Econometricks: Short Guides to Econometrics*

Davud Rostam-Afschar†

December 1, 2024

Abstract

Short guides to econometrics illustrate statistical methods and demonstrate how they work in theory and practice. With many examples.

Keywords: Econometrics, Ordinary Least Squares, Maximum Likelihood, Generalized Method of Moments, Probability Theory, Distribution Theory, Frisch-Waugh-Lovell, Monte Carlo Simulation
JEL classification: A20, A23, C01, C10, C12, C13

*These guides were developed based on lectures delivered by Davud Rostam-Afschar at the University of Mannheim. I am grateful for the valuable input provided by numerous cohorts of PhD students at the Graduate School of Economic and Social Sciences, University of Mannheim. I thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for financial support through CRC *TRR 266 Accounting for Transparency* (Davud Rostam-Afschar, Project-ID 403041268). Replication files and updates are available here.

†University of Mannheim, 68131 Mannheim, Germany; GLO; IZA; NeST (e-mail: rostam-afschar@uni-mannheim.de),

Contents

1 Review of Probability Theory

1.1 Introduction

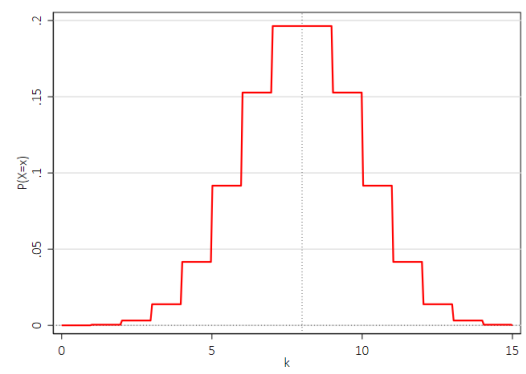
This guide takes a look under the hood of widely used methods in econometrics and beyond. It focuses on Ordinary Least Squares, Maximum Likelihood, Generalized Method of Moments. It shows when and why these methods work with simple examples. This guide also provides an overview of the most important fundamentals of Probability Theory and Distribution Theory on which these methods are based and how to analyze them with the Frisch-Waugh-Lovell decomposition and with Monte Carlo Simulation.

1.2 Probability fundamentals

Discrete and continuous random variables

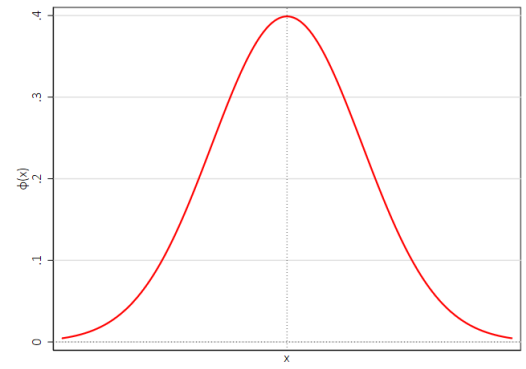
Discrete Random Variable

A random variable X is **discrete** if the set of outcomes x is either finite or countably infinite.



Continuous Random Variable

The random variable X is **continuous** if the set of outcomes x is infinitely divisible and, hence, not countable.



Discrete probabilities

For values x of a discrete random variable X, the **probability mass function** (pmf)

$$f(x) = Prob(X = x).$$

The axioms of probability require

$$0 \leq Prob(X = x) \leq 1,$$

$$\sum_{x} f(x) = 1.$$

Discrete cumulative probabilities

For values x of a discrete random variable X, the **cumulative distribution function**

$$F(x) = \sum_{X \leq x} f(x) = Prob(X \leq x),$$

where

$$f(x_i) = F(x_i) - F(x_{i-1}).$$

Example

Roll of a six-sided die

x	$f(x)$	$F(X \leq x)$
1	$f(1) = 1/6$	$F(X \leq 1) = 1/6$
2	$f(2) = 1/6$	$F(X \leq 2) = 2/6$
3	$f(3) = 1/6$	$F(X \leq 3) = 3/6$
4	$f(4) = 1/6$	$F(X \leq 4) = 4/6$
5	$f(5) = 1/6$	$F(X \leq 5) = 5/6$
6	$f(6) = 1/6$	$F(X \leq 6) = 6/6$

What's the probability that you roll a 5 or higher?
$F(X \geq 5) = 1 - F(X \leq 4) = 1 - 2/3 = 1/3.$

Continuous probabilities

For values x of a continuous random variable X, the probability is zero but the area under $f(x) \geq 0$ in the range form a to b is the **probability density function** (pdf)

$$Prob(a \leq x \leq b) = Prob(a < x < b) = \int_a^b f(x)dx \geq 0.$$

The axioms of probability require

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

$f(x) = 0$ outside the range of x.
The **cumulative distribution function** (cdf) is

$$F(x) = \int_{-\infty}^{x} f(t)dt,$$

$$f(x) = \frac{dF(x)}{dx}.$$

Cumulative distribution function

For continuous and discrete variables, $F(x)$ satisfies

Properties of cdf

- $0 \leq F(x) \leq 1$
- If $x > y$, then $F(x) \geq F(y)$
- $F(+\infty) = 1$
- $F(-\infty) = 0$

and

$$Prob(a < x \leq b) = F(b) - F(a).$$

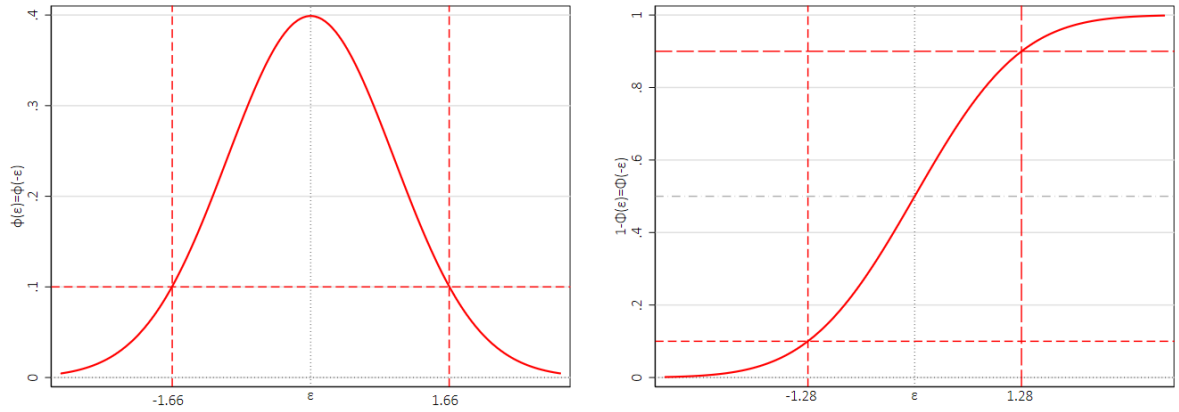
Symmetric distributions

For symmetric distributions

$$f(\mu - x) = f(\mu + x)$$

and

$$1 - F(x) = F(-x).$$



1.3 Mean and variance

Mean of a random variable (Discrete)

The **mean**, or **expected value**, of a discrete random variable is

$$\mu = E[x] = \sum_x x f(x)$$

Example

Roll of a six-sided die

x	$f(x) = 1/n$	$F(X \leq x) = (x - a + 1)/n$
a = 1	$f(1) = 1/6$	$F(X \leq 1) = 1/6$
2	$f(2) = 1/6$	$F(X \leq 2) = 2/6$
3	$f(3) = 1/6$	$F(X \leq 3) = 3/6$
4	$f(4) = 1/6$	$F(X \leq 4) = 4/6$
5	$f(5) = 1/6$	$F(X \leq 5) = 5/6$
b = 6	$f(6) = 1/6$	$F(X \leq 6) = 6/6$

What's the expected value from rolling the dice?
$E[x] = 1/6 + 2/6 + 3/6 + 4/6 + 5/6 + 6/6 = 3.5$.
This is the mean (and the median) of a uniform distribution $(n+1)/2 = (a+b)/2 = 3.5$.

Mean of a random variable (Continuous)

For a continuous random variable x, the expected value is

$$E[x] = \int_x xf(x)dx.$$

Example

The continuous uniform distribution is $1/(b-a)$ for $a \leq x \leq b$ and 0 otherwise.

$$E[x] = \int_a^b \frac{x}{b-a}dx = \frac{1}{b-a}\int_a^b xdx.$$

Antiderivative of x is $x^2/2$

$$E[x] = \frac{1}{b-a}(b^2/2 - a^2/2) = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}.$$

The mean (and the median) is again $(a+b)/2 = 3.5$.

For a function $g(x)$ of x, the expected value is $E[g(x)] = \sum_x g(x)Prob(X = x)$ or $E[g(x)] = \int_x g(x)f(x)dx$. If $g(x) = a+bx$ for constants a and b, then $E[a+bx] = a+bE[x]$.

Variance of a random variable

The **variance** of a random variable $\sigma^2 > 0$ is

$$\sigma^2 = Var[x] = E[(x-\mu)^2] = \begin{cases} \sum_x (x-\mu)^2 f(x) & if\ x\ is\ discrete, \\ \\ \int_x (x-\mu)^2 f(x)dx & if\ x\ is\ continuous. \end{cases}$$

Example

Roll of a six-sided die. What's the variance $V[x]$ from rolling the dice?
The probability of observing x, $Pr(X = x) = 1/n$, is discretely uniformly distributed

$$E[x] = \frac{n+1}{2};\ (E[x])^2 = \frac{(n+1)^2}{4}.$$

$$E[x^2] = \sum_x Pr(X = x) = \frac{1}{n}\sum_{x=1}^{n} x^2 = \frac{(n+1)(2n+1)}{6} \text{ due to the sequence sum of squares.}$$

$$V[x] = E[x^2] - (E[x])^2.$$

$$V[x] = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12} = (6^2-1)/12 \approx 2.92.$$

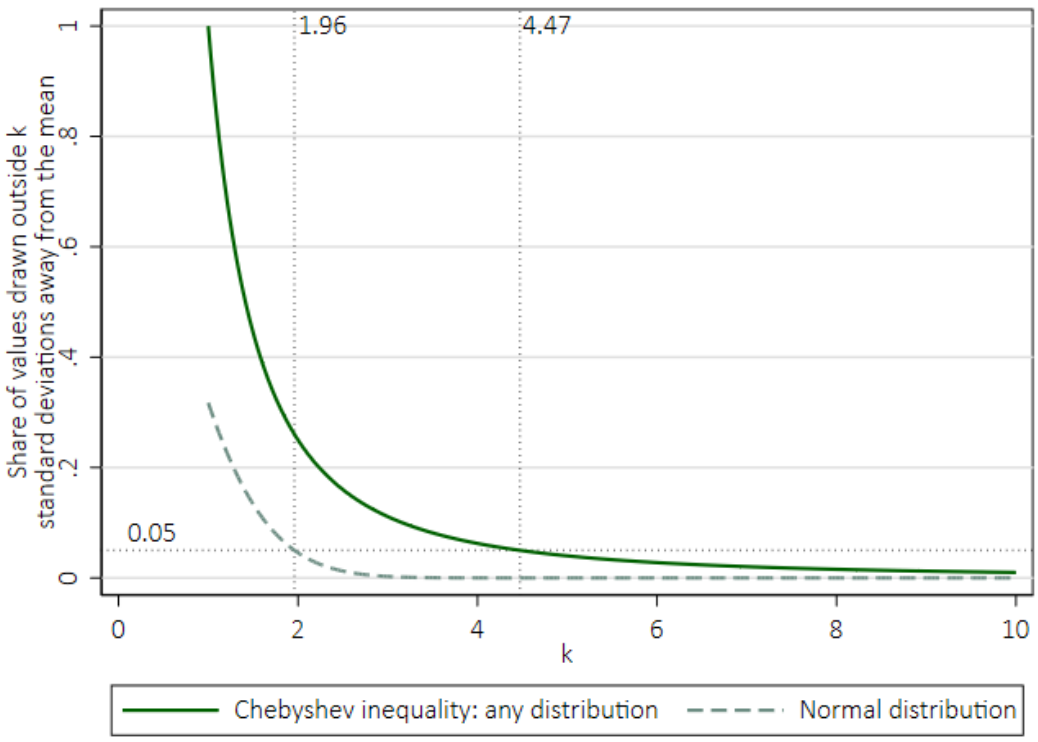
Chebychev inequality

For any random variable x and any positive constant $k > 1$,

$$\Pr(\mu - k\sigma < x < \mu + k\sigma) \geq 1 - \frac{1}{k^2}.$$

Share outside k standard deviations.

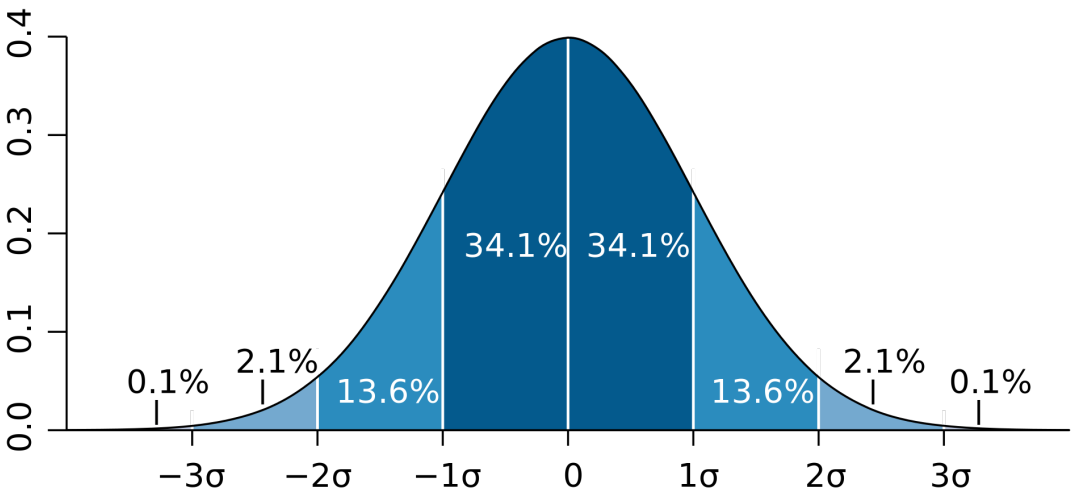
If x is normally distributed, the bound is $1 - (2\Phi(k) - 1)$.



95% of the observations are within 1.96 standard deviations for normally distributed

x. If x is not normal, 95% are at most within 4.47 standard deviations.

Normal coverage



1.4 Moments of a random variable

Central moments of a random variable

The central moments are

$$\mu_r = E[(x - \mu)^r].$$

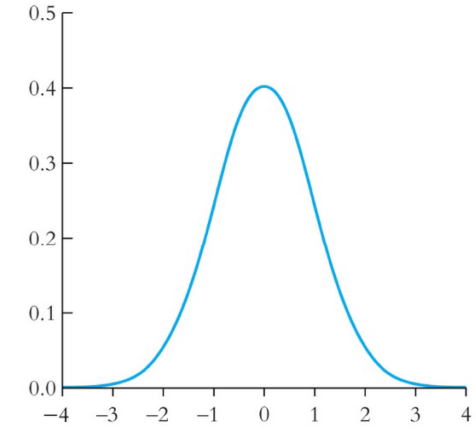
Example

Moments: Two measures often used to describe a probability distribution are

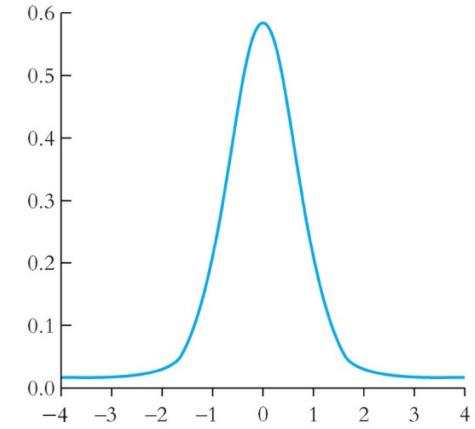
- expectation $= E[(x - \mu)^1]$
- variance $= E[(x - \mu)^2]$
- skewness $= E[(x - \mu)^3]$
- kurtosis $= E[(x - \mu)^4]$

The skewness is zero for symmetric distributions.

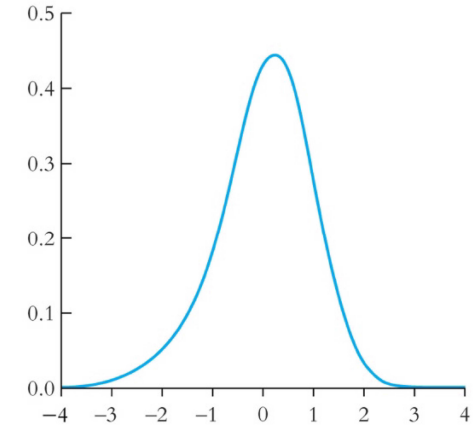
Higher order moments



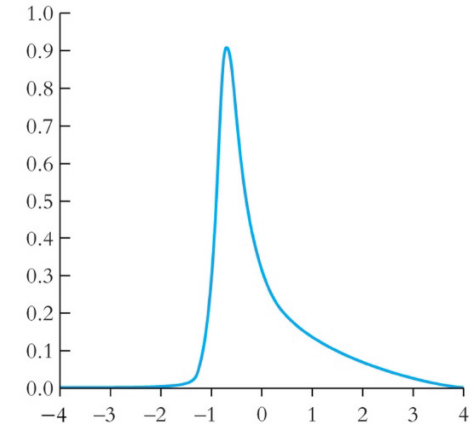
(a) Skewness = 0, kurtosis = 3



(b) Skewness = 0, kurtosis = 20



(c) Skewness = −0.1, kurtosis = 5



(d) Skewness = 0.6, kurtosis = 5

Moment generating function

For the random variable X, with probability density function $f(x)$, if the function

$$M(t) = E[e^{tx}].$$

exists, then it is the **moment generating function**(MGF).

- Often simpler alternative to working directly with probability density functions or cumulative distribution functions
- Not all random variables have moment-generating functions

The nth moment is the nth derivative of the moment-generating function, evaluated at

$t = 0$.

Example

The MGF for the standard normal distribution with $\mu = 0, \sigma = 1$ is

$$M_z(t) = e^{\mu t + \sigma^2 t^2/2} = e^{t^2/2}.$$

If x and y are independent, then the MGF of $x + y$ is $M_x(t)M_y(t)$.

For $x \sim N(\mu, \sigma^2)$ for some $\mu, \sigma > 0$ with moment generating function $M_x(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$, the first moment generating function of x is

$$E[(x-\mu)^1] = M_x{}'(t) = (\mu + \sigma^2 t) \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right).$$

Example

$$\begin{aligned} E[(x-\mu)^1] = M_x{}'(t) &= \frac{d\left[\exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)\right]}{dt} \\ &= \frac{d\left[\mu t + \frac{1}{2}\sigma^2 t^2\right]}{dt} \frac{d\left[\exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)\right]}{d(\mu t + \frac{1}{2}\sigma^2 t^2)} \\ &= (\mu + \sigma^2 t) \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right). \end{aligned}$$

If $x \sim N(0, 1)$,

- the skewness is $E[(x-\mu)^3] = 0$ and
- the kurtosis is $E[(x-\mu)^4] = 3$.

