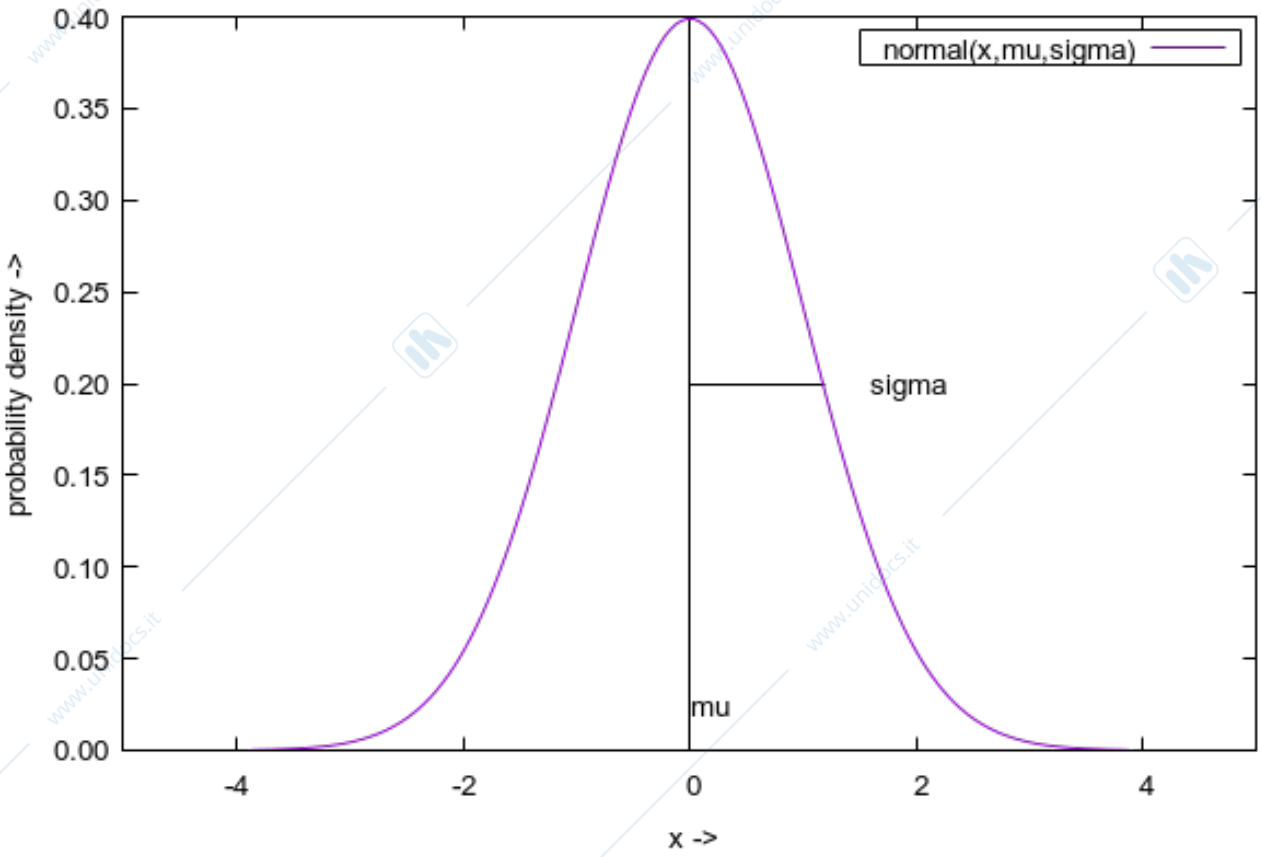
Gaussian distribution

The normal distribution is a bell-shaped distribution that is used most extensively in statistical applications in a wide variety of fields. Its density function is

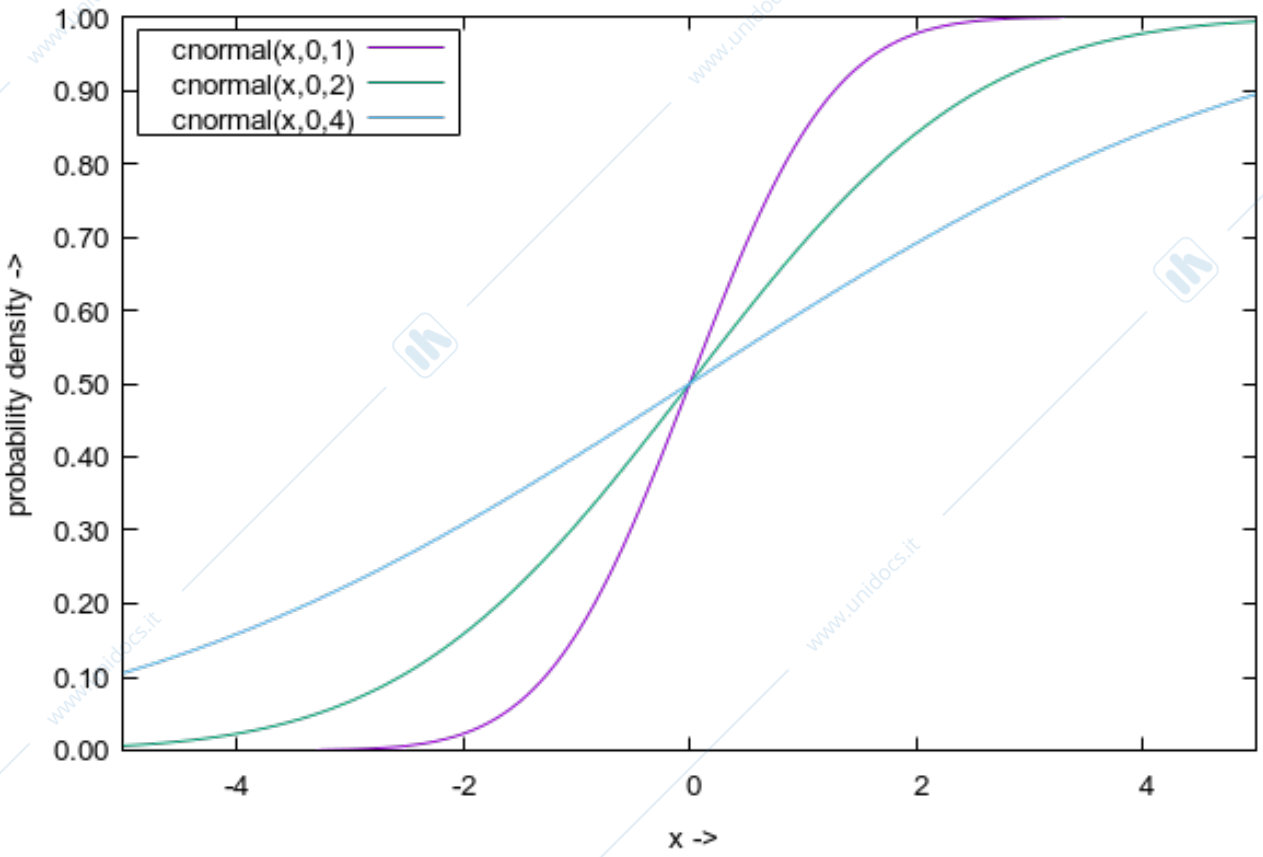
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right], -\infty < x < +\infty$$

Its mean is μ and its variance is σ^2. When x has a normal distribution with mean μ and variance σ^2, we write $x \sim \mathcal{N}(\mu, \sigma^2)$