Example

$$E[(x-\mu)^1] = M_x{}'(t) = (\mu+\sigma^2 t)\exp\left(\mu t+\frac{1}{2}\sigma^2 t^2\right) \text{ with } \mu=0,\sigma=1,t=0 : E[x]=\mu=0$$

$$E[(x-\mu)^2] = M_x{}''(t) = \left(\sigma^2+(\mu+\sigma^2 t)^2\right)\exp\left(\mu t+\frac{1}{2}\sigma^2 t^2\right)$$

$$\text{with } \mu=0,\sigma=1,t=0 : E[(x-\mu)^2]=\sigma^2=1$$

$$E[(x-\mu)^3] = M_x{}'''(t) = \left(3\sigma^2(\mu+\sigma^2 t)+(\mu+\sigma^2 t)^3\right)\exp\left(\mu t+\frac{1}{2}\sigma^2 t^2\right)$$

$$\text{with } \mu=0,\sigma=1,t=0 : E[(x-\mu)^3]=0$$

$$E[(x-\mu)^4] = M_x{}^{(4)}(t) = \left(3\sigma^4+6\sigma^2(\mu+\sigma^2 t)^2+(\mu+\sigma^2 t)^4\right)\exp\left(\mu t+\frac{1}{2}\sigma^2 t^2\right)$$

$$\text{with } \mu=0,\sigma=1,t=0 : E[(x-\mu)^4]=3.$$

Approximating mean and variance

For any two functions $g_1(x)$ and $g_2(x)$,

$$E[g_1(x)+g_2(x)] = E[g_1(x)]+E[g_2(x)].$$

For the general case of a possibly nonlinear $g(x)$,

$$E[g(x)] = \int_x g(x)f(x)dx,$$

and

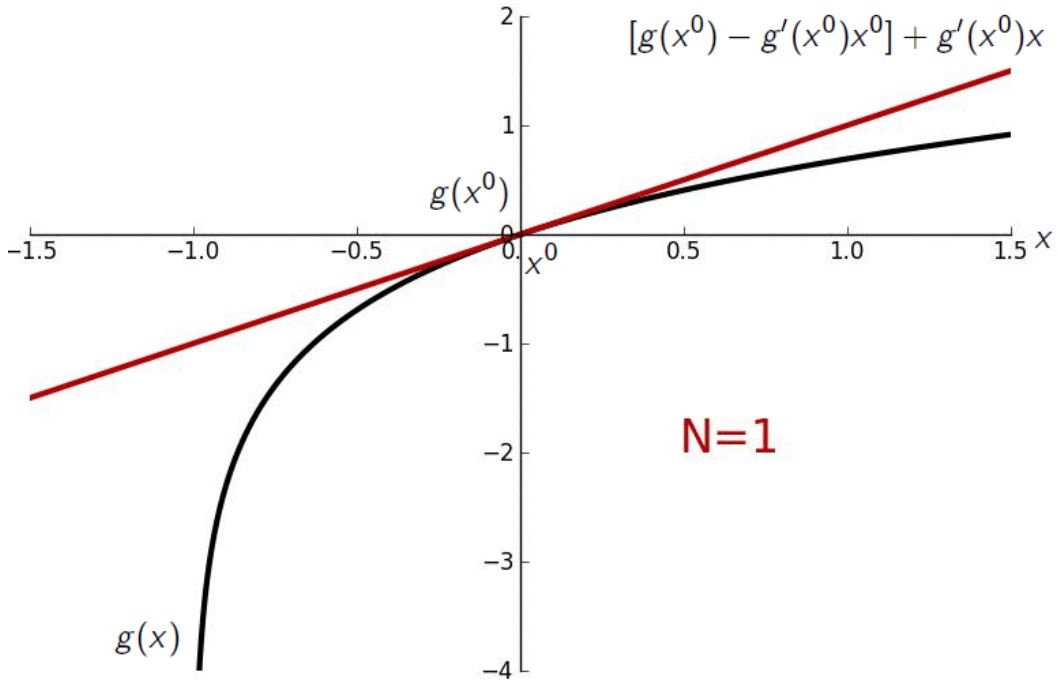
$$Var[g(x)] = \int_x \left(g(x)-E[g(x)]\right)^2 f(x)dx.$$

$E[g(x)]$ and $Var[g(x)]$ can be approximated by a first order linear Taylor series:

First order linear Taylor series

$$g(x) \approx [g(x^0)-g'(x^0)x^0]+g'(x^0)x. \tag{1}$$

Taylor approximation Order 1



A natural choice for the expansion point is $x^0 = \mu = E(x)$. Inserting this value in Eq. (1) gives

$$g(x) \approx [g(\mu) - g'(\mu)\mu] + g'(\mu)x,$$

so that

$$E[g(x)] \approx g(\mu),$$

and

$$Var[g(x)] \approx [g'(\mu)]^2 Var[x].$$

Example

Isoelastic utility. $c_{bad} = 10.00$ Euro; $c_{good} = 100.00$ Euro; probability good outcome 50%

$\mu = E[c] = 1/2 \times c_{bad} + 1/2 \times c_{good}$ =55.00 Euro

$$u(c) = c^{1/2}$$

$u(\mu) = 7.42$ approximates $E[u(c)] = 1/2 \times 10^{1/2} + 1/2 \times 100^{1/2} = 6.58$

Example

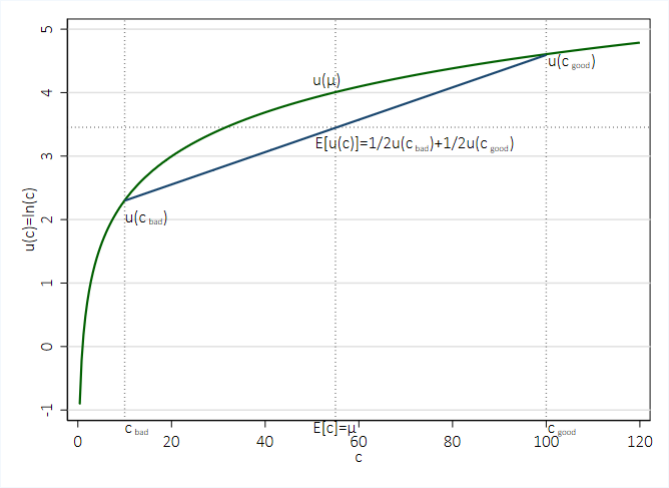
Isoelastic utility.

$c_{bad} = 10.00$ Euro; $c_{good} = 100.00$ Euro; probability good outcome 50%; $\mu = 55.00$ Euro

$$u(c) = \ln(c)$$

$$u(\mu) = 4.01 \text{ approx.}$$

$$E[u(c)] = 1/2 \times \ln(10) + 1/2 \times \ln(100) = 3.45$$



Jensen's inequality:
$E[g(x)] \leq g(E[x])]$ if $g''(x) < 0$.

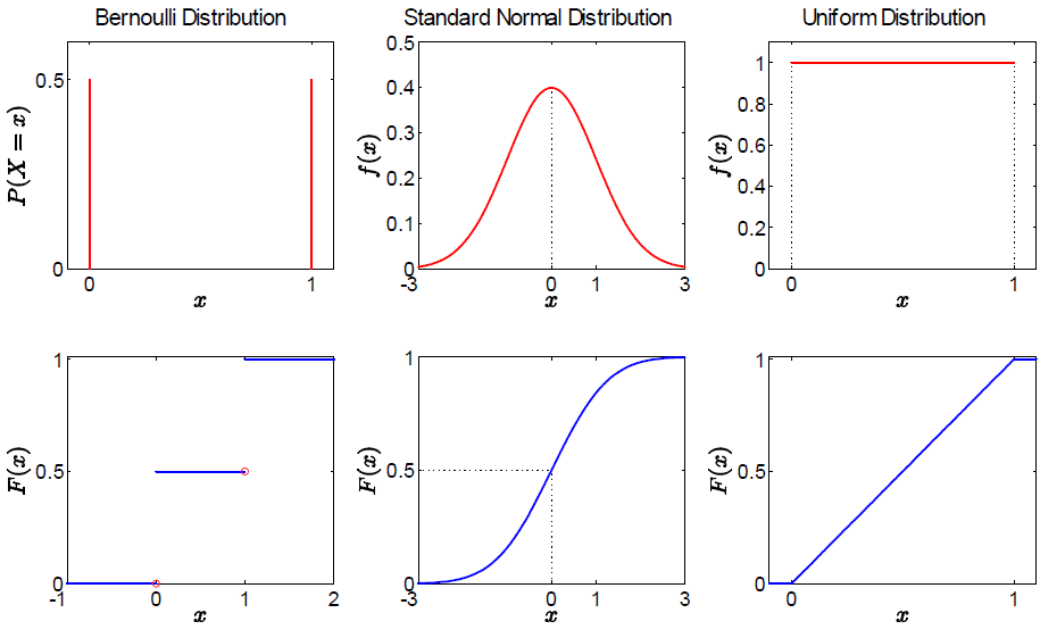
$V[u(c)] \approx (1/55)^2((10-55)^2 + (100-55)^2) = 1.34$
$V[u(c)] = (\ln(10) - E[u(c)])^2 + (\ln(100) - E[u(c)])^2 = 2.65$

1.5 Useful rules

- $Var[x] = E[x^2] - \mu^2$
- $E[x^2] = \sigma^2 + \mu^2$
- If a and b constants, $Var[a + bx] = b^2 Var[x]$
- $Var[a] = 0$
- If $g(x) = a + bx$ and a and b are constants, $E[a + bx] = a + bE[x]$

- Coverage $\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$
- Skewness $= E[(x - \mu)^3]$
- Kurtosis $= E[(x - \mu)^4]$
- For symmetric distributions $f(\mu - x) = f(\mu + x)$; $1 - F(x) = F(-x)$
- $E[g(x)] \approx g(\mu)$
- $Var[g(x)] \approx [g'(\mu)]^2 Var[x]$

2 Specific Distributions



Discrete distributions

Bernoulli distribution

The **Bernoulli distribution** for a single binomial outcome (trial) is

$$\begin{aligned} Prob(x=1) &= p, \\ Prob(x=0) &= 1-p, \end{aligned}$$

where $0 \leq p \leq 1$ is the probability of success.

- $E[x] = p$ and
- $V[x] = E[x^2] - E[x]^2 = p - p^2 = p(1-p)$.

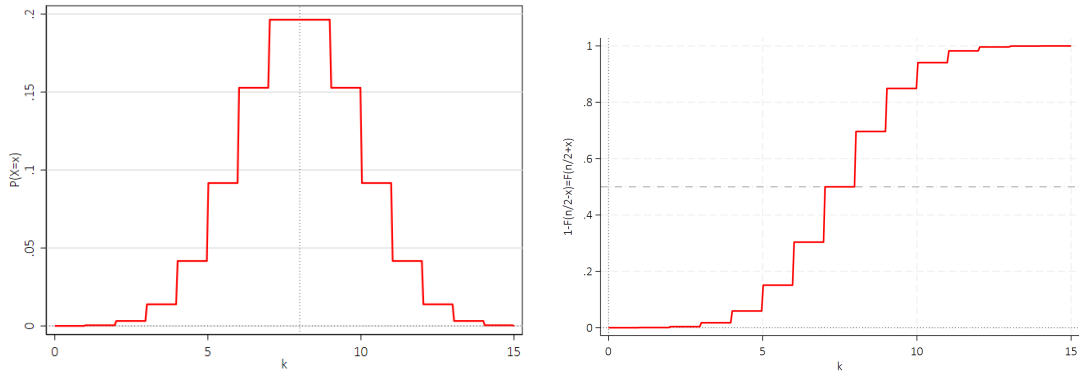
The distribution for x successes in n trials is the **binomial distribution**,

$$Prob(X=x) = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x} \quad x = 0, 1, \ldots, n.$$

The mean and variance of x are

- $E[x] = np$ and
- $V[x] = np(1-p)$.

Example of a binomial $[n = 15, p = 0.5]$ distribution:



Poisson distribution

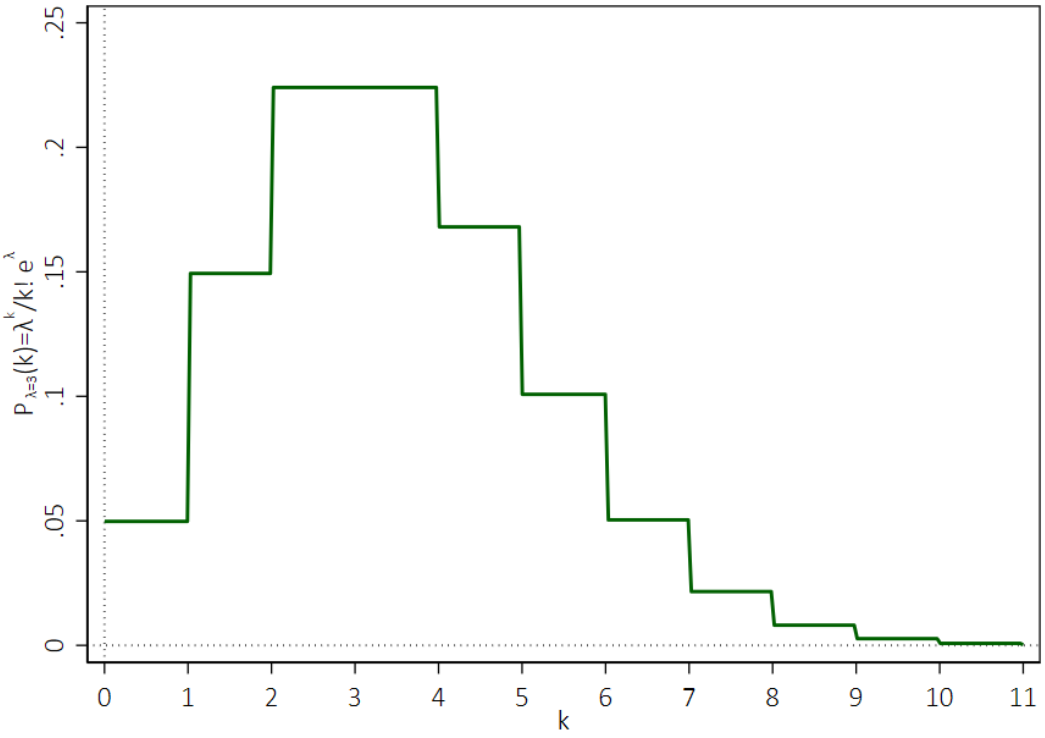
The limiting form of the binomial distribution, $n \to \infty$, is the **Poisson distribution**,

$$Prob(X = x) = \frac{e^{\lambda}\lambda^{x}}{x!}.$$

The mean and variance of x are

- $E[x] = \lambda$ and
- $V[x] = \lambda$.

Example of a Poisson [3] distribution:



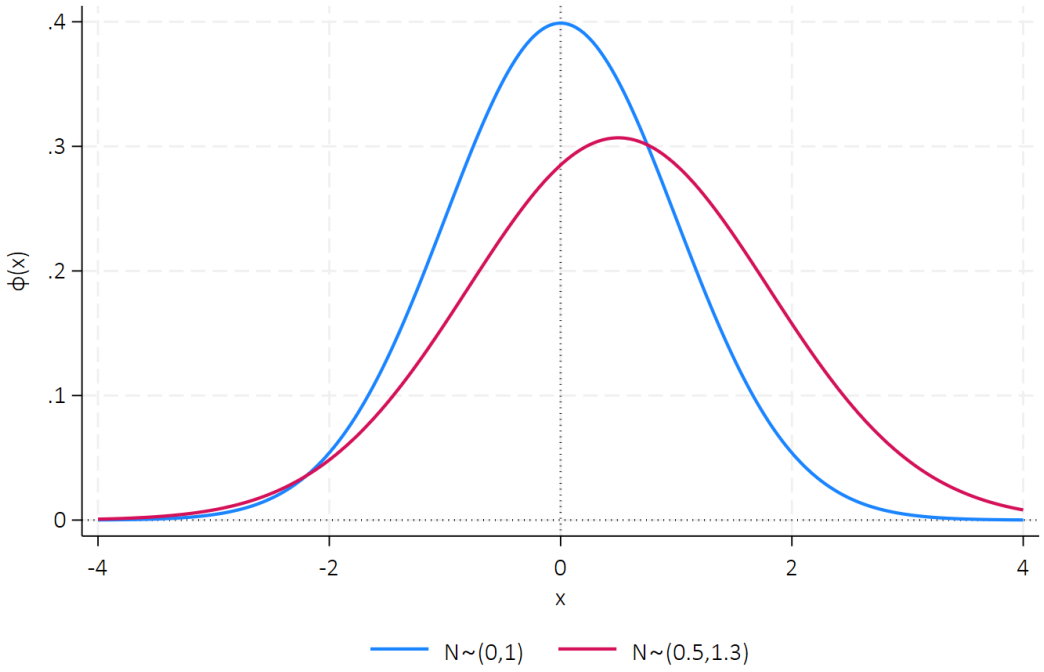
2.1 Normal distribution

The normal distribution

Random variable $x \sim N[\mu, \sigma^2]$ is distributed according to the **normal distribution** with mean μ and standard deviation σ obtained as

$$f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}.$$

The density is denoted $\phi(x)$ and the cumulative distribution function is denoted $\Phi(x)$ for the standard normal. Example of a standard normal, $(x \sim N[0, 1])$, and a normal with mean 0.5 and standard deviation 1.3:



2.2 Method of transformations

Transformation of random variables

Continuous variable x may be transformed to a discrete variable y. Calculate the mean of variable x in the respective interval:

$$\begin{aligned} Prob(Y = \mu_1) &= P(-\infty < X \leq a), \\ Prob(Y = \mu_2) &= P(a < X \leq b), \\ Prob(Y = \mu_3) &= P(b < X \leq \infty). \end{aligned}$$

Method of transformations

If x is a continuous random variable with pdf $f_x(x)$ and if $y = g(x)$ is a continuous monotonic function of x, then the density of y is obtained by

$$Prob(y \leq b) = \int_{-\infty}^{b} f_x(g^{-1}(y))|g^{-1\prime}(y)|dy.$$

With $f_y(y) = f_x(g^{-1}(y))|g^{-1\prime}](y)|dy$, this equation can be written as

$$Prob(y \leq b) = \int_{-\infty}^{b} f_y(y)dy.$$

Example

If $x \sim N[\mu, \sigma^2]$, then the distribution of $y = g(x) = \frac{x-\mu}{\sigma}$ is found as follows:

$$g^{-1}(y) = x = \sigma y + \mu$$

$$g^{-1\prime}(y) = \frac{dx}{dy} = \sigma$$

Therefore with $f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}[(g^{-1}(y)-\mu)^2/\sigma^2]}|g^{-1\prime}(y)|$

$$f_y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-[(\sigma y+\mu)-\mu]^2/2\sigma^2}|\sigma| = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}.$$

Properties of the normal distribution

- Preservation under linear transformation:
  If $x \sim N[\mu, \sigma^2]$, then $(a + bx) \sim N[a + b\mu, b^2\sigma^2]$.
- Convenient transformation $a = -\mu/\sigma$ and $b = 1/\sigma$:
  The resulting variable $z = \frac{(x-\mu)}{\sigma}$ has the standard normal distribution with density

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

- If $x \sim N[\mu, \sigma^2]$, then $f(x) = \frac{1}{\sigma}\phi[\frac{x-\mu}{\sigma}]$
- $Prob(a \leq x \leq b) = Prob\left(\frac{a-\mu}{\sigma} \leq \frac{x-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right)$
- $\phi(-z) = 1 - \phi(z)$ and $\Phi(-x) = 1 - \Phi(x)$ because of symmetry

If $z \sim N[0, 1]$, then $z^2 \sim \chi^2[1]$ with pdf $\frac{1}{\sqrt{2\pi y}} e^{-y/2}$.

Example

$$f_x(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$y = g(x) = x^2$$

$g^{-1}(y) = x = \pm\sqrt{y}$ there are two solutions to g_1, g_2.

$$g^{-1\prime}(y) = \frac{dx}{dy} = \pm 1/2y^{-1/2}$$

$$f_y(y) = f_x(g_1^{-1}(y))|g_1^{-1\prime}(y)| + f_x(g_2^{-1}(y))|g_2^{-1\prime}(y)|$$

$$f_y(y) = f_x(\sqrt{y})|1/2y^{-1/2}| + f_x(-\sqrt{y})| - 1/2y^{-1/2}|$$

$$f_y(y) = \frac{1}{2\sqrt{2\pi y}} e^{-\frac{y}{2}} + \frac{1}{2\sqrt{2\pi y}} e^{-\frac{y}{2}} = \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}}$$

Distributions derived from the normal

- If $z \sim N[0, 1]$, then $z^2 \sim \chi^2[1]$ with $E[z^2] = 1$ and $V[z^2] = 2$.
- If $x_1, ..., x_n$ are n independent $\chi^2[1]$ variables, then

$$\sum_{i=1}^{n} x_i \sim \chi^2[n].$$

	Normal
Parameters	$\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_{>0}$
Support	$x \in \mathbb{R}$
PDF	$\phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
CDF	$\Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$
Mean	μ
Median	μ
Mode	μ
Variance	σ^2
Skewness	0
Ex. Kurtosis	0
MGF	$\exp(\mu t + \sigma^2 t^2/2)$

- PDF denotes probability density function, CDF cumulative distribution function, MGF moment-generating function.
- μ mean (location), σ, s (scale).
- Excess Kurtosis is defined as Kurtosis minus 3.
- The Gauss error function is $\operatorname{erf} z = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, \mathrm{d}t$.

- If $z_i, i = 1, ..., n$, are independent $N[0, 1]$ variables, then

$$\sum_{i=1}^{n} z_i^2 \sim \chi^2[n].$$

- If $z_i, i = 1, ..., n$, are independent $N[0, \sigma^2]$ variables, then

$$\sum_{i=1}^{n} \left(\frac{z_i}{\sigma}\right)^2 \sim \chi^2[n].$$

- If x_1 and x_2 are independent χ^2 variables with n_1 and n_2 degrees of freedom, then

$$x_1 + x_2 \sim \chi^2[n_1 + n_2].$$

2.3 The χ^2 distribution

The χ^2 distribution

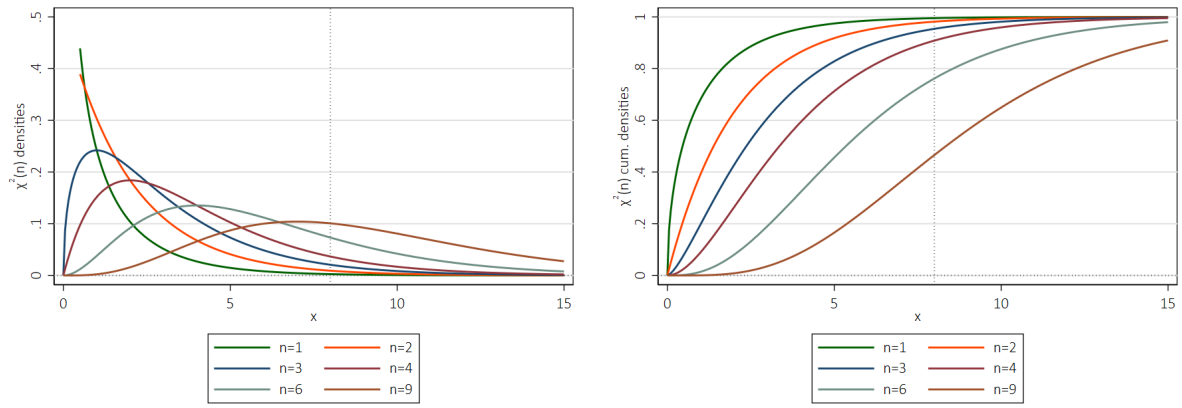
Random variable $x \sim \chi^2[n]$ is distributed according to the **chi-squared distribution** with n degrees of freedom

$$f(x|n) = \frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma\left(\frac{n}{2}\right)},$$

where Γ is the Gamma-distribution (more below).