PDF gaussian distribution



Cumulative Distribution Function gaussian distribution



A few different Normal distribution

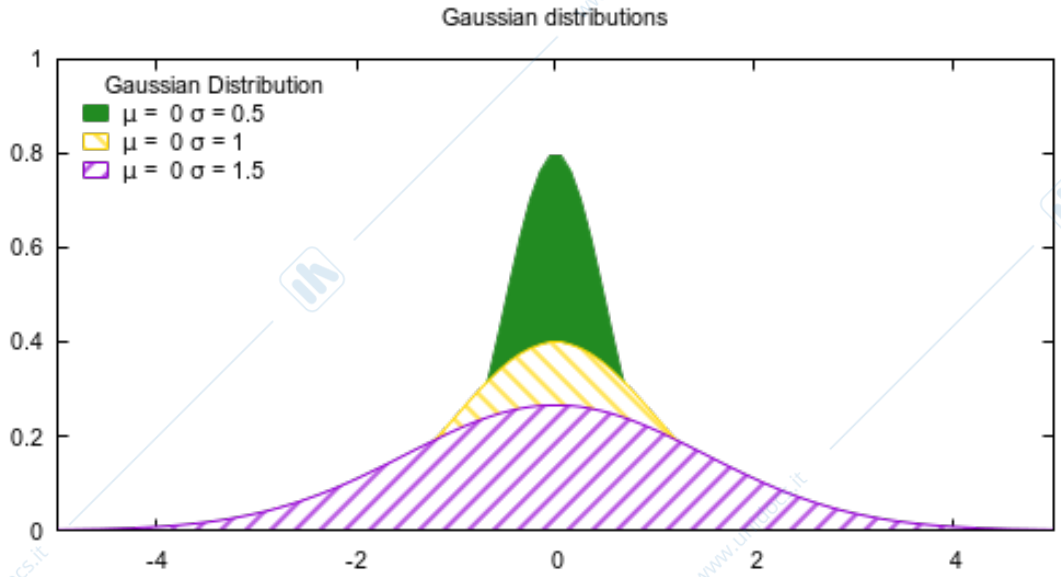


Figure 1: Gaussian distributions

A few different Normal distribution

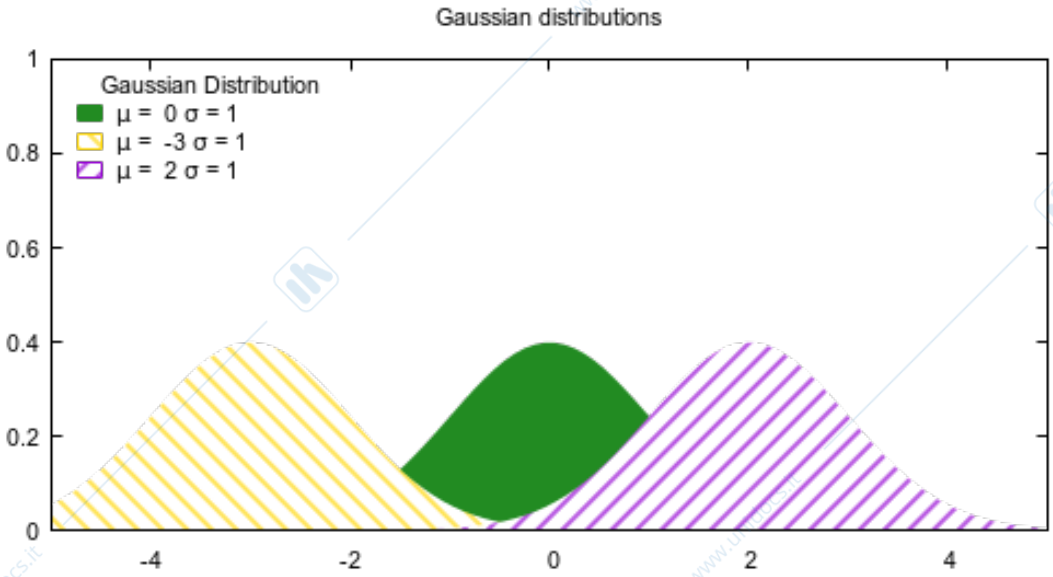


Figure 2: Gaussian distributions

Standard normal distribution

The expected value $\mathbb{E}(X)$ is

$$
\begin{aligned}
&= \int_{-\infty}^{+\infty} x f_X(x)\,\mathrm{d}x \\
&= (2\pi)^{-1/2} \int_{-\infty}^{+\infty} x \exp\left[-\frac{1}{2}x^2\right] \mathrm{d}x \\
&= (2\pi)^{-1/2} \int_{-\infty}^{0} x \exp\left[-\frac{1}{2}x^2\right] \mathrm{d}x + (2\pi)^{-1/2} \int_{0}^{\infty} x \exp\left[-\frac{1}{2}x^2\right] \mathrm{d}x
\end{aligned}
$$

We can solve

$$\int_0^\infty x \exp\left[-\frac{1}{2}x^2\right] \mathrm{d}x$$

by substitutions. Take

$$\begin{aligned} u &= -\frac{1}{2}x^2 \\ \mathrm{d}u &= -x\,\mathrm{d}x \\ x\,\mathrm{d}x &= -\mathrm{d}u \end{aligned}$$

Substituting we get

$$-\int_0^\infty \exp[u]\,\mathrm{d}u = -\exp[u] = -\exp\left[-\frac{1}{2}x^2\right]\Big|_0^\infty$$