- $E[x] = n$
- $V[x] = 2n$

Example of a $\chi^2[3]$ distribution:



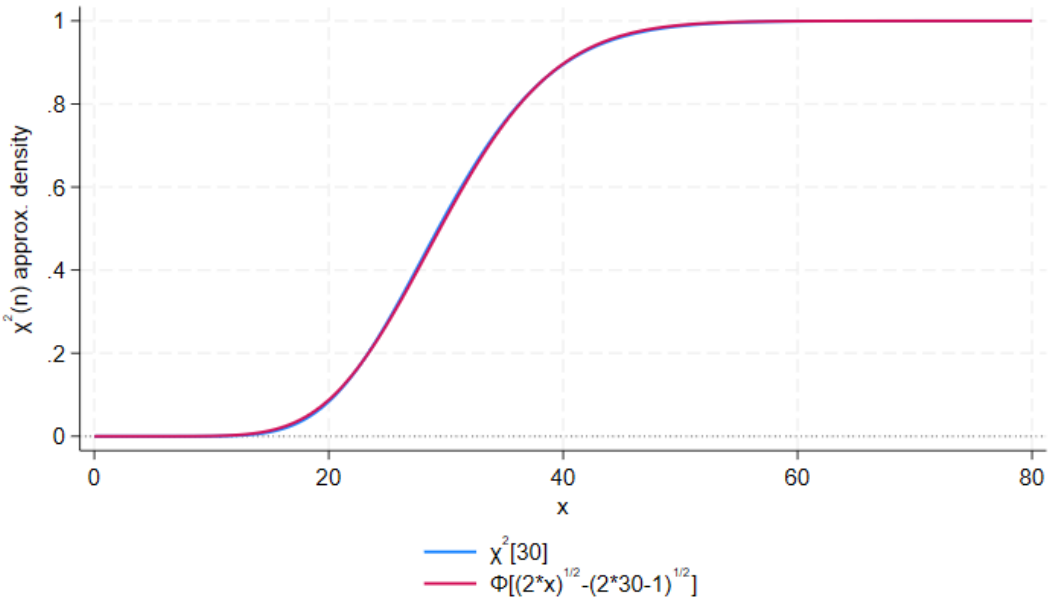
Approximating a χ^2

For degrees of freedom greater than 30 the distribution of the chi-squared variable x is approx.

$$z = (2x)^{1/2} - (2n-1)^{1/2},$$

which is approximately standard normally distributed. Thus,

$$Prob(\chi^2[n] \leq a) \approx \Phi[(2a)^{1/2} - (2n-1)^{1/2}].$$



	χ^2
Parameters	$n \in \mathbb{N}_{>0}$
Support	$x \in \mathbb{R}_{>0}$ if $n = 1$, else $x \in \mathbb{R}_{\geq 0}$
PDF	$\frac{1}{2^{n/2}\Gamma(n/2)}\, x^{n/2-1} e^{-x/2}$
CDF	$\frac{1}{\Gamma(n/2)}\, \gamma\left(\frac{n}{2},\, \frac{x}{2}\right)$
Mean	n
Median	No simple closed form
Mode	$\max(n-2, 0)$
Variance	$2n$
Skewness	$\sqrt{8/n}$
Ex. Kurtosis	$\frac{12}{n}$
MGF	$(1-2t)^{-n/2}$ for $t < \frac{1}{2}$

- n, n_1, n_2 known as degrees of freedom.
- Regularized incomplete beta function $I(x, a, b) = \frac{B(x,a,b)}{B(a,b)}$ with $B(x, a, b) = \int_0^x t^{a-1}\,(1-t)^{b-1}\,dt$.

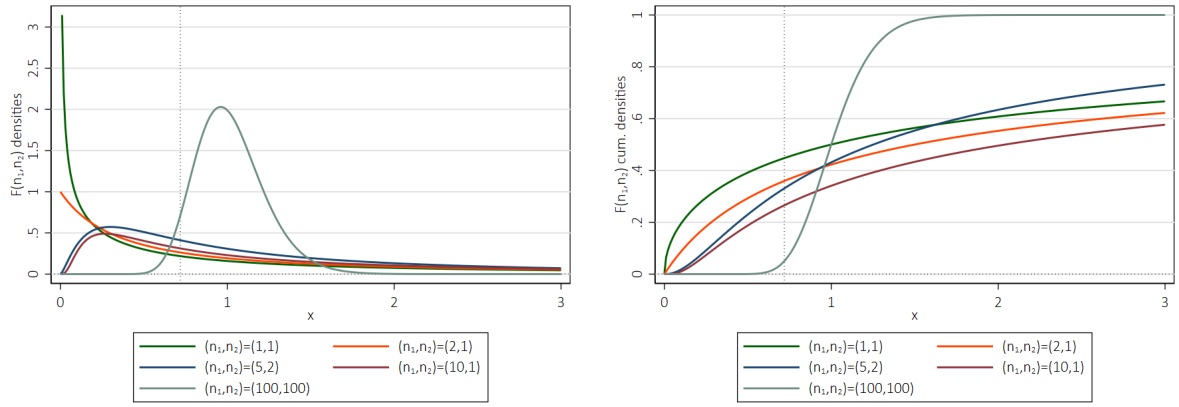
2.4 The F-distribution

The F-distribution

If x_1 and x_2 are two independent chi-squared variables with degrees of freedom parameters n_1 and n_2, respectively, then the ratio

$$F[n_1, n_2] = \frac{x_1/n_1}{x_2/n_2}$$

has the **F distribution** with n_1 and n_2 degrees of freedom.



	F
Parameters	$n_1,\, n_2 \in \mathbb{N}_{>0}$
Support	$x \in \mathbb{R}_{>0}$ if $n_1 = 1$, else $x \in \mathbb{R}_{\geq 0}$
PDF	$n_1^{\frac{n_1}{2}}\, n_2^{\frac{n_2}{2}}\, \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})}\, \frac{x^{\frac{n_1}{2}-1}}{(n_1 x+n_2)^{\frac{n_1+n_2}{2}}}$
CDF	$I\left(\frac{n_1 x}{n_1 x+n_2},\, \frac{n_1}{2},\, \frac{n_2}{2}\right)$
Mean	$\frac{n_2}{n_2-2}$ for $n_2 > 2$
Median	No simple closed form
Mode	$\frac{n_1-2}{n_1}\, \frac{n_2}{n_2+2}$ for $n_1 > 2$
Variance	$\frac{2\, n_2^2\, (n_1+n_2-2)}{n_1 (n_2-2)^2 (n_2-4)}$ for $n_2 > 4$
Skewness	$\frac{(2n_1+n_2-2)\sqrt{8(n_2-4)}}{(n_2-6)\sqrt{n_1(n_1+n_2-2)}}$ for $n_2 > 6$
Ex. Kurtosis	$12 \frac{n_1(5n_2-22)(n_1+n_2-2)+(n_2-4)(n_2-2)^2}{n_1(n_2-6)(n_2-8)(n_1+n_2-2)}$ for $n_2 > 8$
MGF	does not exist

- n, n_1, n_2 known as degrees of freedom.
- Regularized incomplete beta function $I(x, a, b) = \frac{B(x,a,b)}{B(a,b)}$ with $B(x,\, a, b) = \int_0^x t^{a-1}\,(1-t)^{b-1}\, dt$.

2.5 The student t-distribution

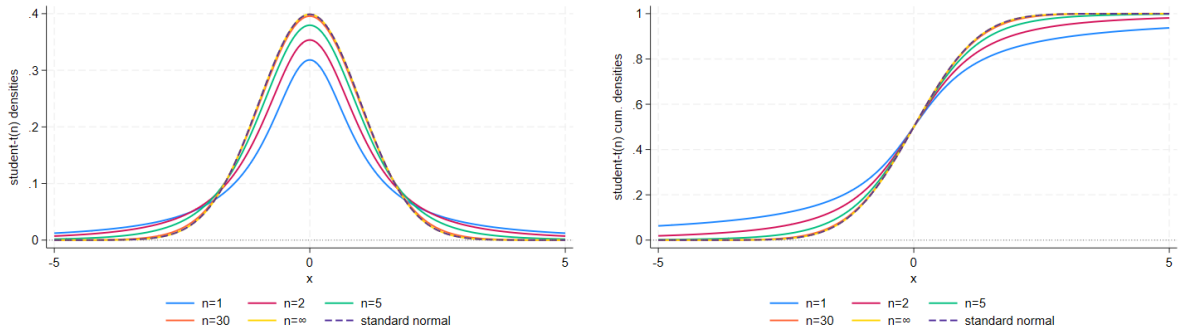
The student t-distribution

If x_1 is an $N[0,1]$ variable, often denoted by z, and x_2 is $\chi^2[n_2]$ and is independent of x_1, then the ratio

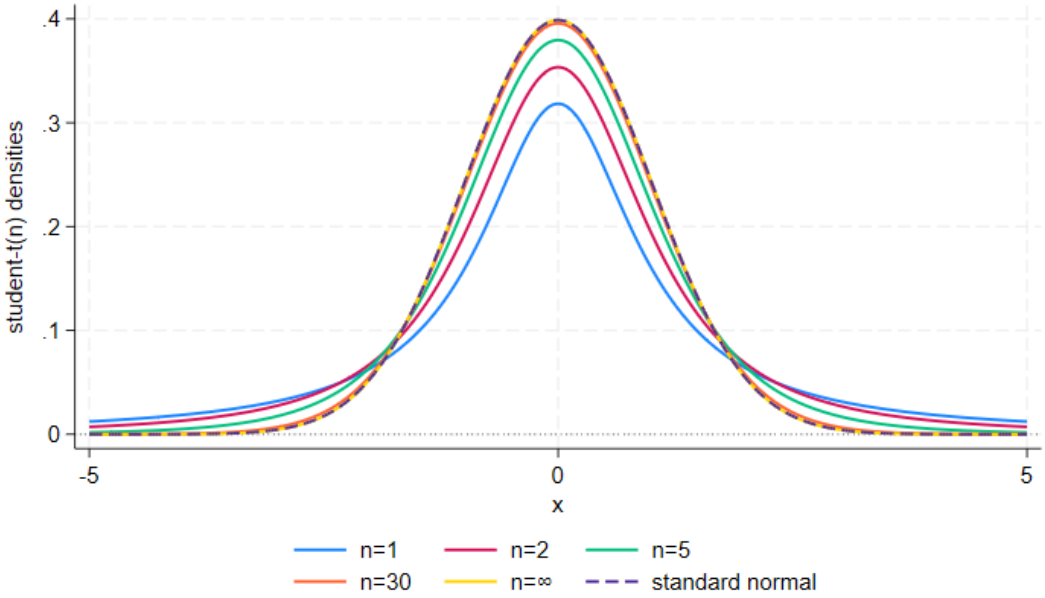
$$t[n_2] = \frac{x_1}{\sqrt{x_2/n_2}}.$$

has the **t distribution** with n_2 degrees of freedom.

Example for the t distributions with 3 and 10 degrees of freedom with the standard normal distribution.



Comparing (2.4) with $n_1 = 1$ and (2.5), if $t \sim t[n]$, then $t^2 \sim F[1,n]$.
The $t[30]$ approx. the standard normal



	t
Parameters	$n \in \mathbb{R}_{>0}$
Support	$x \in \mathbb{R}$
PDF	$\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\ \Gamma\left(\frac{n}{2}\right)} \left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$
CDF	$\frac{1}{2} + x\ \Gamma\left(\frac{n+1}{2}\right) \times \frac{{}_2F_1\left(\frac{1}{2},\ \frac{n+1}{2};\ \frac{3}{2};\ -\frac{x^2}{n}\right)}{\sqrt{\pi n}\ \Gamma\left(\frac{n}{2}\right)}$
Mean	0 for $n > 1$
Median	0
Mode	0
Variance	$\frac{n}{n-2}$ for $n > 2$, ∞ for $1 < n \leq 2$
Skewness	0 for $n > 3$
Ex. Kurtosis	$\frac{6}{n-4}$ for $n > 4, \infty$ for $2 < n \leq 4$
MGF	does not exist

- n denote degrees of freedom.
- ${}_2F_1(\cdot,\cdot;\cdot;\cdot)$ is a particular instance of the hypergeometric function.

2.6 The lognormal distribution

The lognormal distribution

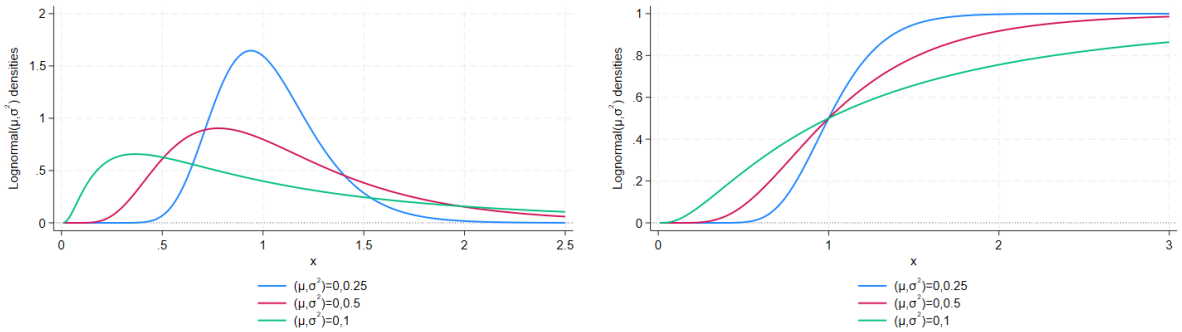
The **lognormal distribution**, denoted $LN[\mu, \sigma^2]$, has been particularly useful in modeling the size distributions.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2}[(\ln x - \mu)/\sigma]^2}, \qquad x > 0$$

A lognormal variable x has

- $E[x] = e^{\mu+\sigma^2/2}$, and
- $Var[x] = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$.

If $y \sim LN[\mu, \sigma^2]$, then $\ln y \sim N[\mu, \sigma^2]$.



	Log-normal
Parameters	$\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_{>0}$
Support	$x \in \mathbb{R}_{>0}$
PDF	$\frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x-\mu\ )^2}{2\sigma^2}\right)$
CDF	$\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{\ln x-\mu}{\sigma\sqrt{2}}\right)\right]$ $= \Phi\left(\frac{\ln(x)-\mu}{\sigma}\right)$
Mean	$\exp\left(\mu+\frac{\sigma^2}{2}\right)$
Median	$\exp(\mu)$
Mode	$\exp\left(\mu-\sigma^2\right)$
Variance	$[\exp(\sigma^2)-1]\exp\left(2\mu+\sigma^2\right)$
Skewness	$[\exp\left(\sigma^2\right)+2]\sqrt{\exp(\sigma^2)-1}$
Ex. Kurtosis	$1\exp\left(4\sigma^2\right)+2\exp\left(3\sigma^2\right)+3\exp\left(2\sigma^2\right)-6$
MGF	not determined by its moments

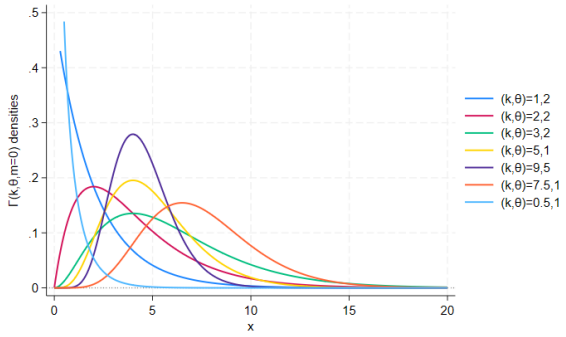
2.7 The gamma distribution

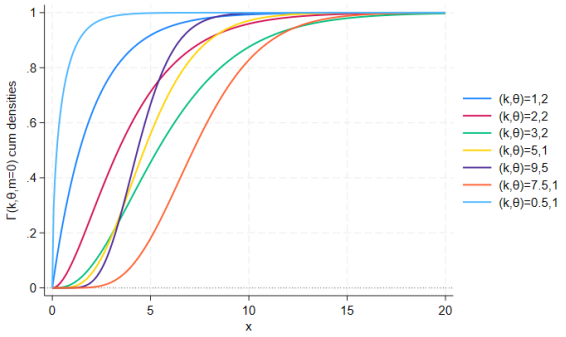
The gamma distribution

The general form of the **gamma distribution** is

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta x} x^{\alpha-1}, \qquad x \geq 0, \beta = 1/\theta > 0, \alpha = k > 0.$$

Many familiar distributions are special cases, including the **exponential distribution**($\alpha = 1$) and **chi-squared**($\beta = 1/2, \alpha = n/2$). The **Erlang distribution** results if α is a positive integer. The mean is α/β, and the variance is α/β^2. The **inverse gamma distribution** is the distribution of $1/x$, where x has the gamma distribution.





	Γ	Γ
Parameters	$k > 0 \in \mathbb{R}$ (shape),	$\alpha > 0 \in \mathbb{R}$ (shape),
	$\theta > 0 \in \mathbb{R}$ scale	$\beta > 0 \in \mathbb{R}$ (rate)
Support	$x \in \mathbb{R}(0, \infty)$	$x \in \mathbb{R}(0, \infty)$
PDF	$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta}$	$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$
CDF	$F(x) = \frac{1}{\Gamma(k)} \gamma\left(k, \frac{x}{\theta}\right)$	$F(x) = \frac{1}{\Gamma(\alpha)} \gamma(\alpha, \beta x)$
Mean	$k\theta$	$\frac{\alpha}{\beta}$
Median	No simple closed form	No simple closed form
Mode	$(k-1)\theta$ for $k \geq 1$, 0 for $k < 1$	$\frac{\alpha-1}{\beta}$ for $\alpha \geq 1$, 0 for $\alpha < 1$
Variance	$k\theta^2$	$\frac{\alpha}{\beta^2}$
Skewness	$\frac{2}{\sqrt{k}}$	$\frac{2}{\sqrt{\alpha}}$
Ex. Kurtosis	$\frac{6}{k}$	$\frac{6}{\alpha}$
MGF	$(1 - \theta t)^{-k}$ for $t < \frac{1}{\theta}$	$\left(1 - \frac{t}{\beta}\right)^{-\alpha}$ for $t < \beta$

- $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, \mathrm{d}t, \qquad \Re(z) > 0$, for complex numbers with a positive real part.
- lower incomplete gamma function is $\gamma(s, x) = \int_0^x t^{s-1}\, e^{-t}\, dt$, for complex numbers with a positive real part.

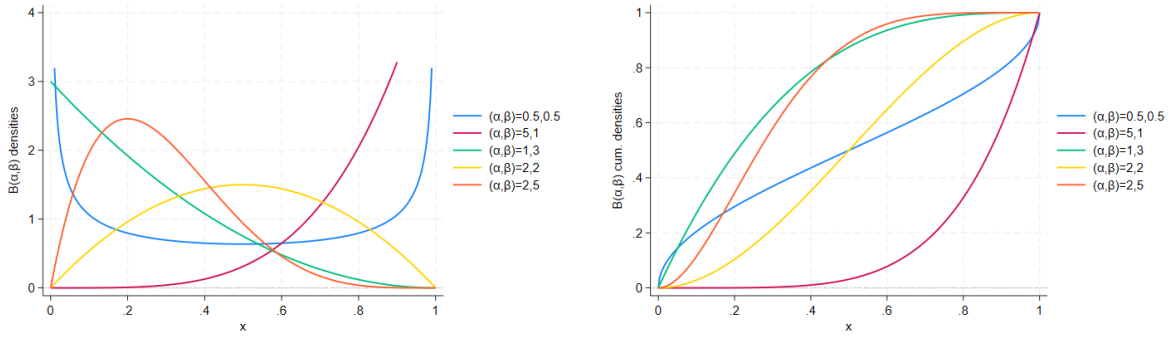
2.8 The beta distribution

The beta distribution

For a variable constrained between 0 and $c > 0$, the **beta distribution** has proved useful. Its density is

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{x}{c}\right)^{\alpha-1} \left(1-\frac{x}{c}\right)^{\beta-1} \frac{1}{c}, \qquad 0 \le x \le 1.$$

It is symmetric if $\alpha = \beta$, asymmetric otherwise. The mean is $ca/(\alpha+\beta)$, and the variance is $c^2\alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2]$.



	B
Parameters	$\alpha, \beta \in \mathbb{R}_{>0}$
Support	$x \in [0,1]$ or $x \in (0,1)$
PDF	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$
CDF	$I(x,\, \alpha, \beta)$
Mean	$\frac{\alpha}{\alpha+\beta}$
Median	$I^{[-1]}_{\frac{1}{2}}(\alpha,\beta) \approx \frac{\alpha-\frac{1}{3}}{\alpha+\beta-\frac{2}{3}}$ for $\alpha, \beta > 1$
Mode	*
Variance	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Skewness	$\frac{2\,(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}$
Ex. Kurtosis	$\frac{6[(\alpha-\beta)^2(\alpha+\beta+1)-\alpha\beta(\alpha+\beta+2)]}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)}$
MGF	$1+\sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r}\right)\frac{t^k}{k!}$

- $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and Γ is the Gamma function.
- $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,\mathrm{d}t, \qquad \Re(z) > 0$, for complex numbers with a positive real part.
- Regularized incomplete beta function $I(x,\, a, b) = \frac{B(x,\, a, b)}{B(a,b)}$ with $B(x,\, a, b) = \int_0^x t^{a-1}\,(1-t)^{b-1}\,dt$.
- $^*\frac{\alpha-1}{\alpha+\beta-2}$ for $\alpha, \beta > 1$; any value in $(0,1)$ for $\alpha, \beta = 1$; $\{0,1\}$ (bimodal) for $\alpha, \beta < 1$; 0 for $\alpha \leq 1, \beta > 1$; 1 for $\alpha > 1, \beta \leq 1$.