So that

$$\begin{aligned} &= (2\pi)^{-1/2}\left[-\exp\left[-\frac{1}{2}x^2\right]\right]\Big|_{-\infty}^{0} + (2\pi)^{-1/2}\left[-\exp\left[-\frac{1}{2}x^2\right]\right]\Big|_0^{+\infty} \\ &= (2\pi)^{-1/2}[-1+0] + (2\pi)^{-1/2}[0+1] = 0 \end{aligned}$$

Now for the variance we have that $\mathbb{E}[X^2]$ is

$$
\begin{aligned}
&= \int_{-\infty}^{+\infty} x^2 f_X(x)\,\mathrm{d}x \\
&= (2\pi)^{-1/2} \int_{-\infty}^{+\infty} x^2 \exp\left[-\frac{1}{2}x^2\right] \mathrm{d}x \\
&= (2\pi)^{-1/2} \left\{ \int_{-\infty}^{0} x \left(x \exp\left[-\frac{1}{2}x^2\right] \right) \mathrm{d}x + \right. \\
&\quad \left. + \int_{0}^{\infty} x \left(x \exp\left[-\frac{1}{2}x^2\right] \right) \mathrm{d}x \right\}
\end{aligned}
$$

Integrating by parts

$$\int_{-\infty}^{0} x \left(x \exp\left[-\frac{1}{2} x^2 \right] \right) \mathrm{d}x,$$

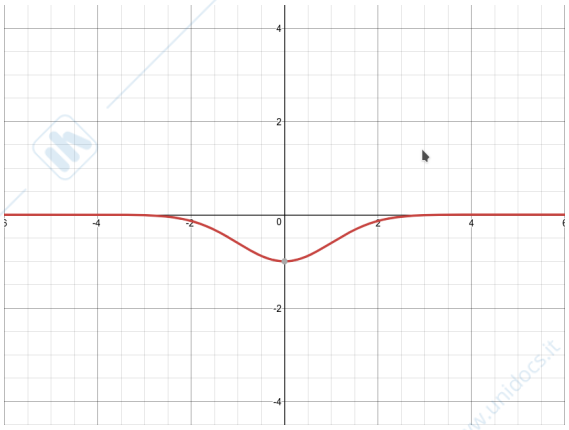
we get

$$\begin{aligned} f(x) &= x \\ \frac{\mathrm{d}f}{\mathrm{d}x} &= 1 \\ g(x) &= -\exp\left[-\frac{1}{2} x^2 \right] \\ \frac{\mathrm{d}g}{\mathrm{d}x} &= x \exp\left[-\frac{1}{2} x^2 \right] \mathrm{d}x \end{aligned}$$

$$f(x)g(x) - \int f'(x)g(x)\,\mathrm{d}x$$

$$
\begin{aligned}
&= (2\pi)^{-1/2}\left\{\left[-x\exp\left[-\frac{1}{2}x^2\right]\right]_{-\infty}^{0}+\int_{-\infty}^{0}\exp\left[-\frac{1}{2}x^2\right]\mathrm{d}x+\right.\\
&\quad\left.+\left[-x\exp\left[-\frac{1}{2}x^2\right]\right]_{0}^{\infty}+\int_{0}^{\infty}\exp\left[-\frac{1}{2}x^2\right]\mathrm{d}x\right\}\\
&= (2\pi)^{-1/2}\left\{[0-0]+[0-0]+\right.\\
&\quad\left.+\int_{-\infty}^{0}\exp\left[-\frac{1}{2}x^2\right]\mathrm{d}x+\int_{0}^{\infty}\exp\left[-\frac{1}{2}x^2\right]\mathrm{d}x\right\}\\
&= (2\pi)^{-1/2}\int_{-\infty}^{+\infty}\exp\left[-\frac{1}{2}x^2\right]\mathrm{d}x=\int_{-\infty}^{+\infty}f_X(x)\,\mathrm{d}x=1.
\end{aligned}
$$

$$-\exp(-\frac{1}{2}x^2)$$



while $\mathbb{E}[X]^2 = 0$, so that $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 1 - 0$.
We can write $x \sim \mathcal{N}(0,1)$ (**standard normal**), then for example, $P(x \geq 0) = 0.5,\ P(-2 \leq x \leq 2) \approx 0.96$