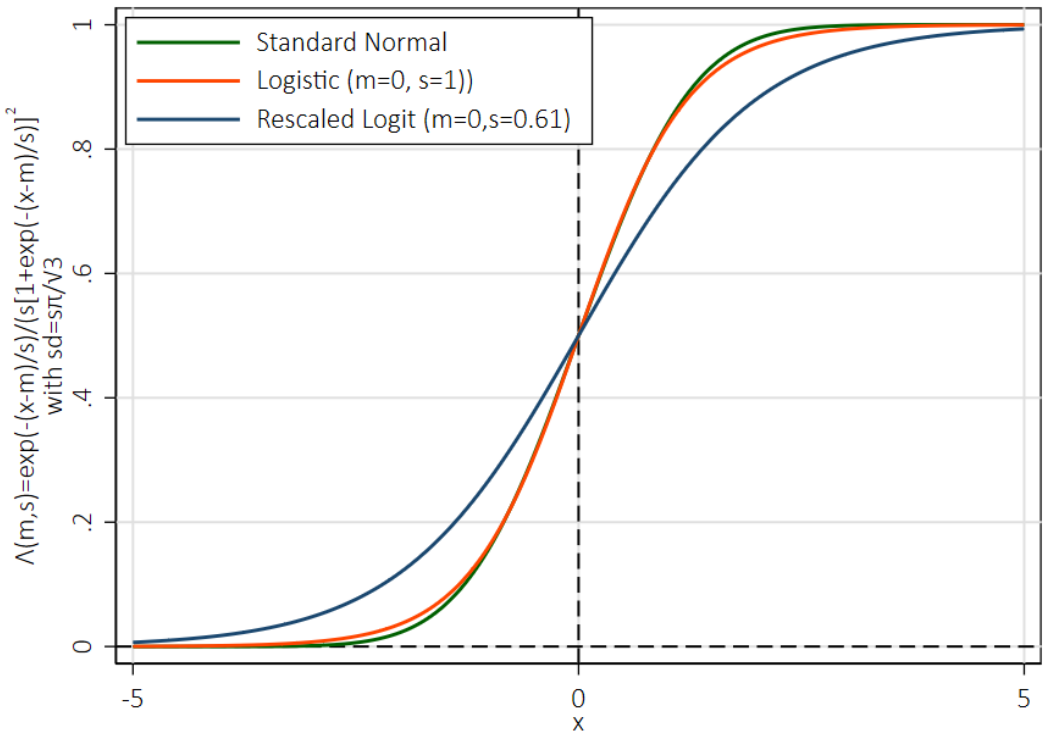
2.9 The logistic distribution

The logistic distribution

The **logistic distribution** is an alternative if the normal cannot model the mass in the tails; the cdf for a logistic random variable with $\mu = 0, s = 1$ is

$$F(x) = \Lambda(x) = \frac{1}{1 + e^{-x}}.$$

The density is $f(x) = \Lambda(x)[1 - \Lambda(x)]$. The mean and variance of this random variable are zero and $\sigma^2 = \pi^2/3$.



	Logistic
Parameters	$\mu \in \mathbb{R}$, $s \in \mathbb{R}_{>0}$
Support	$x \in \mathbb{R}$
PDF	$\lambda\left(\frac{x-\mu}{s}\right) = \frac{e^{-(x-\mu)/s}}{s\left(1+e^{-(x-\mu)/s}\right)^2}$
CDF	$\Lambda\left(\frac{x-\mu}{s}\right) = \frac{1}{1+e^{-(x-\mu)/s}}$
Mean	μ
Median	μ
Mode	μ
Variance	$\frac{s^2\pi^2}{3}$
Skewness	0
Ex. Kurtosis	6/5
MGF	$e^{\mu t}B(1-st, 1+st)$ for $t \in (-1/s, 1/s)$

2.10 The Wishart distribution

The Wishart distribution

The **Wishart distribution** describes the distribution of a random matrix obtained as

$$f(\boldsymbol{W}) = \sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)'.$$

where x_i is the ith of nK element random vectors from the multivariate normal distribution with mean vector, μ, and covariance matrix, Σ. The density of the Wishart random matrix is

$$f(\boldsymbol{W}) = \frac{\exp\left[-\frac{1}{2}trace(\Sigma^{-1}\boldsymbol{W})\right]|\boldsymbol{W}|^{-\frac{1}{2}(n-K-1)}}{2^{nK/2}|\Sigma|^{K/2}\pi^{K(K-1)/4}\prod_{j=1}^{K}\Gamma\left(\frac{n+1-j}{2}\right)}.$$

The mean matrix is $n\Sigma$. For the individual pairs of elements in $\boldsymbol{W}$,

$$Cov[w_{ij}, w_{rs}] = n(\sigma_{ir}\sigma_{js} + \sigma_{is}\sigma_{jr}).$$

The Wishart distribution is a multivariate extension of χ^2 distribution. If $\boldsymbol{W} \sim W(n, \sigma^2)$, then $\boldsymbol{W}/\sigma^2 \sim \chi^2[n]$.

3 Review of Distribution Theory

3.1 Joint and marginal bivariate distributions

Bivariate distributions

For observations of two discrete variables $y \in \{1, 2\}$ and $x \in \{1, 2, 3\}$, we can calculate

- the frequencies $n_{x,y}$,

freq. $n_{x,y}$	$y = 1$	$y = 2$	$f(x) = n_x/N$
$x = 1$	1	2	$3/10$
$x = 2$	1	2	$3/10$
$x = 3$	0	4	$4/10$
$f(y) = n_y/N$	$2/10$	$8/10$	1

- the frequencies $n_{x,y}$,
- conditional distributions $f(y|x)$ and $f(x|y)$,
- joint distributions $f(x, y)$, and
- marginal distributions $f_y(y)$ and $f_x(x)$.

freq. $n_{x,y}$	$y = 1$	$y = 2$	$f(x) = n_x/N$
$x = 1$	1	2	$3/10$
$x = 2$	1	2	$3/10$
$x = 3$	0	4	$4/10$
$f(y) = n_y/N$	$2/10$	$8/10$	1

cond. distr. $f(y\|x)$	$y = 1$	$y = 2$	$\sum_y$
$f(y\|x = 1)$	$1/3$	$2/3$	1
$f(y\|x = 2)$	$1/3$	$2/3$	1
$f(y\|x = 3)$	0	1	1
$f(y\|x = 1, x = 2, x = 3)$	$1/5$	$4/5$	1

cond. distr. $f(x\|y)$	$f(x\|y = 1)$	$f(x\|y = 2)$	$f(x\|y = 1, y = 2)$
$x = 1$	$1/2$	$1/4$	$3/10$
$x = 2$	$1/2$	$1/4$	$3/10$
$x = 3$	0	$1/2$	$4/10$
$\sum_x$	1	1	1

joint distr. $f(x, y)$	$f(x, y = 1)$	$f(x, y = 2)$	marginal pr. $f_x(x)$
$f(x = 1, y)$	$1/10$	$2/10$	$3/10$
$f(x = 2, y)$	$1/10$	$2/10$	$3/10$
$f(x = 3, y)$	0	$4/10$	$4/10$
marginal pr. $f_y(y)$	$2/10$	$8/10$	1

3.2 The joint density function

The joint density function

Two random variables X and Y have **joint density function**

- if x and y are discrete

$$f(x,y) = Prob(a \leq x \leq b, c \leq y \leq d) = \sum_{a \leq x \leq b} \sum_{c \leq y \leq d} f(x,y)$$

- if x and y are continuous

$$f(x,y) = Prob(a \leq x \leq b, c \leq y \leq d) = \int_a^b \int_c^d f(x,y)dxdy$$

Example

With $a = 1, b = 2, c = 2, d = 2$ and the following $f(x,y)$

joint distr.		
$f(x,y)$	$f(x, y = 1)$	$f(x, y = 2)$
$f(x = 1, y)$	$1/10$	$2/10$
$f(x = 2, y)$	$1/10$	$2/10$
$f(x = 3, y)$	0	$4/10$

$Prob(1 \leq x \leq 2, 2 \leq y \leq 2) = f(y = 2, x = 1) + f(y = 2, x = 2) = 2/5.$

For values x and y of two discrete random variable X and Y, the **probability distribution**

$$f(x,y) = Prob(X = x, Y = y).$$

The axioms of probability require

$$f(x,y) \geq 0,$$

$$\sum_x \sum_y f(x,y) = 1.$$

If X and Y are continuous,

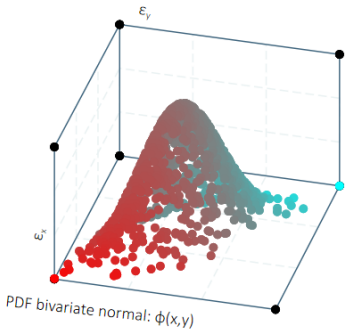
$$\int_x \int_y f(x,y)dxdy = 1.$$

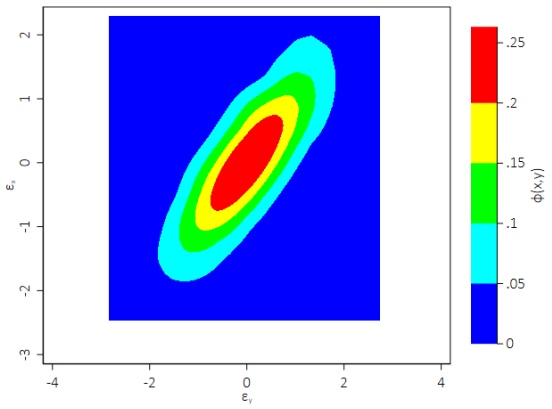
bivariate normal distribution

The bivariate normal distribution is the joint distribution of two normally distributed variables. The density is

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-1/2[(\epsilon_x^2+\epsilon_y^2-2\rho\epsilon_x\epsilon_y)/(1-\rho^2)]},$$

where $\epsilon_x = \frac{x-\mu_x}{\sigma_x}$, and $\epsilon_y = \frac{y-\mu_y}{\sigma_y}$.





3.3 The joint cumulative density function

The joint cumulative density function.

The probability of a joint event of X and Y have **joint cumulative density function**

- if x and y are discrete

$$F(x,y) = Prob(X \leq x, Y \leq y) = \sum_{X \leq x} \sum_{Y \leq y} f(x,y)$$

- if x and y are continuous

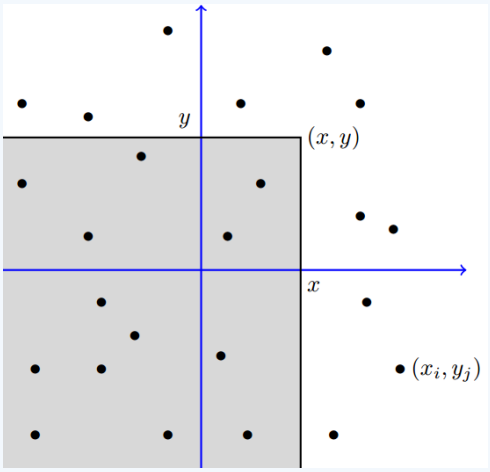
$$F(x,y) = Prob(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(t,s)dsdt$$

Example

With $x = 2, y = 2$ and the following $f(x, y)$

$f(x, y)$	$f(x, y = 1)$	$f(x, y = 2)$
$f(x = 1, y)$	1/10	2/10
$f(x = 2, y)$	1/10	2/10
$f(x = 3, y)$	0	4/10

$Prob(X \leq 2, Y \leq 2) = f(x = 1, y = 1) + f(x = 2, y = 1) + f(x = 1, y = 2) + f(x = 2, y = 2) = 3/5.$



Cumulative probability distribution

For values x and y of two discrete random variable X and Y, the **cumulative probability distribution**

$$F(x, y) = Prob(X \leq x, Y \leq y).$$

The axioms of probability require

$$0 \leq F(x, y) \leq 1,$$

$$F(\infty, \infty) = 1,$$

$$F(-\infty, y) = 0,$$

$$F(x, -\infty) = 0.$$

The marginal probabilities can be found from the joint cdf

$$f_x(x) = P(X \leq x) = Prob(X \leq x, Y \leq \infty) = F(x, \infty).$$

3.4 The marginal probability density

The marginal probability density

To obtain the marginal distributions $f_x(x)$ and $f_y(y)$ from the joint density $f(x, y)$, it is necessary to sum or integrate out the other variable. For example,

- if x and y are discrete
$$f_x(x) = \sum_y f(x, y),$$
- if x and y are continuous
$$f_x(x) = \int_y f(x, s)ds.$$

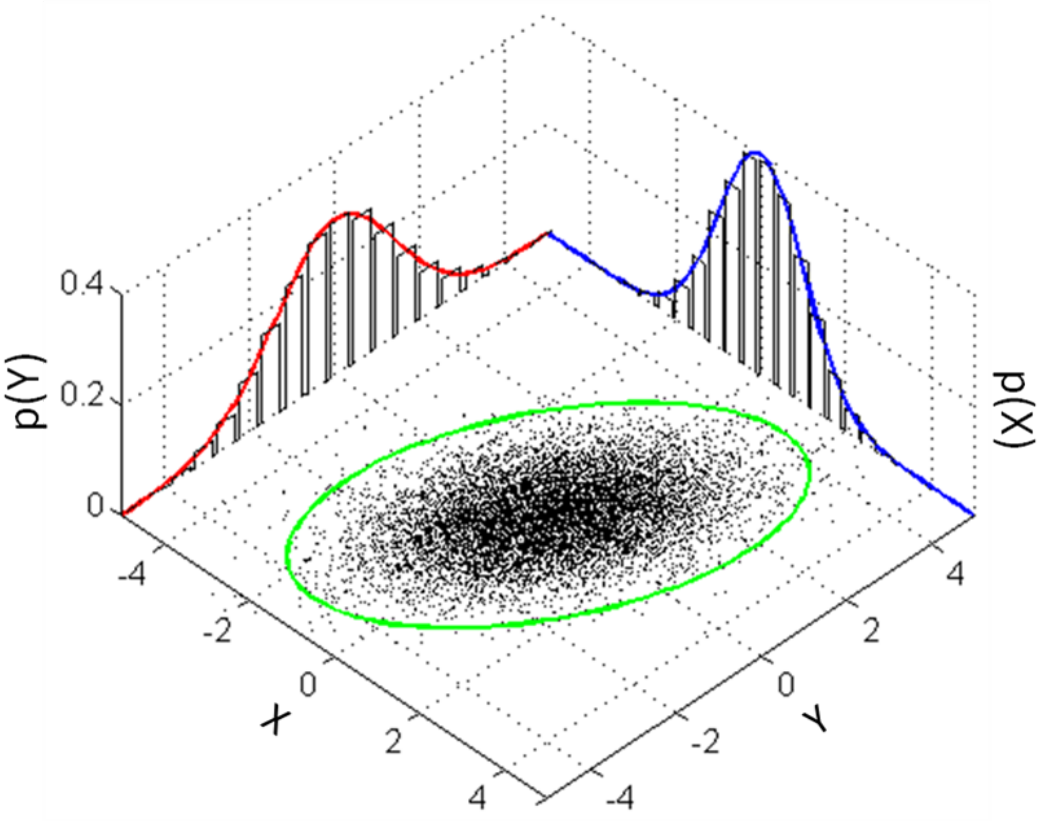
Example

$f(x,y)$	$f(x,y=1)$	$f(x,y=2)$	$f_x(x)$
$f(x=1,y)$	$1/10$	$2/10$	$3/10$
$f(x=2,y)$	$1/10$	$2/10$	$3/10$
$f(x=3,y)$	0	$4/10$	$4/10$
$f_y(y)$	$2/10$	$8/10$	1

$$f_x(x=1) = f(x=1,y=1) + f(x=1,y=2) = 3/10.$$

$$f_y(y=2) = f(x=1,y=2) + f(x=2,y=2) + f(x=3,y=2) = 4/5.$$

The bivariate normal distribution



Why do we care about marginal distributions?

Means, variances, and higher moments of the variables in a joint distribution are defined with respect to the marginal distributions.

- **Expectations**
  If x and y are discrete

$$E[x] = \sum_x x f_x(x) = \sum_x x \left[\sum_y f(x,y) \right] = \sum_x \sum_y x f(x,y).$$

  If x and y are continuous

$$E[x] = \int_x x f_x(x) = \int_x \int_y x f(x,y) dy dx.$$

- **Variances**

$$Var[x] = \sum_x (x - E[x])^2 f_x(x) = \sum_x \sum_y (x - E[x])^2 f(x,y).$$

3.5 Covariance and correlation

For any function $g(x, y)$,

$$E[g(x,y)] = \begin{cases} \sum_x \sum_y g(x,y) f(x,y) & \textit{in the discrete case,} \\ \\ \int_x \int_y g(x,y) f(x,x) dy dx & \textit{in the continuous case.} \end{cases}$$

The covariance of x and y is a special case:

$$\begin{aligned} Cov[x,y] &= E[(x-\mu_x)(y-\mu_y)] \\ &= E[xy] - \mu_x \mu_y = \sigma_{xy} \end{aligned}$$

If x and y are independent, then $f(x, y) = f_x(x) f_y(y)$ and

$$\begin{aligned} \sigma_{xy} &= \sum_x \sum_y f_x(x) f_y(y)(x-\mu_x)(y-\mu_y) \\ &= \sum_x (x-\mu_x) f_x(x) \sum_y (y-\mu_y) f_y(y) = E[x-\mu_x] E[y-\mu_y] = 0. \end{aligned}$$

- correlation $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$
- $\sigma_{xy} = 0$ does not imply independence (except for bivariate normal).

Independence: Pdf and cdf from marginal densities

- Two random variables are statistically independent if and only if their joint density is the product of the marginal densities:

$$f(x,y) = f_x(x) f_y(y) \Leftrightarrow x \textit{ and } y \textit{ are independent.}$$

- If (and only if) x and y are independent, then the marginal cdfs factors the cdf as well:

$$F(x,y) = F_x(x) F_y(y) = Prob(X \leq x, Y \leq y) = Prob(X \leq x) Prob(Y \leq y).$$

Example

$f(x,y)$	$f(x,y=1)$	$f(x,y=2)$	$f_x(x)$
$f(x=1,y)$	$1/6$	$1/6$	$1/3$
$f(x=2,y)$	$1/6$	$1/6$	$1/3$
$f(x=3,y)$	$1/6$	$1/6$	$1/3$
$f_y(y)$	$1/2$	$1/2$	1

$$f_x(x=3) \times f_y(y=2) = 1/3 \times 1/2 = 1/6.$$

$F(x,y)$	$F(x,y=1)$	$F(x,y=2)$
$F(x=1,y)$	$1/6$	$2/6$
$F(x=2,y)$	$2/6$	$4/6$
$F(x=3,y)$	$3/6$	1

$$\begin{aligned} &P(x \leq 2)P(y \leq 2) \\ &= [f(x=2,y=1) + f(x=2,y=2)] \\ &\times [f(x=1,y=2) + f(x=2,y=2)] \\ &= [1/6 + 1/6][1/6 + 1/6] = 4/36 = 2/18. \end{aligned}$$

3.6 The conditional density function

The conditional density function

The **conditional distribution** over y for each value of x (and vice versa) has conditional densities

$$f(y|x) = \frac{f(x,y)}{f_x(x)} \quad f(x|y) = \frac{f(x,y)}{f_y(y)}.$$

The marginal distribution of x averages the probability of x given y over the distribution of all values of y $f_x(x) = E[f(x|y)f(y)]$. If x and y are independent, knowing the value of y does not provide any information about x, so $f_x(x) = f(x|y)$.