Multivariate Distributions

A **vector-valued** random variable takes on values that are vectors. It can be thought of as several scalar random variables that have a single, joint distribution. A continuous, bivariate r..v. (X_1, X_2) has distribution unction

$$F(x_1, x_2) = \Pr\left((X_1 \leq x_1) \cap (X_2 \leq x_2)\right)$$

For continuous variables, the PDF, if it exists, is the joint density

$$f(x_1, x_2) = \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2}$$

$$F(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(y_1, y_2)\, \mathrm{d}y_1\, \mathrm{d}y_2 \tag{11}$$

If two random variables happen to be independent of each other, then the joint density function is just the product of the two univariate densities

$$F(x_1, x_2) = F(x_1, \infty)F(\infty, x_2) \tag{12}$$

The first factor is the joint probability that $X_1 \leq x_1$ and $X_2 \leq \infty$. Since the second inequality imposes no constraint, this factor is the probability that $X_1 \leq x_1$. The function $F(x_1, \infty)$, which is called marginal CDF of X_1, is the CDF of X_1 considered by itself.

Conditional Probabilities

Suppose that A and B are any two events. Then the probability of event A conditional on B, or given B, is denoted as $\Pr(A \mid B)$ and is defined implicitly by the equation

$$\Pr(A \cap B) = \Pr(B)\Pr(A \mid B) \tag{13}$$

The idea underlying the definition is that, if we know somehow that the event B has been realized, this knowledge can provide information about whether event A has also been realized. For instance, if A and B are disjoint, and B is realized, then it is certain that A has not been.

Conditional Expectations

- Law of Iterated Expectations

$$\mathbb{E}\left(\mathbb{E}\left(X_1|X_2\right)\right) = \mathbb{E}\left(X_1\right) \tag{14}$$

- Conditional expectation and deterministic function

$$\mathbb{E}\left(X_1 h(X_2)|X_2\right) = h(X_2)\mathbb{E}\left(X_1|X_2\right) \tag{15}$$

for any deterministic function $h(.)$

The Specification of Regression Models

$$y_t = \beta_1 + \beta_2 x_t + u_t \tag{16}$$

The model (16) can be interpreted as a model for the mean of y_t conditional on x_t. Assume that the error term u_t has mean 0 conditional on x_t, we get

$$\mathbb{E}\left(y_t \mid x_t\right) = \beta_1 + \beta_2 x_t + \mathbb{E}\left(u_t \mid x_t\right) = \beta_1 + \beta_2 x_t \tag{17}$$

Information sets

- We are interested in the mean of y_t conditional on a set of potential explanatory variables. This set is called an **information set**, and it is denoted by Ω_t
- We want to condition on **exogenous** variables but not on **endogenous** ones
  - An exogenous variable has its origin outside the model under consideration, while the mechanism generating an endogenous variable is inside the model.
- The consumption function (16) can be correctly specified only if two conditions hold:
  - The mean of the consumption on disposable income is a linear function of the later
  - Consumption is not a variable that contributes to the determination of disposable income

Error Terms

It is essential to make assumptions about the properties of the error terms. The simplest assumptions is that all the error terms have mean 0, come from the same distribution, and are independent of each other.

- The error terms are independently and identically distributed, or IID
- When the observations are ordered by time, it seems plausible that an error term is correlated with neighboring error terms. This phenomenon is called serial correlation. When there is serial correlation, the error terms cannot be IID because they are not independent.
- The variance of the error terms may by systematically larger for some observations than for others. This phenomenon is called heteroskedasticity. When there is heteroskedasticity, the error terms cannot be IID.

Interpretation of the coefficients

β_j can be interpreted as a marginal effect. In particular, β_j is the marginal effect of x_j on y, holding all other explanatory variables constant.

$$y = \beta_1 + \beta_2 x_2 + u_t \tag{18}$$

we can add $\triangle x_2$ and we get

$$y + \triangle y = \beta_1 + \beta_2 (x_2 + \triangle x_2) + u_t \tag{19}$$

$$\begin{aligned} (19)\text{-}(18) &= \triangle y = \beta_2 \triangle x_2 = \\ \beta_2 &= \frac{\triangle y}{\triangle x_2} \\ \beta_1 &= \text{predicted value of } y \text{ when } x_2 = 0 \end{aligned}$$

Simulating Econometric Models

- Fix the sample size, n;
- Choose the parameters (β_1 and β_2) of the deterministic specification;
- Obtain the n values x_t, $n = 1, ..., n$;
- Evaluate the n successive values of the regression function $\beta_1 + \beta_2 x_t$ for $t = 1, ..., n$;
- Choose the probability distribution of the error terms;
- Use a random-number generator to generate the n successive and mutually independent values u_t of the error terms;
- Form the n successive values y_t of the dependent variable by adding the error terms to the values of the regression function.

Simulating Econometric Models

```
s_size<-1000
beta1<-2
beta2<-8
x<-rnorm(s_size, mean = 5, sd = 1)
y <- 1:s_size; 0
    for (i in 1:s_size)
    {
        eps<-rnorm(1, mean = 0, sd = 1)
        y[i]<-beta1+beta2*x[i]+eps
    }
```

Linear and Nonlinear Regression Models

Consider the following models:

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 x_t^2 + u_t \quad (20)$$

$$y_t = \gamma_1 + \gamma_2 \log x_t + u_t \quad (21)$$

$$y_t = \delta_1 + \delta_2 \frac{1}{x_t} + u_t \quad (22)$$

In equation (20), there is an extra param., β_3, which allows $\mathbb{E}(y_t \mid x_t)$ to vary quadratically with x_t whenever β_3 is nonzero. In the model (21) and (22) a nonlinear transformation of x_t is used in place of x_t itself. As a consequences, the relationship between x_t and $\mathbb{E}(y_t \mid x_t)$ is nonlinear.

(20), (21), (22) are all said to be linear regression models, because, even though the mean of y_t may depend nonlinearly on x_t, it always depends linearly on the unknown parameters of the regression function.

Consider the following model

$$y_t = \mathrm{e}^{\beta_1} + x_{t2}^{\beta_2} + x_{t3}^{\beta_3}\mathrm{e}^{v_t} \tag{23}$$

Take the log

$$\log y_t = \beta_1 + \beta_2 \log(x_{t2}) + \beta_3 \log(x_{t3}) + v_t \tag{24}$$

which is a loglinear regression model

How to interpret the coefficients

Three different cases:

- $y_t = \beta_1 + \beta_2 \log(x_t) + u_t$
- $\log(y_t) = \beta_1 + \beta_2 x_t + u_t$
- $\log(y_t) = \beta_1 + \beta_2 \log(x_t) + u_t$

where log is the natural logarithm

Consider the first case

$$y = \beta_1 + \beta_2 \log(x) \tag{25}$$

we can add $\triangle x$ and we get

$$y + \triangle y = \beta_1 + \beta_2 \log(x + \triangle x) \tag{26}$$

we can compute

$$(26)\text{-}(25) = \triangle y = \beta_2[\log(x + \triangle x) - \log(x)]$$

given $\log(x + \triangle x) - \log(x) = \frac{\triangle x}{x}$, we obtain

$$\beta_2 \approx \frac{\triangle y}{\frac{\triangle x}{x}}$$

if x increases 5%, we have $\frac{\triangle x}{x} = 0.05$ so y is $0.05 \times \beta_1$

In the second case we have

$$\log(y) = \beta_1 + \beta_2 x \tag{27}$$

now consider a little change in $\triangle x$, we get

$$\log(y + \triangle y) = \beta_1 + \beta_2(x + \triangle x) \tag{28}$$

$$(28)\text{-}(27) = \log(y + \triangle y) - \log(y) = \beta_1 \triangle x \tag{29}$$

so that

$$\frac{\triangle y}{y} = \beta_2 \triangle x$$

a small change in x produces $100\beta_2\%$ in y

Finally, we have the third case

$$\log(y) = \beta_1 + \beta_2 \log(x) \tag{30}$$

after a little change in x

$$\log(y + \triangle y) = \beta_1 + \beta_2 (\log (x + \triangle x)) \tag{31}$$

as we did before

$$(31)\text{-}(30) = \log(y + \triangle y) - \log(y) = \beta_2[\log(x + \triangle x) - \log(x)] \tag{32}$$

so that

$$\frac{\triangle y}{y} = \beta_2 \frac{\triangle x}{x}$$

a 1% increase in x will give β_2 % in y.