Example

cond. distr.			
$f(x\|y)$	$f(x\|y=1)$	$f(x\|y=2)$	$f(x\|y=1,y=2)$
$x=1$	$1/2$	$1/4$	$3/10$
$x=2$	$1/2$	$1/4$	$3/10$
$x=3$	0	$1/2$	$4/10$
$\sum_x$	1	1	1

joint distr.			marginal pr.
$f(x,y)$	$f(x,y=1)$	$f(x,y=2)$	$f_x(x)$
$f(x=1,y)$	$1/10$	$2/10$	$3/10$
$f(x=2,y)$	$1/10$	$2/10$	$3/10$
$f(x=3,y)$	0	$4/10$	$4/10$
marginal pr. $f_y(y)$	$2/10$	$8/10$	1

$$f(x=3|y=2) = \frac{f(x=3,y=2)}{f_y(y=2)} = 4/10 \times 10/8 = 1/2.$$

$$f_x(x=2) = E_y[f(x=2|y)f(y)] = f(x=2|y=1)f(y=1) + f(x=2|y=2)f(y=2)$$

$$= 1/2 \times 2/10 + 1/4 \times 8/10 = 1/10 + 2/10 = 3/10.$$

3.7 Conditional mean aka regression

A random variable may always be written as

$$\begin{aligned} y &= E[y|x] + (y - E[y|x]) \\ &= E[y|x] + \epsilon. \end{aligned}$$

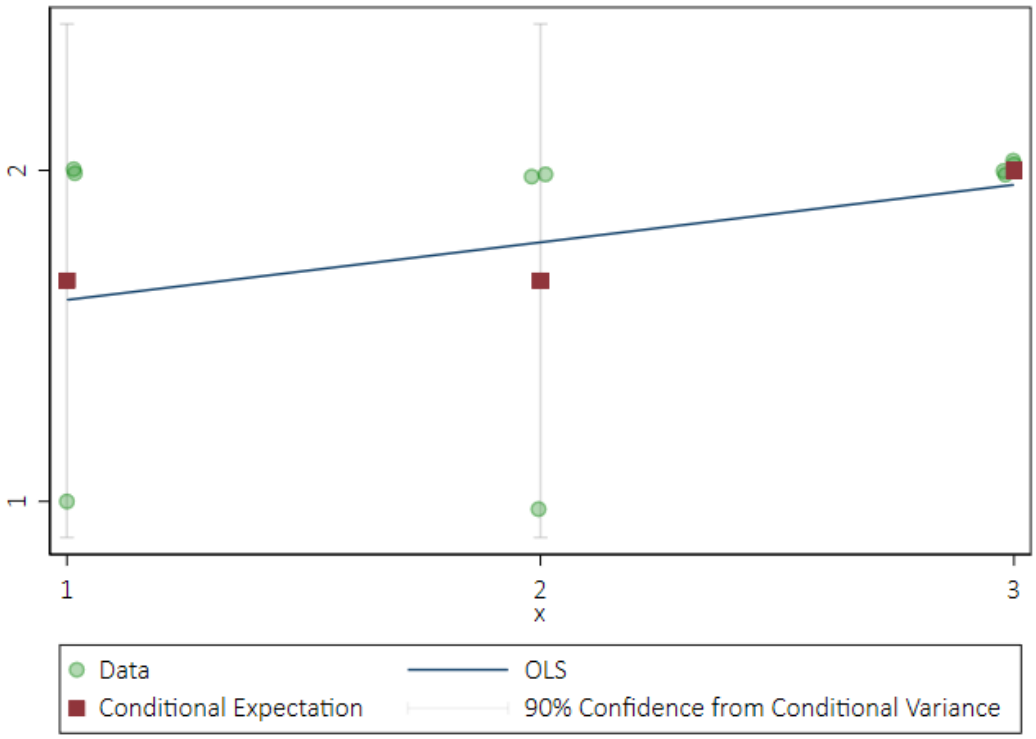
Definition

The regression of y on x is obtained from the **conditional mean**

$$E[y|x] = \begin{cases} \sum_y yf(y|x) & if\ y\ is\ discrete, \\ \\ \int_y yf(y|x)dy & if\ y\ is\ continuous. \end{cases}$$

Predict y at values of x:

$$\sum_y yf(y|x=1) = 1 \times 1/3 + 2 \times 2/3 = 5/3.$$



Conditional variance

A **conditional variance** is the variance of the conditional distribution:

$$Var[y|x] = \begin{cases} \sum_y (y - E[y|x])^2 f(y|x) & if\ y\ is\ discrete, \\ \\ \int_y (y - E[y|x])^2 f(y|x) dy, & if\ y\ is\ continuous. \end{cases}$$

The computation can be simplified by using

$$Var[y|x] = E[y^2|x] - (E[y|x])^2 \geq 0.$$

Decomposition of variance $Var[y] = E_x[Var[y|x]] + Var_x[E[y|x]]$

- When we condition on x, the variance of y reduces on average. $Var[y] \geq E_x[Var[y|x]]$
- $E_x[Var[y|x]]$ is the average of variances **within** each x
- $Var_x[E[y|x]]$ is variance **between** y averages in each x.
- $E[y|x = 1] = 1.67$, $E[y|x = 2] = 1.67$, and $E[y|x = 3] = 2$
- $V[y|x = 1] = 0.22$, $V[y|x = 2] = 0.22$, and $V[y|x = 3] = 0$

Example

$f(y\|x)$	$y=1$	$y=2$	
$f(y\|x=1)$	1/3	2/3	1
$f(y\|x=2)$	1/3	2/3	1
$f(y\|x=3)$	0	1	1

$E[y|x=1] = 1/3\times 1+2/3\times 2 = 5/3$

$E[y|x=2] = 1/3\times 1+2/3\times 2 = 5/3$

$E[y|x=3] = 0\times 1+1\times 2 = 2$

$f(x,y)$	$f(x,y=1)$	$f(x,y=2)$	$f_x(x)$
$f(x=1,y)$	1/10	2/10	3/10
$f(x=2,y)$	1/10	2/10	3/10
$f(x=3,y)$	0	4/10	4/10
$f_y(y)$	2/10	8/10	1

$V[y|x=1] = 1^2\times 1/3+2^2\times 2/3-(5/3)^2 = 2/9$

$V[y|x=2] = 1^2\times 1/3+2^2\times 2/3-(5/3)^2 = 2/9$

$V[y|x=3] = 1^2\times 0+2^2\times 1-2^2 = 0$

alternatively (requiring more differences)

$V[y|x=1] = (1-5/3)^2\times 1/3+(2-5/3)^2\times 2/3 = 2/9$

Average of variances **within** each x, $E[V[y|x]]$ is less or equal total variance $V[y]$.

Example

- Use the conditional mean to calculate $E[y]$:

$$E[y] = E_x[E[y|x]] = E[y|x=1]f(x=1)+E[y|x=2]f(x=2)+E[y|x=3]f(x=3)$$

$$= 5/3\times 3/10+5/3\times 3/10+2\times 4/10 = 9/5.$$

$$E[y] = \sum_y f_y(y) = 1\times 2/10+2\times 8/10 = 9/5.$$

- Variation in y, $V[y|x=1] = 0.22$, $V[y|x=2] = 0.22$, and $V[y|x=3] = 0$ due to variation in x, is on average
  $E[V[y|x]] = 3/10\times 2/9+3/10\times 2/9+4/10\times 0 = 2/15$.

- For each conditional mean $E[y|x=1] = 5/3$, $E[y|x=2] = 5/3$, and $E[y|x=3] = 2$, y varies with
  $V[E[y|x]] = E[(E[y|x])^2]-(E[y|x])^2 = 3/10\times(5/3)^2+3/10\times(5/3)^2+4/10\times(2)^2-(9/5)^2 = 2/75$.

- $E[V[y|x]] + V[E[y|x]] = V[y] = 2/75 + 2/15 = 4/25$.
  With degree of freedom correction $(n-1)$ (as reported in software):
  $E[V[y|x]] + V[E[y|x]] = V[y] = 2/75/(10-1) \times 10 + 2/15/(10-1) \times 10 = 8/45$.

3.8 The bivariate normal

Properties of the bivariate normal

Recall bivariate normal distribution is the joint distribution of two normally distributed variables. The density is

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-1/2[(\epsilon_x^2+\epsilon_y^2-2\rho\epsilon_x\epsilon_y)/(1-\rho^2)]},$$

where $\epsilon_x = \frac{x-\mu_x}{\sigma_x}$, and $\epsilon_y = \frac{y-\mu_y}{\sigma_y}$.

The covariance is $\sigma_{xy} = \rho_{xy}\sigma_x\sigma_y$, where

- $-1 < \rho_{xy} < 1$ is the correlation between x and y
- $\mu_x, \sigma_x, \mu_y, \sigma_y$ are means and standard deviations of the marginal distributions of x or y

If x and y are bivariately normally distributed $(x,y) \sim N_2[\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy}]$

- the marginal distributions are normal

$$f_x(x) = N[\mu_x, \sigma_x^2]$$

$$f_y(y) = N[\mu_y, \sigma_y^2]$$

- the conditional distributions are normal

$$f(y|x) = N[\alpha + \beta x, \sigma_y^2(1-\rho^2)]$$

$$\alpha = \mu_y - \beta\mu_x; \beta = \frac{\sigma_{xy}}{\sigma_x^2}$$

- $f(x,y) = f_x(x)f_x(x)$ if $\rho_{xy} = 0$: x and y are independent if and only if they are uncorrelated

3.9 Useful rules

- $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$
- $E[ax + by + c] = aE[x] + bE[y] + c$
- $Var[ax + by + c] = a^2 Var[x] + b^2 Var[y] + 2abCov[x, y] = Var[ax + by]$
- $Cov[ax + by, cx + dy] = acVar[x] + bdVar[y] + (ad + bc)Cov[x, y]$
- If X and Y are uncorrelated, then $Var[x + y] = Var[x - y] = Var[x] + Var[y]$.
- Linearity
$$E[ax + by|z] = aE[x|z] + bE[y|z].$$
- Adam's Law / Law of Iterated Expectation
$$E[y] = E_x[E[y|x]]$$
- Adam's general Law / Law of Iterated Expectation
$$E[y|g_2(g_1(x))] = E[E[y|g_1(x)]|g_2(g_1(x))]$$
- Independence
If x and y are independent, then
$$E[y] = E[y|x],$$
$$E[g_1(x)g_2(y)] = E[g_1(x)]E[g_2(y)].$$
- Taking out what is known
$$E[g_1(x)g_2(y)|x] = g_1(x)E[g_2(y)|x].$$
- Projection of y by $E[y|x]$, such that orthogonal to $h(x)$
$$E[(y - E[y|x])h(x)] = 0.$$
- Keeping just what is needed (y predictable from x needed, not residual)
$$E[xy] = E[xE[y|x]].$$

- Eve's Law (EVVE) / Law of Total Variance

$$Var[y] = E_x[Var[y|x]] + Var_x[E[y|x]]$$

- ECCE law / Law of Total Covariance

$$Cov[x, y] = E_z[Cov[y, x|z]] + Cov_z[E[x|z], E[y|z]]$$

- $Cov[x, y] = Cov_x[x, E[y|x]] = \int_x (x - E[x])\, E[y|x] f_x(x) dx.$
- If $E[y|x] = \alpha + \beta x$, then $\alpha = E[y] - \beta E[x]$ and $\beta = \frac{Cov[x,y]}{Var[x]}$
- Regression variance $Var_x[E[y|x]]$, because $E[y|x]$ varies with x
- Residual variance $E_x[Var[y|x]] = Var[y] - Var_x[E[y|x]]$, because y varies around the conditional mean
- Decomposition of variance $Var[y] = Var_x[E[y|x]] + E_x[Var[y|x]]$
- Coefficient of determination $= \frac{\text{regression variance}}{\text{total variance}}$
- If $E[y|x] = \alpha + \beta x$ and if $Var[y|x]$ is a constant, then

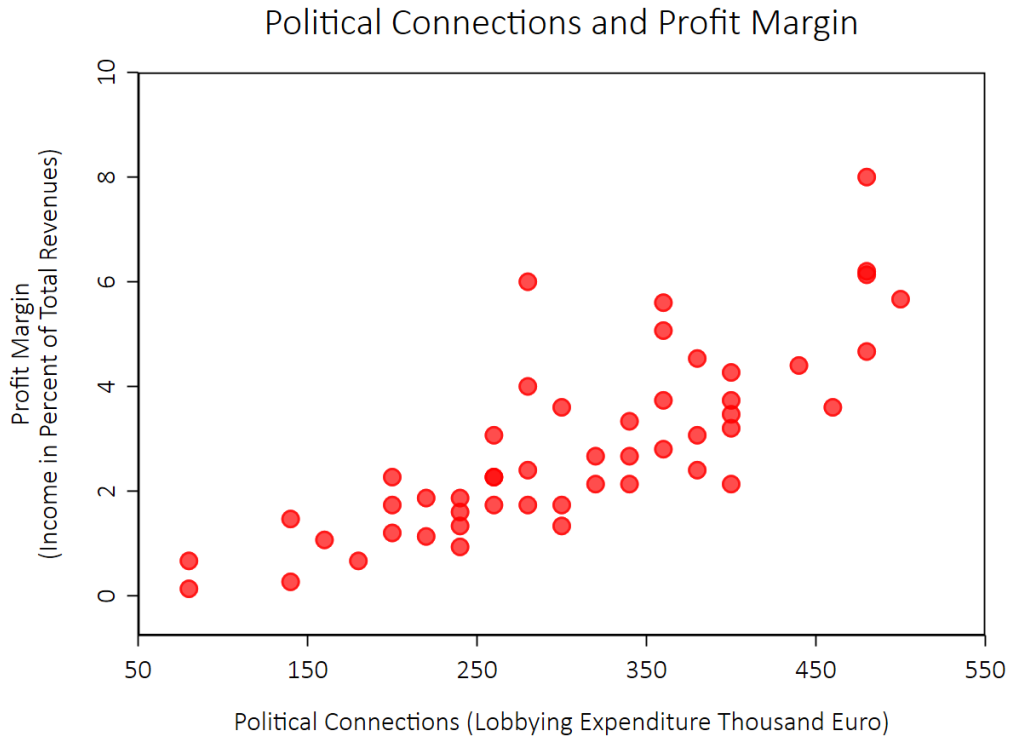
$$Var[y|x] = Var[y]\left(1 - Corr^2[y, x]\right) = \sigma_y^2\left(1 - \sigma_{xy}^2\right)$$

4 The Least Squares Estimator

4.1 What is the Relationship between Two Variables?

Political Connections and Firms

Firm profits increase with the degree of political connections

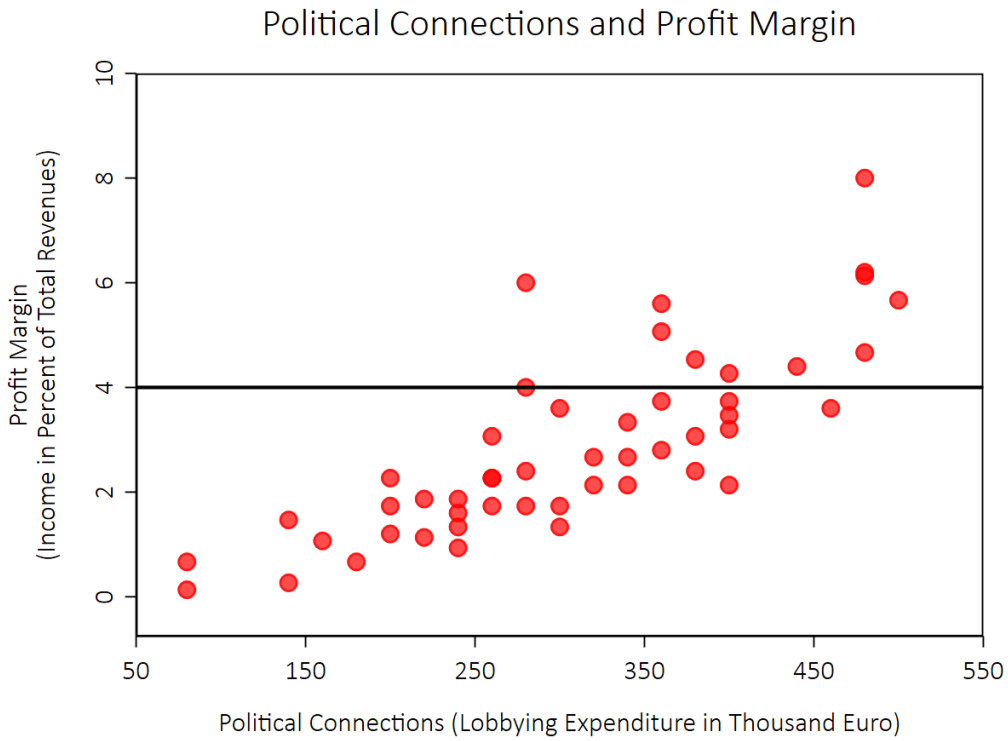


- Learn how to represent relationships between two or more variables
- How to quantify and predict effects of shocks and policy changes
- Show properties of the OLS estimator in small & large samples
- Apply Monte Carlo Simulations to assess properties of OLS

4.2 The Econometric Model

Specification of a Linear Regression

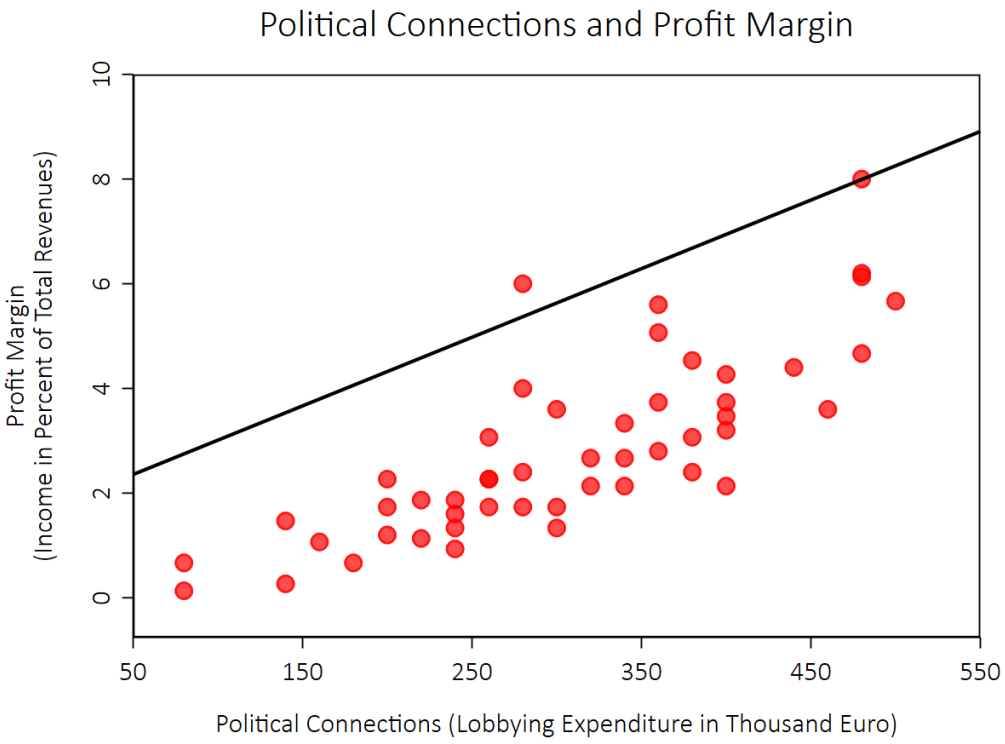
- dependent variable y_i = profits of firm i
- explanatory variables $x_{i1}, \ldots, x_{iK}$ $k = 1, \ldots K$ political connections, other firm characteristics
- $x_{i0} = 1$ is a constant
- parameters to be estimated $\beta_0, \beta_1, \ldots, \beta_K$ are $K + 1$
- u_i is called the error term



$$y_i = (\beta_0 = 4) + (\beta_1 = 0)x_{i1} + u_i.$$

- dependent variable y_i = profits of firm i
- explanatory variables $x_{i1}, \ldots, x_{iK}$ $k = 1, \ldots K$ political connections, other firm characteristics
- $x_{i0} = 1$ is a constant
- parameters to be estimated $\beta_0, \beta_1, \ldots, \beta_K$ are $K + 1$

- u_i is called the error term



$$y_i = (\beta_0 = 2.36) + (\beta_1 = 0.01)x_{i1} + u_i.$$

How Were the Data Generated?

The *data generating process* is fully described by a set of assumptions.

The Five Assumptions of the Econometric Model

- LRM1: Linearity
- LRM2: Simple random sampling
- LRM3: Exogeneity
- LRM4: Error variance
- LRM5: Identifiability

Data Generating Process: Linearity

LRM1: Linearity

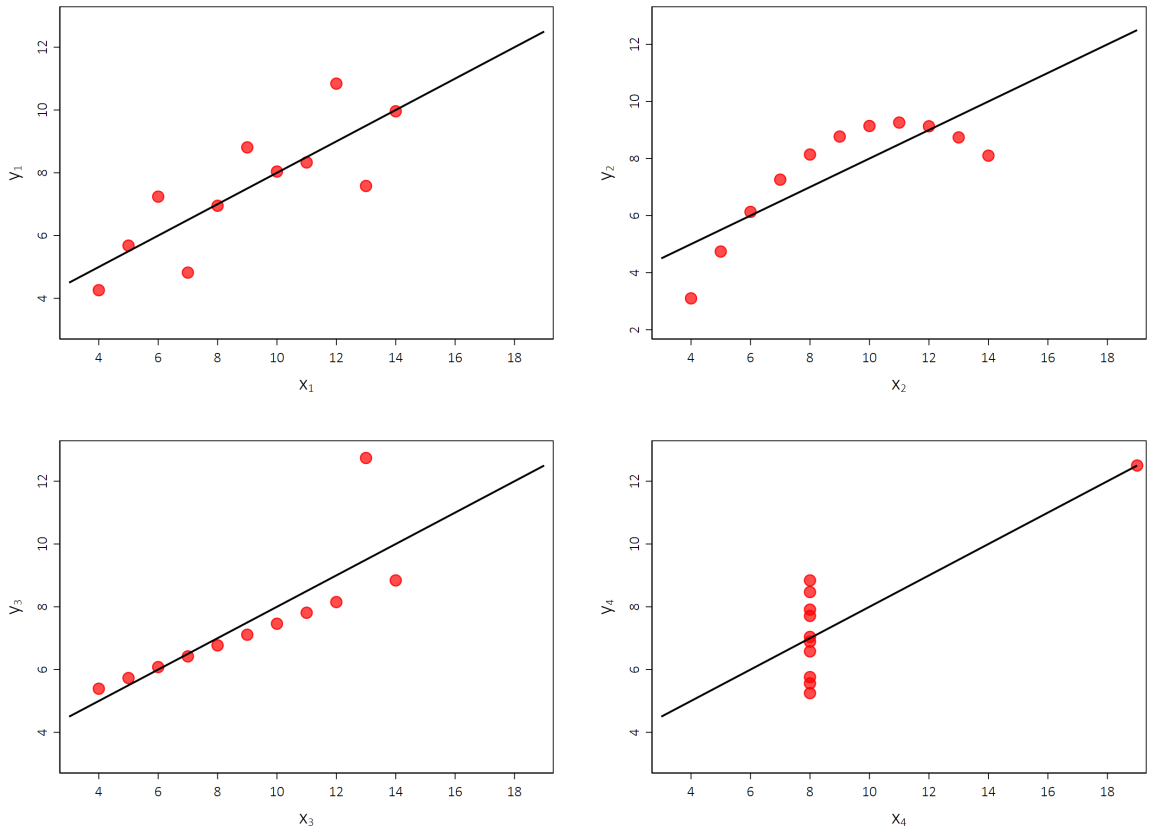
$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_K x_{iK} + u_i \text{ and } E(u_i) = 0.$$

LRM1 assumes that the

- functional relationship is linear in parameters β_k
- error term u_i enters additively
- parameters β_k are constant across individual firms i and $j \neq i$.

Anscombe's Quartet

Figure 1: All four sets are identical when examined using linear statistics, but very different when graphed. Correlation between x and y is 0.816. Linear Regression y = 3.00 + 0.50x.



Data Generating Process: Random Sampling

LRM2: Simple Random Sampling

$$\{x_{i1}, \ldots, x_{iK}, y_i\}_{i=1}^{N} \quad \text{i.i.d. (independent and identically distributed)}$$

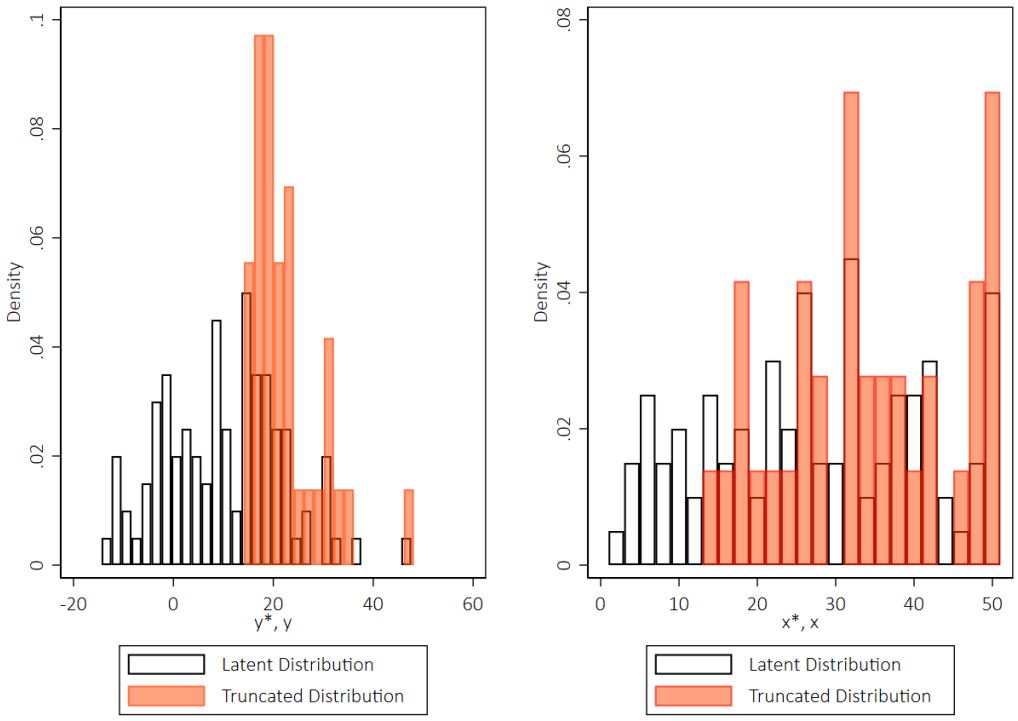
LRM2 means that

- observation i has no information content for observation $j \neq i$
- all observations i come from the same distribution

This assumption is guaranteed by simple random sampling provided there is no systematic non-response or truncation.

Density of Population and Truncated Sample

Figure 2: Distribution of a dependent variable and an independent variable truncated at $y^* = 15$.



Data Generating Process: Exogeneity

LRM3: Exogeneity

1. $$u_i|x_{i1}, \ldots, x_{iK} \sim N(0, \sigma_i^2)$$

   LRM3a assumes that the error term is normally distributed conditional on the explanatory variables.

2. $$u_i \perp x_{ik} \quad \forall k \quad \text{(independent)}, pdf_{u,x}(u_i x_{ik}) = pdf_u(u_i) pdf_x(x_{ik})$$

   LRM3b means that the error term is independent of the explanatory variables.

3. $$E(u_i|x_{i1}, \ldots, x_{iK}) = E(u_i) = 0 \quad \text{(mean independent)}$$

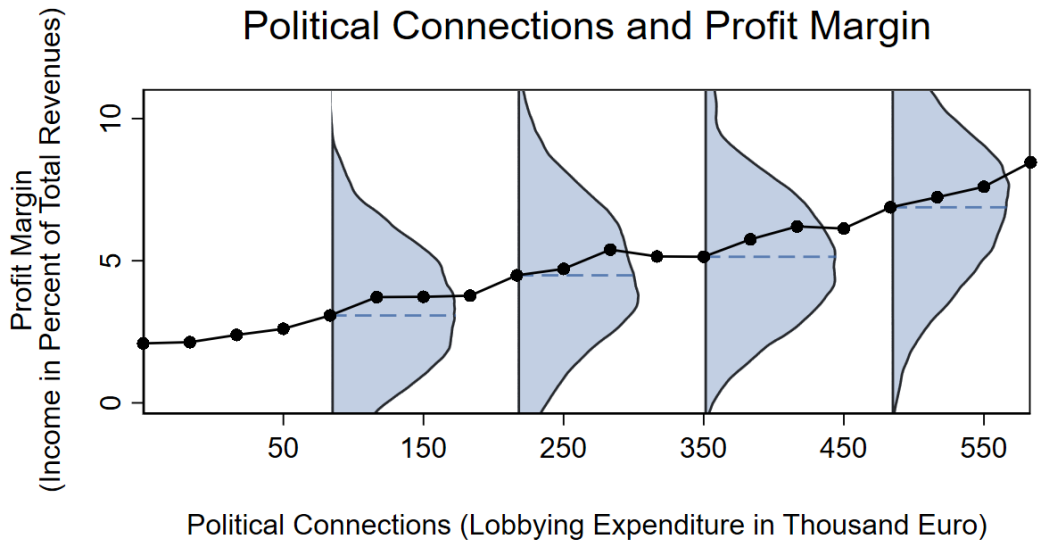
   LRM3c states that the *mean* of the error term is independent of explanatory variables.

4. $$cov(x_{ik}, u_i) = 0 \quad \forall k \quad \text{(uncorrelated)}$$

   LRM3d means that the error term and the explanatory variables are uncorrelated.

LRM3a or LRM3b imply LRM3c and LRM3d. LRM3c implies LRM3d.

Figure 3: Distributions of the dependent variable conditional on values of an independent variable.



Weaker exogeneity assumption if interest only in, say, x_{i1}:

Conditional Mean Independence $E(u_i|x_{i1}, x_{i2}, \ldots, x_{iK}) = E(u_i|x_{i2}, \ldots, x_{iK})$

Given the control variables $x_{i2}, \ldots, x_{iK}$, the mean of u_i does not depend on the variable of interest x_{i1}.

Data Generating Process: Error Variance

LRM4: Error Variance

1.
$$V(u_i|x_{i1},\ldots,x_{iK}) = \sigma^2 < \infty \quad \text{(homoskedasticity)}$$

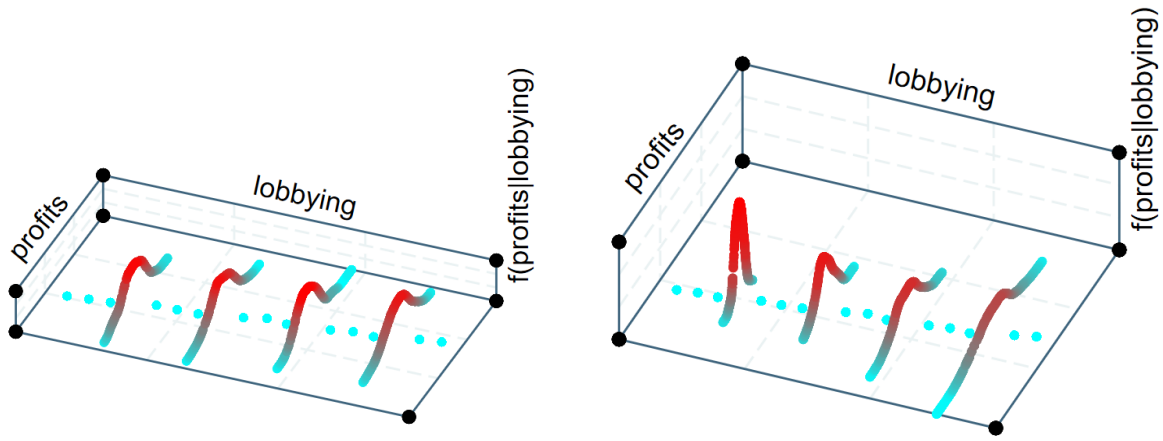
LRM4a means that the variance of the error term is a constant.

2.
$$V(u_i|x_{i1},\ldots,x_{iK}) = \sigma_i^2 = g(x_{i1},\ldots,x_{iK}) < \infty \quad \text{(cond. heteroskedasticity)}$$

LRM4b allows the variance of the error term to depend on a function g of the explanatory variables.

Heteroskedasticity

Figure 4: The simple regression model under homo- and heteroskedasticity. $Var(profits|lobbying, employees)$ increasing with *lobbying*.



Data Generating Process: Identifiability

LRM5: Identifiability

$$(x_{i0}, x_{i1}, \ldots, x_{iK}) \text{ are not linearly dependent}$$

$$0 < V(x_{ik}) < \infty \quad \forall k > 0$$

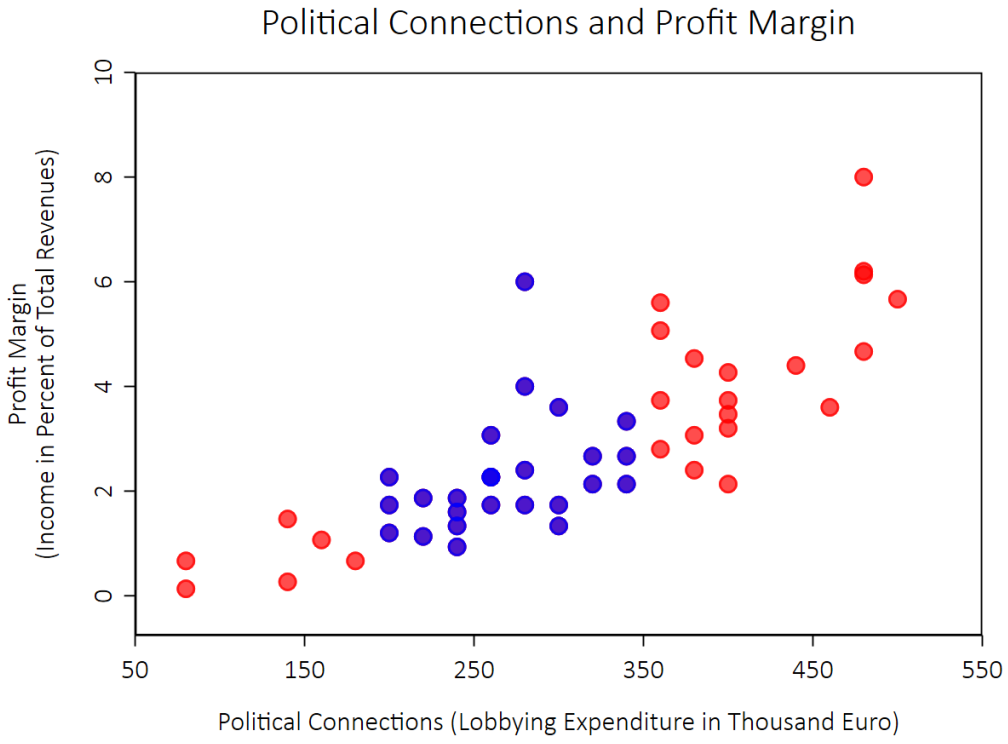
LRM5 assumes that

- the regressors are not *perfectly collinear*, i.e. no variable is a linear combination of the others
- all regressors (but the constant) have strictly positive variance both in expectations and in the sample and not too many extreme values.

LRM5 means that every explanatory variable adds additional information.

The Identifying Variation from x_{ik}

Figure 5: The number of red and blue dots is the same. Using which would you get a more accurate regression line?



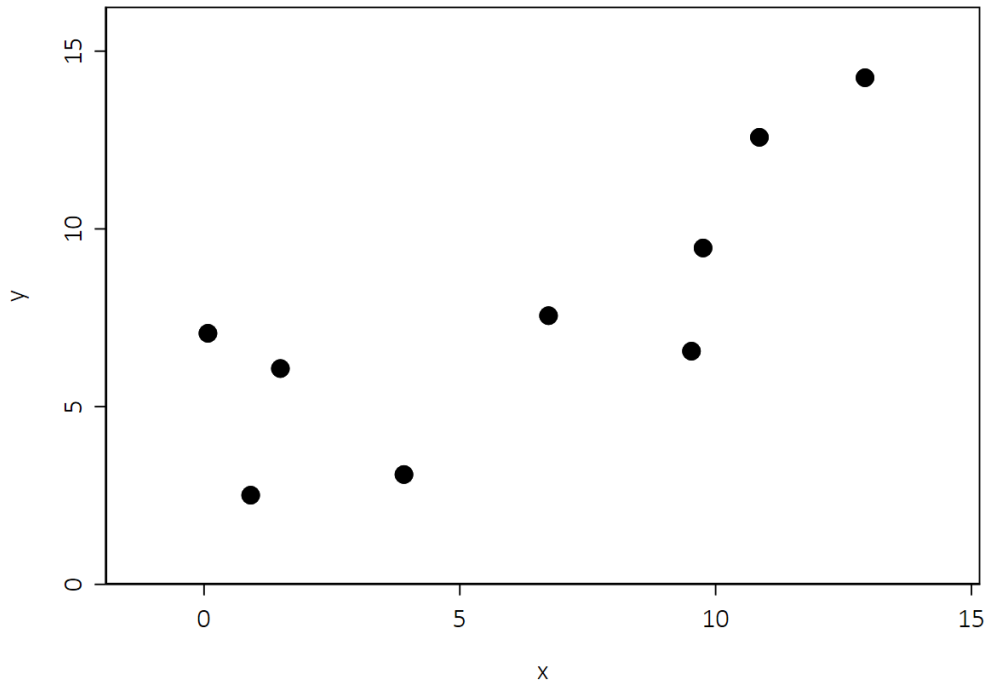
4.3 Estimation with OLS

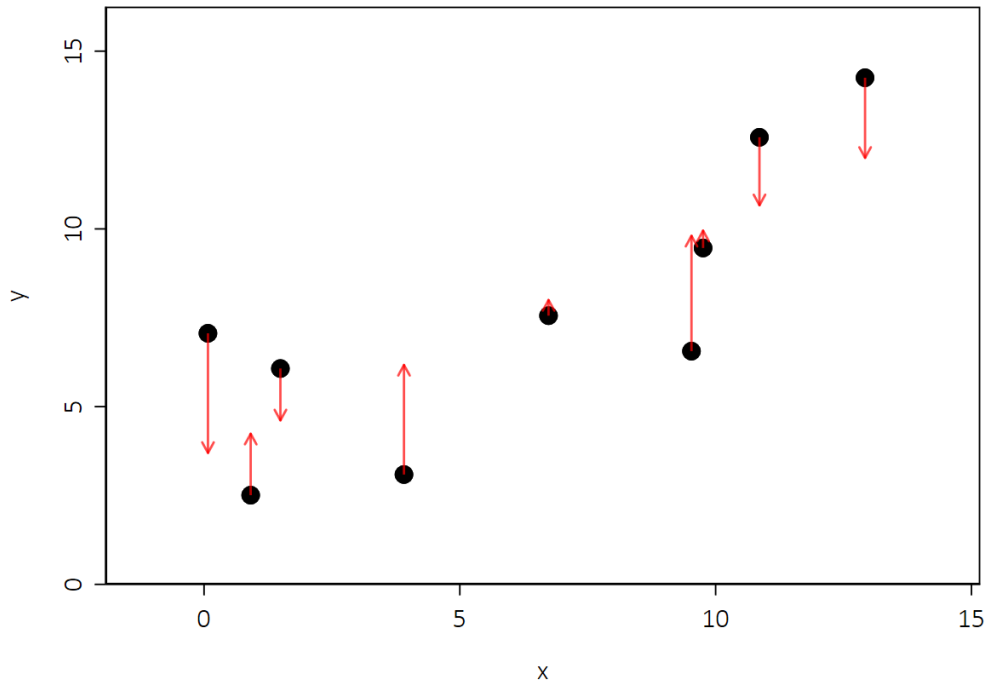
Ordinary least squares (OLS) minimizes the squared distances (SD) between the observed and the predicted dependent variable y:

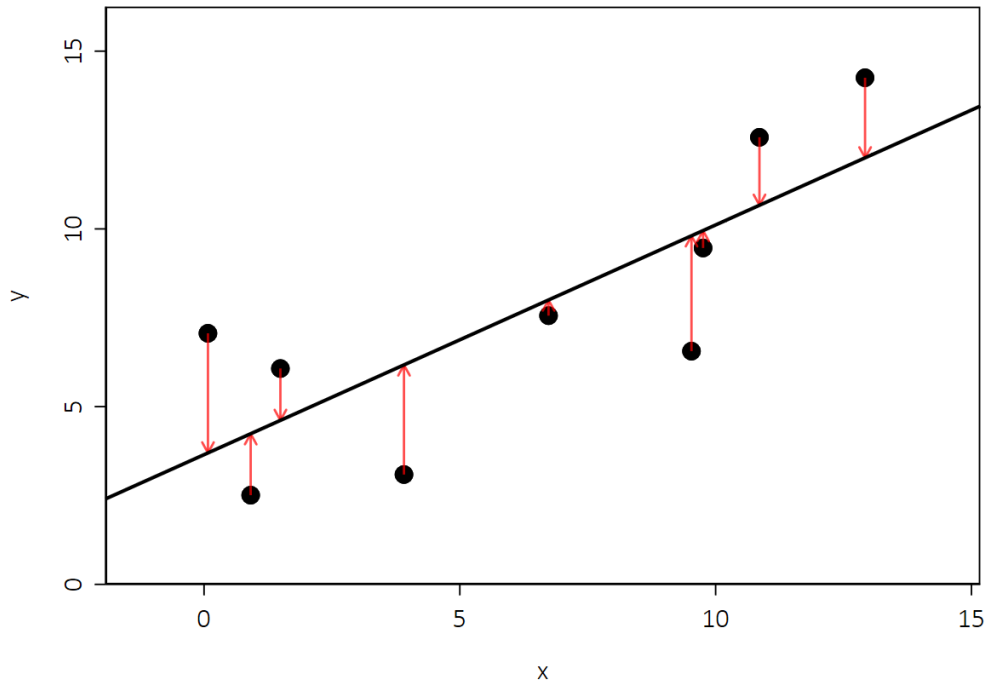
$$\min_{\beta_0,\ldots,\beta_K} SD(\beta_0,\ldots,\beta_K),$$

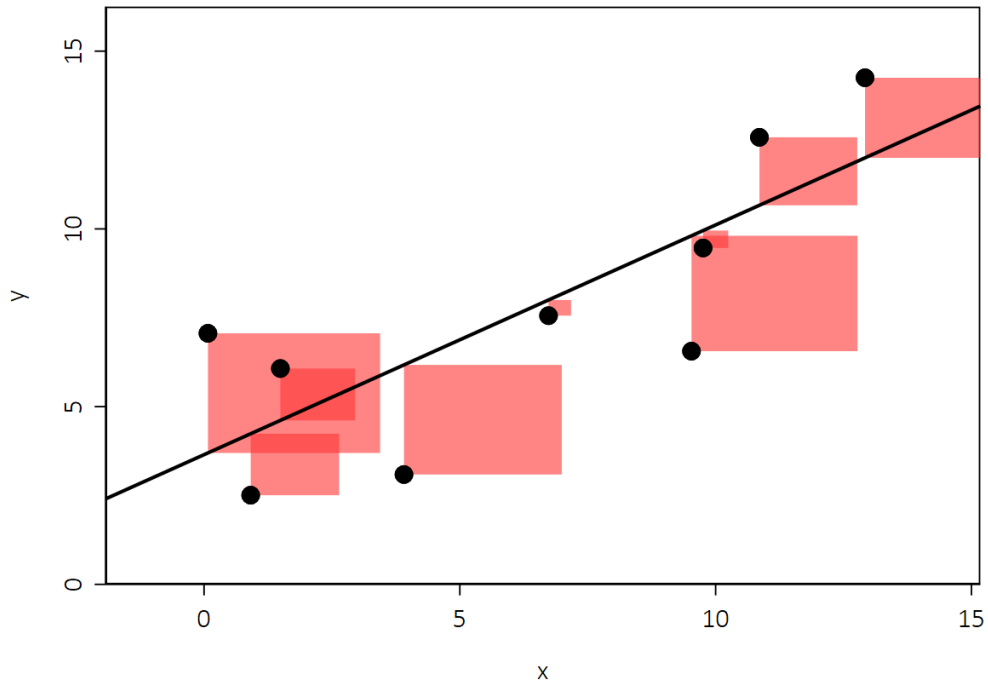
$$\text{where } SD = \sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_K x_{iK})]^2.$$

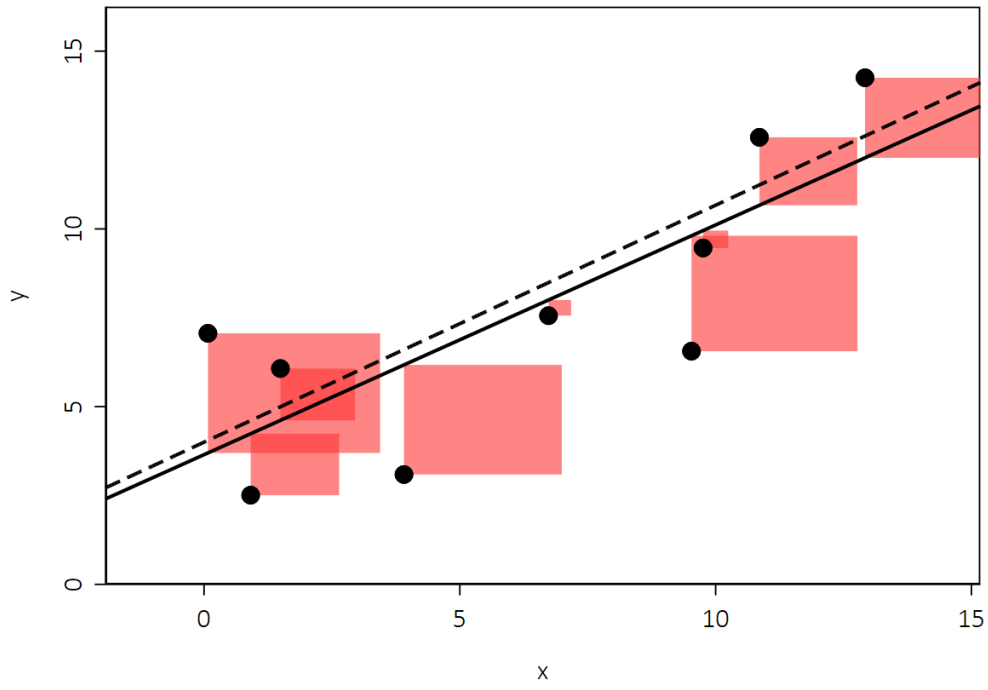
How to Describe the Relationship Best?




y
x
0
5
10
15


y
x
0
5
10
15


y
x
0
5
10
15


y
x
0
5
10
15

Invention of OLS

Legendre to Jacobi (Paris, 30 November 1827, Plackett, 1972): "*...How can Mr. Gauss have dared to tell you that the greater part of your theorems were known to him...?*

... this is the same man ... who wanted to appropriate in 1809 the method of least squares published in 1805.