The Multiple Regression Model in Matrix Notation

$$\begin{aligned} y_1 &= \beta_1 + \beta_2 x_1 + u_1 \\ y_2 &= \beta_1 + \beta_2 x_2 + u_2 \\ \vdots &\quad \vdots \quad \vdots \quad \vdots \\ y_n &= \beta_1 + \beta_2 x_n + u_n \end{aligned}$$

This can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

or

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \tag{33}$$

The Multiple Regression Model in Matrix Notation

$$y_i = \beta_1 + \beta_2 x_{2i} + ... + \beta_k x_{ki} + u_i \; i = 1, 2, ..., n \tag{34}$$

This can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} & & x_{k1} \\ x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & & \vdots \\ x_{1n} & x_{2n} & & x_{kn} \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

or

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \tag{35}$$

- It is common in statistics to speak of the **population** from which a sample is drawn. Recall the use of **population mean** as as a synonym for the mathematical expectation.
- The **sample mean** is an estimate of the population mean
- A better concept is that of **data-generating process** (DGP). We mean whatever the mechanism is at work in the real world of economic activity giving rise to the numbers in our sample. A data-generating process is thus the analog in econometrics of a population in biostatistics. Samples may be drawn from a DGP as they may be drawn from a population.

Estimating the simple linear regression model

A natural way to estimate parameters is to replace population means by sample means. This technique is called the **method of moments**.
In order to apply this method to regression models, we must use the facts that population moments are expectations, and that regression models are specified in terms of the conditional expectations of the error terms.

Let see how the principle of replacing population means by sample means works for the simple linear regression model. The error term for observation t is

$$u_t = y_t - \beta_1 - \beta_2 x_t \tag{36}$$

and according to our model, the expectation of this error term is zero $\mathbb{E}(u_t)$.

Since we have n error terms for a sample of size n, we can consider the sample mean of the error terms:

$$\frac{1}{n}\sum_{t=1}^{n} u_t = \frac{1}{n}\sum_{t=1}^{n} (y_t - \beta_1 - \beta_2 x_t) \tag{37}$$

We would like to set this sample mean equal to zero.

Suppose that $\beta_2 = 0$. This reduces the number of parameters in the model to just one. In that case, there is just one value of α which allows the right-hand side of equation (37) to equal zero.
The equation defining this value is

$$\frac{1}{n}\sum_{t=1}^{n}(y_t - \beta_1) = 0 \tag{38}$$

This estimate is just the mean of the observed values of the dependent variable,

$$\widehat{\beta}_1 = \frac{1}{n}\sum_{t=1}^{n} y_t \tag{39}$$

How does method of moments work re-introducing β_2? The equation 38 would become

$$\frac{1}{n}\sum_{t=1}^{n}(y_t - \beta_1 - \beta_2 x_t) = 0 \tag{40}$$

This is one equation and two unknowns!
To obtain another equation we use the following argument. Our model specifies that the mean of u_t is 0 conditional on the explanatory variable x_t. The conditional mean assumption implies that not only $E(u_t) = 0$, but that $\mathbb{E}(x_t u_t) = 0$ as well. By the law of the iterated expectations, we have

$$\mathbb{E}(x_t u_t) = \mathbb{E}\left(\mathbb{E}\left(x_t u_t \mid x_t\right)\right) = \mathbb{E}\left(x_t \mathbb{E}\left(u_t \mid x_t\right)\right) = 0 \tag{41}$$

Thus we can supplement (40) by the following equation, which replaces the population mean in (41) by the corresponding sample mean,

$$\frac{1}{n}\sum_{t=1}^{n} x_t \left(y_t - \beta_1 - \beta_2 x_t\right) = 0 \tag{42}$$

The equations (40) and (42) are two linear equations in two unknowns, β_1 and β_2. Solving these equations yields the MM estimates

Since β_1 and β_2 do not depend on t, these two equations can be written as (system of Normal equations)

$$\begin{aligned} \beta_1 + \left(\frac{1}{n}\sum_{t=1}^{n} x_t\right)\beta_2 &= \frac{1}{n}\sum_{t=1}^{n} y_t \\ \left(\frac{1}{n}\sum_{t=1}^{n} x_t\right)\beta_1 + \left(\frac{1}{n}\sum_{t=1}^{n} x_t^2\right)\beta_2 &= \frac{1}{n}\sum_{t=1}^{n} x_t y_t \end{aligned}$$