— Other examples will be found in other places, but a man of honour should refrain from imitating them."

Figure 6: Watercolor caricature of Legendre by Boilly (1820), the only existing portrait known.

Invention of OLS

Legendre to Jacobi (Paris, 30 November 1827, Plackett, 1972): "*...How can Mr. Gauss have dared to tell you that the greater part of your theorems were known to him...?*

... this is the same man ... who wanted to appropriate in 1809 the method of least squares published in 1805.

— Other examples will be found in other places, but a man of honour should refrain from imitating them."

Figure 7: Portrait of Gauss by Jensen (1840).

Estimation with OLS

For the bivariate regression model, the OLS estimators of β_0 and β_1 are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_{i1} - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_{i1} - \bar{x})^2} = \frac{cov(x,y)}{var(x)}$$

$$\hat{\beta}_1 = cov(x,y)/(s_x s_x) = Rs_y/s_x,$$

where $R \equiv cov(x,y)/(s_x s_y)$ is **Pearson's correlation coefficient** with s_z denoting the standard deviation of z.

OLS estimator Measures Linear Correlation

Equivalently,

$$R = s_x/s_y\hat{\beta}_1 = \frac{\hat{\beta}_1 \sum_{i=1}^N (x_{i1} - \bar{x})}{\sum_{i=1}^N (y_i - \bar{y})} = \frac{\sum_{i=1}^N (\hat{\beta}_1 x_{i1} - \hat{\beta}_1 \bar{x})}{\sum_{i=1}^N (y_i - \bar{y})}.$$

Squaring gives

$$R^2 = \frac{\sum_{i=1}^N (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^N (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^N \hat{u}_i^2}{\sum_{i=1}^N (y_i - \bar{y})^2}.$$

R^2 as measure of the **goodness of fit**:
The fit improves with the fraction of the sample variation in y that is explained by the x.

The Case with K Explanatory Variables

The more general case with K explanatory variables is

$$\underset{(K+1)\times 1}{\hat{\beta}} = \underset{(K+1)\times(K+1)}{(X'X)^{-1}} \underset{(K+1)\times N}{X'} \underset{N\times 1}{y}$$

Given the OLS estimator, we can predict the

- dependent variable by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_K x_{iK}$
- the error term by $\hat{u}_i = y_i - \hat{y}_i$.

$\hat{u}_i$ is called the *residual*.

Adjusted $R^2 = 1 - \frac{N-1}{N-K-1} \frac{\sum_{i=1}^N \hat{u}_i^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$.

Figure 8: Scatter cloud visualized with
GRAPH3D for Stata.

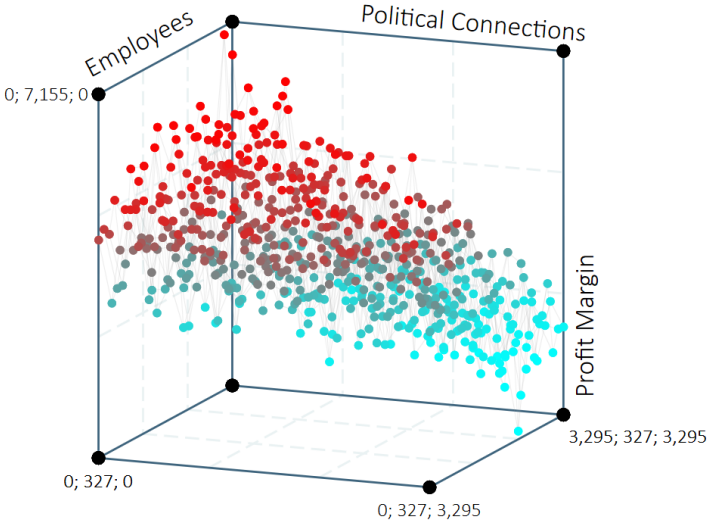
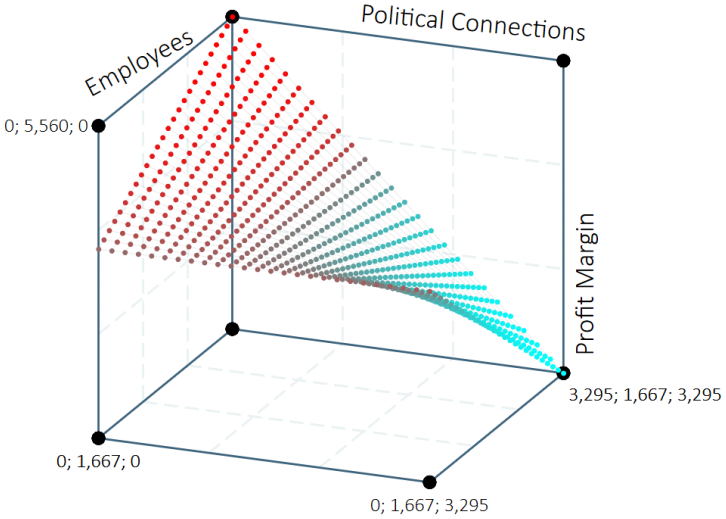


Figure 9: OLS surface visualized with
GRAPH3D for Stata.



4.4 Properties of the OLS Estimator in the Small and in the Large

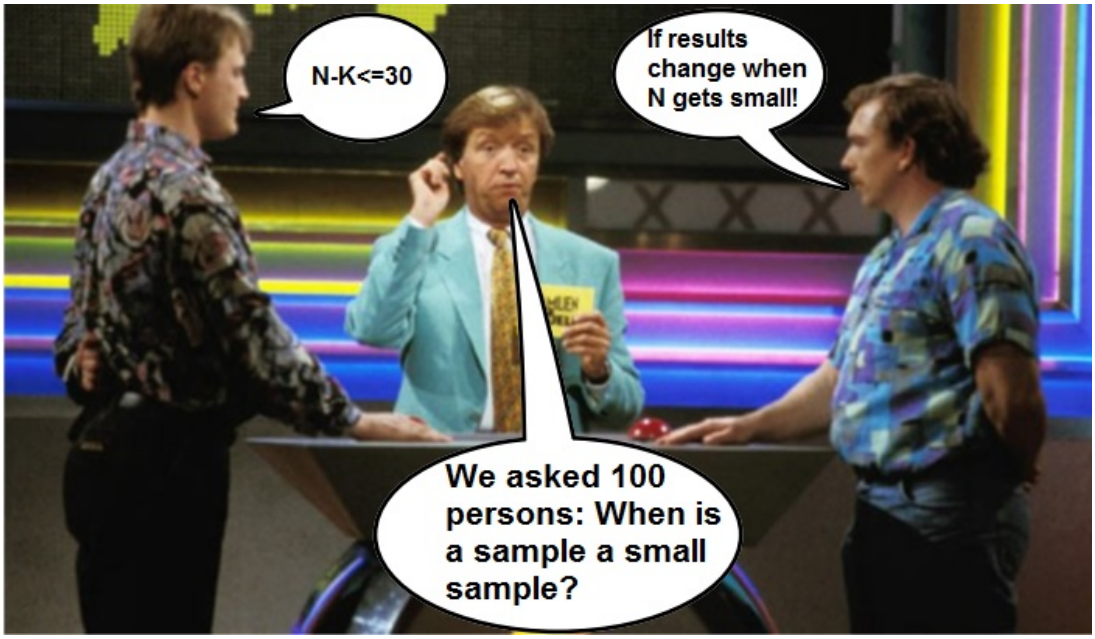
Properties of the OLS Estimator

- *Small sample properties of* $\hat{\beta}$
  - unbiased
  - normally distributed
  - efficient

- *Large sample properties of* $\hat{\beta}$
  - consistent
  - approx. normal
  - asymptotically efficient

Figure 10: What is a small sample?
Familien-Duell
Light Entertainment.

Figure 11: What is a small sample? (Wooldridge, 2009, p. 755): "But large sample approximations have been known to work well for sample sizes as small as $N = 20$." *Source:* Familien-Duell Grundy Light Entertainment.



Unbiasedness and Normality of $\hat{\beta}_k$

Assuming LRM1, LRM2, LRM3a, LRM4, and LRM5,
the following properties can be established even for small samples.

- The OLS estimator of β is **unbiased**.

$$E(\hat{\beta}_k|x_{11},\ldots,x_{NK}) = \beta_k.$$

- The OLS estimator is (multivariate) **normally distributed**.

$$\hat{\beta}_k|x_{11},\ldots,x_{NK} \sim N(\beta_k, V(\hat{\beta}_k)).$$

- Under homoskedasticity (LRM4a)
  the variance $\widehat{V}(\hat{\beta}_k|x_{11},\ldots,x_{NK})$ can be **unbiasedly** estimated.

Variance of $\hat{\beta}_k$ and Efficiency

- For the bivariate regression model, it is estimated as

$$\widehat{V} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \text{ with}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}\hat{u}_i^2}{N-K-1}.$$

- Gauß-Markov-Theorem: under homoskedasticity (LRM4a)
  $\hat{\beta}_k$ is the **BLUE** (best linear unbiased estimator, e.g., non-linear least squares biased).

- $\widehat{V}(\hat{\beta}_k)$ inflates with
  - **micronumerosity** (small sample size)
  - **multicollinearity** (high (but not perfect) correlation between two or more of the independent variables).

Unbiasedness

- The OLS estimator of β is *unbiased.*
  Plug $y = X\beta + u$ into the formula for $\hat{\beta}$ and then use the law of iterated expectation to first take expectation with respect to u conditional on X and then take the unconditional expectation:

$$\mathrm{E}[\,\hat{\beta}] = E_{X,u}\Big[(X'X)^{-1}X'(X\beta + u)\Big]$$

$$= \beta + E_{X,u}\Big[(X'X)^{-1}X'u\Big]$$
$$= \beta + E_X\Big[E_{u|X}\Big[(X'X)^{-1}X'u|X\Big]\Big]$$
$$= \beta + E_X\Big[(X'X)^{-1}X'E_{u|X}[u|X]\Big]$$
$$= \beta,$$

where $E[u|X] = 0$ by assumptions of the model.

Variance

- The OLS estimator β has variance $\widehat{V}(\hat{\beta}_k|x_{11},\ldots,x_{NK}) = \sigma^2(X'X)^{-1}$
  Let $\sigma^2 I$ denote the covariance matrix of u. Then,

$$E[\,(\hat{\beta}-\beta)(\hat{\beta}-\beta)'] = E\Big[((X'X)^{-1}X'u)((X'X)^{-1}X'u)'\Big]$$
$$= E\Big[(X'X)^{-1}X'uu'X(X'X)^{-1}\Big]$$
$$= E\Big[(X'X)^{-1}X'\sigma^2X(X'X)^{-1}\Big]$$
$$= E\Big[\sigma^2(X'X)^{-1}X'X(X'X)^{-1}\Big]$$
$$= \sigma^2(X'X)^{-1},$$

where we used the fact that $\hat{\beta}-\beta$ is just an affine transformation of u by the matrix $(X'X)^{-1}X'$.

Estimator for Variance

For a simple linear regression model, where $\beta = [\beta_0, \beta_1]'$ (β_0 is the y-intercept and β_1 is the slope), one obtains

$$\sigma^2(X'X)^{-1} = \sigma^2\left(\sum x_i x_i'\right)^{-1}$$
$$= \sigma^2\left(\sum (1,x_i)'(1,x_i)\right)^{-1}$$
$$= \sigma^2\left(\sum \begin{pmatrix} 1 x_i \\ x_i x_i^2 \end{pmatrix}\right)^{-1}$$
$$= \sigma^2\begin{pmatrix} N\sum x_i \\ \sum x_i \sum x_i^2 \end{pmatrix}^{-1}$$

$$= \sigma^2 \cdot \frac{1}{N\sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 - \sum x_i \\ -\sum x_i N \end{pmatrix}$$

$$= \sigma^2 \cdot \frac{1}{N\sum_{i=1}^{N}(x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2 - \sum x_i \\ -\sum x_i N \end{pmatrix}$$

$$Var(\beta_1) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}.$$

Parameter Values for Simulations

Monte Carlo Simulations show the distribution of the estimate. Suppose the data generating process is

$$y_i = \beta_0 + \beta_1 x_{i1} + u_i.$$

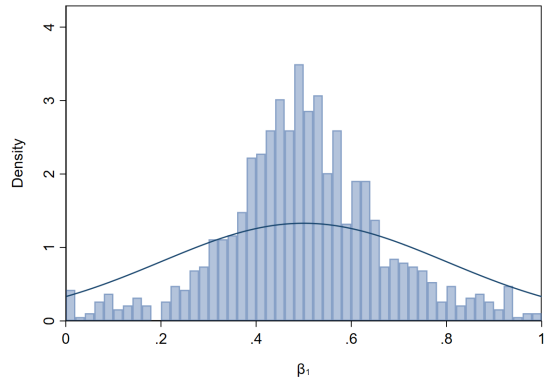
- $\beta_0 = 2.00$
- $\beta_1 = 0.5$
- $u_i \sim N(0.00, 1.00)$
- $N = 3, N = 5, N = 10,$
  $N = 25, N = 100, N = 1000$

Try it yourself...

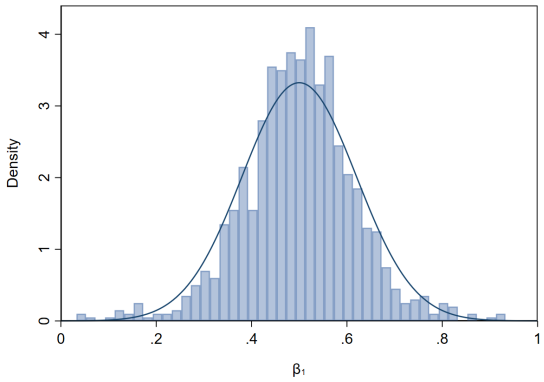
How to Establish Asymptotic Properties of $\hat{\beta}_k$?

Law of Large Numbers

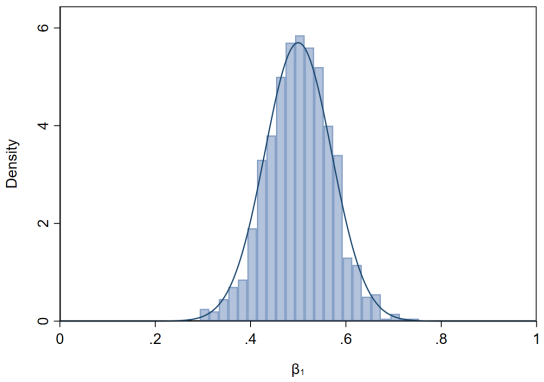
As N increases, the distribution of $\hat{\beta}_k$ becomes more tightly centered around β_k.



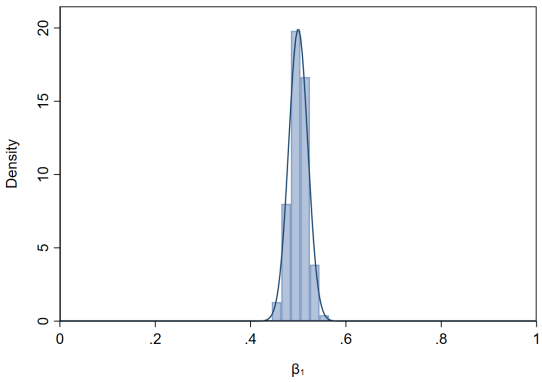
(a) N=3



(b) N=5



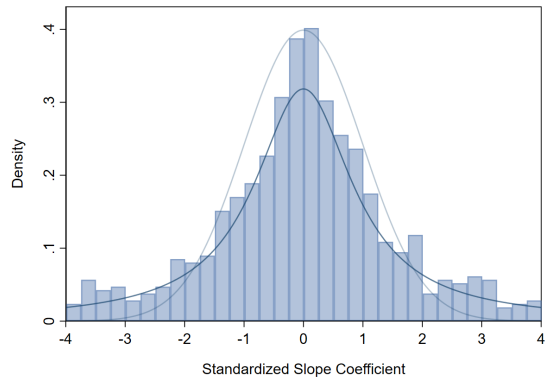
(c) N=10



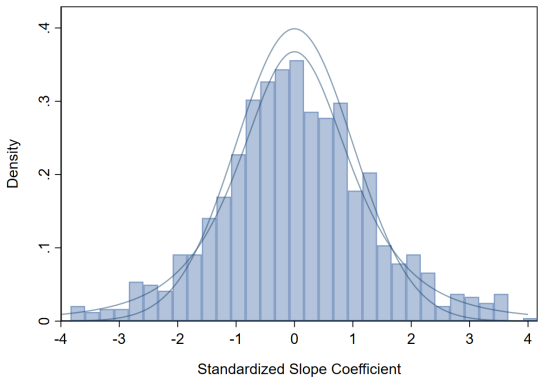
(d) N=100

Central Limit Theorem

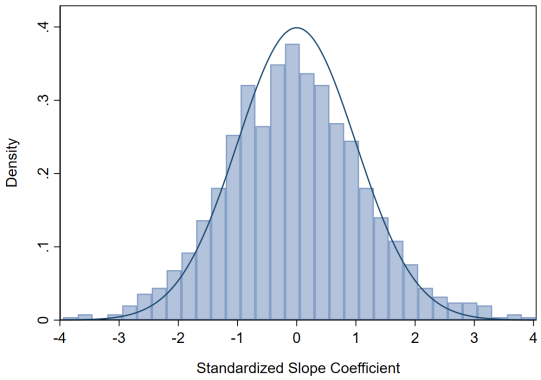
As N increases, the distribution of $\hat{\beta}_k$ becomes normal (starting from a t-distribution).



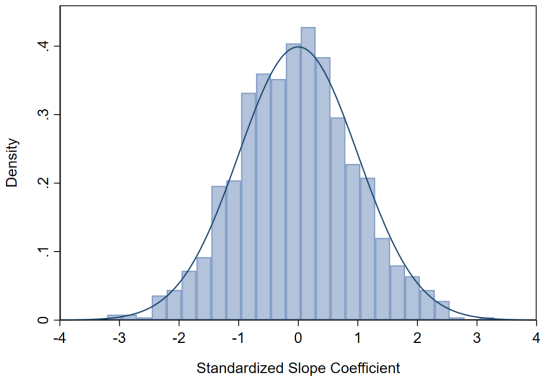
(a) N=3



(b) N=5



(c) N=10



(d) N=100

Consistency, Asymptotically Normality

Assuming LRM1, LRM2, LRM3d, LRM4a or LRM4b, and LRM5 the following properties can be established using law of large numbers and central limit theorem for large samples.

- The OLS estimator is **consistent**:

$$plim\hat{\beta}_k = \beta_k.$$

That is, for all $\varepsilon > 0$

$$\lim_{N\to\infty} \Pr\left(|\hat{\beta}_k - \beta_k| > \varepsilon\right) = 0.$$

- The OLS estimator is **asymptotically normally distributed**

$$\sqrt{N}(\hat{\beta}_k - \beta_k) \xrightarrow{d} N(0, Avar(\hat{\beta}_k) \times N)$$

(Avar means asymptotic variance)

- The OLS estimator is **approximately normally distributed**

$$\hat{\beta}_k \overset{A}{\sim} N\left(\beta_k, Avar(\hat{\beta}_k)\right)$$

Efficiency and Asymptotic Variance

For the bivariate regression under LRM4a (homoskedasticity) it can be **consistently** estimated as

$$\widehat{Avar}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{N}(x_{i1} - \bar{x})^2},$$

with

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{u}_i^2}{N-2}.$$

Under LRMb (heteroskedasticity), $Avar(\hat{\beta})$ can be **consistently** estimated as the *robust* or *Eicker-Huber-White* estimator.

The robust variance estimator is calculated as

$$\widehat{Avar}(\hat{\beta}_1) = \frac{\sum_{i=1}^{N} \hat{u}_i^2 (x_{i1} - \bar{x})^2}{\left[\sum_{i=1}^{N}(x_{i1} - \bar{x})^2\right]}.$$

Note: In practice we can almost never be sure that the errors are homoskedastic and should therefore always use robust standard errors.