Multiplying both equations by n, we have

$$\begin{bmatrix} n & \sum_{t=1}^{n} x_t \\ \sum_{t=1}^{n} x_t & \sum_{t=1}^{n} x_t^2 \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{n} y_t \\ \sum_{t=1}^{n} x_t y_t \end{bmatrix}$$

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \begin{bmatrix} \sum_{t=1}^{n} y_t \\ \sum_{t=1}^{n} x_t y_t \end{bmatrix}$$

and

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \begin{bmatrix} n & \sum_{t=1}^{n} x_t \\ \sum_{t=1}^{n} x_t & \sum_{t=1}^{n} x_t^2 \end{bmatrix}$$

So, we get

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{43}$$

The $\widehat{\boldsymbol{\beta}}$ that solves the above is called the **ordinary least squares** (OLS) estimator for the linear regression model.

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{44}$$

Estimating the Multiple Linear Regression Model

Since each explanatory variables is required to be in the information set Ω_t, we have

$$\mathbb{E}\left(x_{ti}u_t\right) = 0$$

So, we have

$$\frac{1}{n}\sum_{t=1}^{n} x_{ti}\left(y_t - \boldsymbol{X}_t\boldsymbol{\beta}\right) = 0 \tag{45}$$

where $\boldsymbol{X}_t$ denotes the th row of $\boldsymbol{X}$. Or in matrix form

$$\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) = \boldsymbol{0} \tag{46}$$

$$X^{\mathrm{T}}y = \begin{bmatrix} x_1^{\mathrm{T}} \\ x_2^{\mathrm{T}} \\ \vdots \\ x_t^{\mathrm{T}} \end{bmatrix} y = \begin{bmatrix} x_1^{\mathrm{T}}y \\ x_2^{\mathrm{T}}y \\ \vdots \\ x_k^{\mathrm{T}}y \end{bmatrix}$$

$$X^{\mathrm{T}}X = \begin{bmatrix} x_1^{\mathrm{T}} \\ x_2^{\mathrm{T}} \\ \vdots \\ x_t^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} x_1 & x_2 & \cdots & x_k \end{bmatrix} = \begin{bmatrix} x_1^{\mathrm{T}}x_1 & x_1^{\mathrm{T}}x_2 & \cdots & x_1^{\mathrm{T}}x_k \\ x_2^{\mathrm{T}}x_1 & x_2^{\mathrm{T}}x_2 & \cdots & x_2^{\mathrm{T}}x_k \\ \vdots & \vdots & \ddots & \vdots \\ x_k^{\mathrm{T}}x_1 & x_k^{\mathrm{T}}x_2 & \cdots & x_k^{\mathrm{T}}x_k \end{bmatrix}$$

Least-Squares Estimation

$$SSR(\boldsymbol{\beta}) = \sum_{t=1}^{n} (y_t - \boldsymbol{X}_t \boldsymbol{\beta})^2 \tag{47}$$

Consider our original model $y_t = \beta_1 + \beta_2 x_t$, when $\beta_2 = 0$, the above expression (47) becomes

$$SSR(\beta_1) = \sum_{t=1}^{n} (y_t - \beta_1)^2 = \sum_{t=1}^{n} y_t^2 + n\beta_1^2 - 2\beta_1 \sum_{t=1}^{n} y_t \tag{48}$$

The first-order condition for a minimum of SSR

$$\frac{\partial SSR}{\partial \beta_1} = 2\beta_1 n - 2\sum_{t=1}^{n} y_t = 0 \tag{49}$$

In this case the matrix $\boldsymbol{X} = \boldsymbol{\iota}$.
Therefore $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{\iota}^{\mathrm{T}}\boldsymbol{\iota} = n$, and $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{\iota}^{\mathrm{T}}\boldsymbol{y} = \sum_{t=1}^{n} y_t$.

$$\beta_1 = \frac{1}{n}\sum_{t=1}^{n} y_t = \left(\boldsymbol{\iota}^{\mathrm{T}}\boldsymbol{\iota}\right)^{-1}\boldsymbol{\iota}^{\mathrm{T}}\boldsymbol{y} \tag{50}$$

This is the MM estimator for the model with $\beta_2 = 0$. The sample mean is just a special case of the formula (44)

$$\mathrm{SSR}\left(\boldsymbol{\beta}\right) = \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) \tag{51}$$

The first order condition for minimizing the SSR (51) can be written as (43)