Sketch of Proof for Asymptotic Properties

- The OLS estimator of $\hat{\beta}$ is consistent and asymptotic normal

  Estimator $\hat{\beta}$ can be written as: $\hat{\beta} = \left(\frac{1}{N}X'X\right)^{-1}\frac{1}{N}X'y = \beta + \left(\frac{1}{N}X'X\right)^{-1}\frac{1}{N}X'u = \beta + \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} x_i u_i\right)$

  We can use the law of large numbers to establish that : $\frac{1}{N}\sum_{i=1}^{N} x_i x_i' \xrightarrow{p} \mathrm{E}[x_i x_i'] = \frac{Q_{xx}}{N}$, $\frac{1}{N}\sum_{i=1}^{N} x_i u_i \xrightarrow{p} \mathrm{E}[x_i u_i] = 0$

  By Slutsky's theorem and continuous mapping theorem these results can be combined to establish consistency of estimator $\hat{\beta}$: $\hat{\beta} \xrightarrow{p} \beta + Q_{xx}^{-1} \cdot 0 = \beta$

  The central limit theorem tells us that: $\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i u_i \xrightarrow{d} \mathcal{N}(0, V)$, where $V = \mathrm{Var}[x_i u_i] = \mathrm{E}[\, u_i^2 x_i x_i'\,] = \mathrm{E}\left[\, \mathrm{E}[u_i^2|x_i]\, x_i x_i'\,\right] = \sigma^2 \frac{Q_{xx}}{N}$

Applying Slutsky's theorem again we'll have:

$$\sqrt{N}(\hat{\beta}-\beta)=\left(\frac{1}{N}\sum_{i=1}^{N}x_i x_i'\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}x_i u_i\right) \xrightarrow{d} Q_{xx}^{-1}N\cdot\mathcal{N}\big(0,\sigma^2\frac{Q_{xx}}{N}\big)$$
$$=\mathcal{N}\big(0,\sigma^2 Q_{xx}^{-1}N\big)$$

OLS Properties in the Small and in the Large

Set of assumptions	(1)	(2)	(3)	(4)	(5)	(6)
LRM1: linearity	f	u	l	f	i	l	l	e	d	
LRM2: simple random sampling	f	u	l	f	i	l	l	e	d	
LRM5: identifiability	f	u	l	f	i	l	l	e	d	
LRM4: error variance						
- LRM4a: homoskedastic	✓	✓	✓	×	×	×
- LRM4b: heteroskedastic	×	×	×	✓	✓	✓
LRM3: exogeneity						
- LRM3a: normality	✓	×	×	✓	×	×
- LRM3b: independent	✓	✓	×	×	×	×
- LRM3c: mean indep.	✓	✓	✓	✓	✓	×
- LRM3d: uncorrelated	✓	✓	✓	✓	✓	✓
Small sample properties of $\hat{\beta}$						
- unbiased	✓	✓	✓	✓	✓	×
- normally distributed	✓	×	×	✓	×	×
- efficient	✓	✓	✓	×	×	×
Large sample properties of $\hat{\beta}$						
- consistent	✓	✓	✓	✓	✓	✓
- approx. normal	✓	✓	✓	✓	✓	✓
- asymptotically efficient	✓	✓	✓	×	×	×

- *Notes:* ✓ = fulfilled, × = violated

Tests in Small Samples I

Assume LRM1, LRM2, LRM3a, LRM4a, and LRM5. A simple null hypotheses of the form $H_0 : \beta_k = q$ is tested with the **t-test**.

If the null hypotheses is true, the t-statistic

$$t = \frac{\hat{\beta}_k - q}{\widehat{se}(\hat{\beta}_k)} \sim t_{N-K-1}$$

follows a t-distribution with $N - K - 1$ degrees of freedom. The standard error is $\widehat{se}(\hat{\beta}_k) = \sqrt{\hat{V}(\hat{\beta}_k)}$.

For example, to perform a two-sided test of H_0 against the alternative hypotheses $H_A : \beta_k \neq q$ on the 5% significance level, we calculate the t-statistic and compare its absolute value to the 0.975-quantile of the t-distribution. With $N = 30$ and $K = 2$, H_0 is rejected if $|t| > 2.052$.

Tests in Small Samples II

A null hypotheses of the form $H_0 : r_{j1}\beta_1 + \ldots + r_{jK}\beta_K = q_j$, in matrix notation $H_0 : R\beta = q$, with J linear restrictions $j = 1 \ldots J$ is jointly tested with the **F-test**.

If the null hypotheses is true, the F-statistic follows an F distribution with J numerator degrees of freedom and $N - K - 1$ denominator degrees of freedom:

$$F = \frac{\left(R\hat{\beta} - q\right)' \left[R\hat{V}(\hat{\beta}|X)R'\right]^{-1} \left(R\hat{\beta} - q\right)}{J} \sim F_{J,N-K-1}.$$

For example, to perform a two-sided test of H_0 against the alternative hypotheses $H_A : r_{j1}\beta_1 + \ldots + r_{jK}\beta_K \neq q_j$ for all j at the 5% significance level, we calculate the F-statistic and compare it to the 0.95-quantile of the F-distribution.

With $N = 30, K = 2$ and $J = 2$, H_0 is rejected if $F > 3.35$. We cannot perform two-sided F-tests because the F distribution has one tail.

Tests in Small Samples III

Only under homoskedasticity (LRM4a), the F-statistic can also be computed as

$$F = \frac{(R^2 - R^2_{\text{restricted}})/J}{(1 - R^2)/(N - K - 1)} \sim F_{J,N-K-1},$$

where $R^2_{\text{restricted}}$ is estimated by restricted least squares which minimizes $SD(\beta)$ s.t. $r_{j1}\beta_1 + \ldots + r_{jK}\beta_K \neq q_j$ for all j.

Exclusionary restrictions of the form $H_0 : \beta_k = 0, \beta_m = 0, \ldots$ are a special case of $H_0 : r_{j1}\beta_1 + \ldots + r_{jK}\beta_K = q_j$ for all j. In this case, restricted least squares is simply estimated as a regression were the explanatory variables $k, m, \ldots$ are excluded, e.g. a regression with a constant only.

If the F distribution has degrees of freedom (df) 1 as the numerator df, and $N - K - 1$ as the denominator df, then it can be shown that $t^2 = F(1, N - K - 1)$.

Confidence Intervals in Small Samples

Assuming LRM1, LRM2, LRM3a, LRM4a, and LRM5, we can construct confidence intervals for a particular coefficient β_k. The $(1 - \alpha)$ confidence interval is given by

$$\left(\hat{\beta}_k - t_{(1-\alpha/2),(N-K-1)}\widehat{se}(\hat{\beta}_k), \hat{\beta}_k + t_{(1-\alpha/2),(N-K-1)}\widehat{se}(\hat{\beta}_k)\right),$$

where $t_{(1-\alpha/2),(N-K-1)}$ is the $(1 - \alpha/2)$ quantile of the t-distribution with $(N - K - 1)$ degrees of freedom. For example, the 95% confidence interval with $N = 30$ and $K = 2$ is $\left(\hat{\beta}_k - 2.052\widehat{se}(\hat{\beta}_k), \hat{\beta}_k + 2.052\widehat{se}(\hat{\beta}_k)\right)$.

Recall: α is the maximum acceptable probability of a Type I error.

Null hypothesis (H_0)	is valid (Innocent)	is invalid (Guilty)
Reject H_0	**Type I ($\alpha = 0.05$) error**	Correct outcome
I think he is guilty!	False positive	True positive
	Convicted!	Convicted!
Don't reject H_0	Correct outcome	**Type II (β) error**
I think he is innocent!	True negative	False negative
	Freed!	Freed!

Asymptotic Tests

Assume LRM1, LRM2, LRM3d, LRM4a or LRM4b, and LRM5. A simple null hypotheses of the form $H_0 : \beta_k = q$ is tested with the **z-test**. If the null hypotheses is true, the z-statistic

$$z = \frac{\hat{\beta}_k - q}{\widehat{se}(\hat{\beta}_k)} \overset{A}{\sim} N(0, 1)$$

follows approximately the standard normal distribution. The standard error is $\widehat{se}(\hat{\beta}_k) = \sqrt{\widehat{Avar}(\hat{\beta}_k)}$.

For example, to perform a two sided test of H_0 against the alternative hypotheses $H_A : \beta_k \neq q$ on the 5% significance level, we calculate the z-statistic and compare its absolute value to the 0.975-quantile of the standard normal distribution. H_0 is rejected if $|z| > 1.96$.

We talk about the Wald test later...

Confidence Intervals in Large Samples

Assuming LRM1, LRM2, LRM3d, LRM5, and LRM4a or LRM4b, we can construct confidence intervals for a particular coefficient β_k. The $(1-\alpha)$ confidence interval is given by

$$\left(\hat{\beta}_k - z_{(1-\alpha/2)}\widehat{se}(\hat{\beta}_k), \hat{\beta}_k + z_{(1-\alpha/2)}\widehat{se}(\hat{\beta}_k)\right)$$

where $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution.

For example, the 95% confidence interval is $\left(\hat{\beta}_k - 1.96\widehat{se}(\hat{\beta}_k), \hat{\beta}_k + 1.96\widehat{se}(\hat{\beta}_k)\right)$.

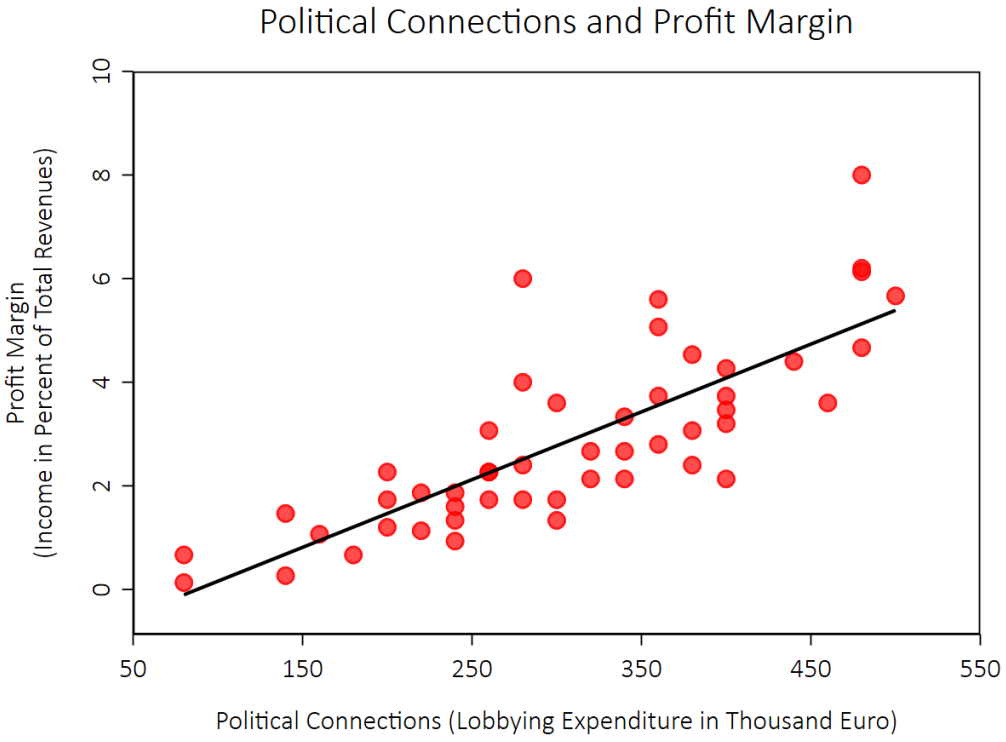
OLS Properties in the Small and in the Large

Set of assumptions	(1)	(2)	(3)	(4)	(5)	(6)
LRM1: linearity	f u l	f u l	f i l	f i l	e d	
LRM2: simple random sampling	f u l	f u l	f i l	f i l	e d	
LRM5: identifiability	f u l	f u l	f i l	f i l	e d	
LRM4: error variance						
- LRM4a: homoskedastic	✓	✓	✓	×	×	×
- LRM4b: heteroskedastic	×	×	×	✓	✓	✓
LRM3: exogeneity						
- LRM3a: normality	✓	×	×	✓	×	×
- LRM3b: independent	✓	✓	×	×	×	×
- LRM3c: mean indep.	✓	✓	✓	✓	✓	×
- LRM3d: uncorrelated	✓	✓	✓	✓	✓	✓
Small sample properties of $\hat{\beta}$						
- unbiased	✓	✓	✓	✓	✓	×
- normally distributed	✓	×	×	✓	×	×
- efficient	✓	✓	✓	×	×	×
t-test, F-test	✓	×	×	×	×	×
Large sample properties of $\hat{\beta}$						
- consistent	✓	✓	✓	✓	✓	✓
- approx. normal	✓	✓	✓	✓	✓	✓
- asymptotically efficient	✓	✓	✓	×	×	×
z-test, Wald test	✓	✓	✓	✓*	✓*	✓*

- *Notes:* ✓ = fulfilled, × = violated, * = corrected standard errors.

4.5 Politically Connected Firms: Causality or Correlation?

Arguments For Causality of Effect



Econometric methods need to address concerns, including:

- **Misspecification:** Results robust to different functional forms
- **Errors-in-variables:** little concern with administrative data
- **External validity:** Similar effect found in independent studies.

Arguments Against Causality of Effect

- **Omitted variable bias:**
  e.g., business acumen
  → Panel data models
- **Sample selection bias:**
  lobbying expenditures only observed if in transparency register.
  → Selection correction models
- **Simultaneous causality:**

- profits may be higher because of political connections
- firms may become connected because of their high profits

All of those concerns may be addressed with
→instrumental variable models. What would be a good instrument/experiment?

5 Simplifying Linear Regressions using Frisch-Waugh-Lovell

5.1 Frisch-Waugh-Lovell theorem in equation algebra

From the multivariate to the bivariate regression

Regress y_i on two explanatory variables, where x_i^2 is the variable of interest and x_i^1 (or further variables) are not of interest.

$$y_i = \beta_0 + \beta_2 x_i^2 + \beta_1 x_i^1 + \varepsilon_i.$$

Surprising and useful result:

- We can obtain **exactly the same** coefficients and residuals from a regression of two demeaned variables
$$\tilde{y}_i = \beta_0 + \beta_2 \tilde{x}_i^2 + \varepsilon_i.$$
- We can obtain **exactly the same** coefficient and residuals from a regression of two residualized variables
$$\varepsilon_i^y = \beta_2 \varepsilon_i^2 + \varepsilon_i.$$

Why is the decomposition useful?

Allows breaking a multivariate model with K independent variables into K bivariate models.

- Relationship between two variables from a multivariate model can be shown in a two-dimensional scatter plot
- Absorbs fixed effects to reduce computation time (see reghdfe for Stata)
- Allows to separate variability between the regressors (multicollinearity) and between the residualized variable $\tilde{x}_i^2$ and the dependent variable y_i.
- Understand biases in multivariate models tractably.

How to decompose y_i and x_i^2?

Partial out x_i^1 from y_i and from x_i^2.

- Regress x_i^2 on all x_i^1 and get residuals ε_i^2:
$$x_i^2 = \gamma_0 + \gamma_1 x_i^1 + \varepsilon_i^2,$$

this implies $Cov(x_i^1, \varepsilon_i^2) = 0$,

- Regress y_i on all x_i^1 and get residuals ε_i^y:

$$y_i = \delta_0 + \delta_1 x_i^1 + \varepsilon_i^y.$$

This implies $Cov(x_i^1, \varepsilon_i^y) = 0$.

From the residuals and the constants γ_0 and δ_0 generate

- $\tilde{x}_i^2 = \gamma_0 + \varepsilon_i^2$,
- $\tilde{y}_i = \delta_0 + \varepsilon_i^y$.

Finally,

$$\tilde{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{x}_i^2 + \tilde{\varepsilon}_i = \beta_0 + \beta_2 \tilde{x}_i^2 + \varepsilon_i.$$

Decomposition theorem

Decomposition theorem

For multivariate regressions and detrended regressions, e.g.,

$$y_i = \beta_0 + \beta_2 x_i^2 + \beta_1 x_i^1 + \varepsilon_i,$$

$$\tilde{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{x}_i^2 + \tilde{\varepsilon}_i,$$

the same regression coefficients will be obtained with any non-empty subset of the explanatory variables, such that

$$\tilde{\beta}_1 = \beta_2 \text{ and also } \tilde{\varepsilon}_i = \varepsilon_i.$$

Examining either set of residuals will convey precisely the same information about the properties of the unobservable stochastic disturbances.

Detrended variables

Show that

$$\begin{aligned} y_i &= \beta_0 + \beta_2 x_i^2 + \beta_1 x_i^1 + \varepsilon_i \\ &= \tilde{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{x}_i^2 + \tilde{\varepsilon}_i. \end{aligned} \tag{2}$$

Plug in the variables $y_i = \delta_0 + \delta_1 x_i^1 + \varepsilon_i^y$ and $x_i^2 = \gamma_0 + \gamma_1 x_i^1 + \varepsilon_i^2$ in the equation (2)

$$\begin{aligned} y_i &= \delta_0 + \delta_1 x_i^1 + \varepsilon_i^y = \beta_0 + \beta_2(\gamma_0 + \gamma_1 x_i^1 + \varepsilon_i^2) + \beta_1 x_i^1 + \varepsilon_i \\ \tilde{y}_i &= \delta_0 + \varepsilon_i^y = \beta_0 + \beta_2(\gamma_0 + \varepsilon_i^2) + (\beta_2\gamma_1 - \delta_1 + \beta_1) x_i^1 + \varepsilon_i. \end{aligned}$$

Because we partialled out x_i^1 using OLS, x_i^1 is mechanically uncorrelated to ε_i^2 and to ε_i^y. Therefore, the regression coefficient $(\beta_2\gamma_1 - \delta_1 + \beta_1)$ of the partialled out variable x_i^1 is zero. The equation simplifies with $\tilde{x}_i^2 = \gamma_0 + \varepsilon_i^2$ to

$$\tilde{y}_i = \delta_0 + \varepsilon_i^y = \beta_0 + \beta_2(\gamma_0 + \varepsilon_i^2) + \varepsilon_i.$$

Regression anatomy: Only detrending x_i^2 and not y_i. The regression constant, residuals, and the standard errors change but β_2 remains

$$\begin{aligned} y_i = \delta_0 + \delta_1 x_i^1 + \varepsilon_i^y &= (\beta_0 + \delta_1\bar{x}^1) + \beta_2(\gamma_0 + \varepsilon_i^2) + (\varepsilon_i + \delta_1 x_i^1) \\ y_i &= \kappa + \beta_2\tilde{x}^2 + \epsilon_i. \end{aligned}$$

Residualized variables

$$\begin{aligned} \tilde{y}_i = \delta_0 + \varepsilon_i^y &= \beta_0 + \beta_2(\gamma_0 + \varepsilon_i^2) + \varepsilon_i \\ \varepsilon_i^y &= \beta_0 - \delta_0 + \beta_2\gamma_0 + \beta_2\varepsilon_i^2 + \varepsilon_i. \end{aligned}$$

The same result of the FWL Theorem holds as well for a regression of the residualized variables because $\beta_0 = \delta_0 - \beta_2\gamma_0$:

$$\varepsilon_i^y = \beta_2\varepsilon_i^2 + \varepsilon_i.$$

5.2 Projection and residual maker matrices

Partition of $\boldsymbol{y}$

Least squares partitions the vector $\boldsymbol{y}$ into two orthogonal parts

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} = \boldsymbol{P}\boldsymbol{y} + \boldsymbol{M}\boldsymbol{y}.$$

- $n \times 1$ vector of data $\boldsymbol{y}$
- $n \times n$ projection matrix $\boldsymbol{P}$
- $n \times n$ residual maker matrix $\boldsymbol{M}$
- $n \times 1$ vector of residuals $\boldsymbol{e}$

Projection matrix

$$\boldsymbol{Py} = \boldsymbol{Xb} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$$

$$\rightarrow \boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}.$$

Projection matrix

Properties.

- symmetric such that $\boldsymbol{P} = \boldsymbol{P'}$, thus orthogonal
- idempotent such that $\boldsymbol{P} = \boldsymbol{P}^2$, thus indeed a projection
- annihilator matrix $\boldsymbol{PX} = \boldsymbol{X}$

Example for projection matrix

Example

Show $\boldsymbol{PX} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'X} = \boldsymbol{X}$.

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}; \mathbf{X'X} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}; \mathbf{X'X}^{-1} = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1.5 \end{bmatrix};$$

$$\boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 3/2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \end{bmatrix}$$

$$\boldsymbol{PX} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}.$$

Project $\boldsymbol{y}$ on the column space of $\boldsymbol{X}$, i.e. regress $\boldsymbol{y}$ on $\boldsymbol{x}$ and predict $E[\boldsymbol{y}] = \hat{\boldsymbol{y}}$.

$$\boldsymbol{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}; \boldsymbol{Py} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \hat{\boldsymbol{y}} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}.$$

Residual maker matrix

$$\begin{aligned} \boldsymbol{My} &= \boldsymbol{e} = \boldsymbol{y} - \boldsymbol{Xb} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \\ \boldsymbol{My} &= (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'})\boldsymbol{y} \end{aligned}$$

$$\rightarrow \boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'} = (\boldsymbol{I} - \boldsymbol{P}).$$

Residual maker matrix

Properties.

- symmetric such that $\boldsymbol{M} = \boldsymbol{M'}$
- idempotent such that $\boldsymbol{M} = \boldsymbol{M}^2$
- annihilator matrix $\boldsymbol{MX} = 0$
- orthogonal to $\boldsymbol{P}$: $\boldsymbol{PM} = \boldsymbol{MP} = \boldsymbol{0}$.

Example for residual maker matrix

Example

Show $\boldsymbol{MX} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'})\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = \boldsymbol{0}$.

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}; \mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix};$$

$$\boldsymbol{M} = (\boldsymbol{I} - \boldsymbol{P}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & -1/2 \\ 0 & 0 & 0 \\ -1/2 & 0 & 1/2 \end{bmatrix}$$

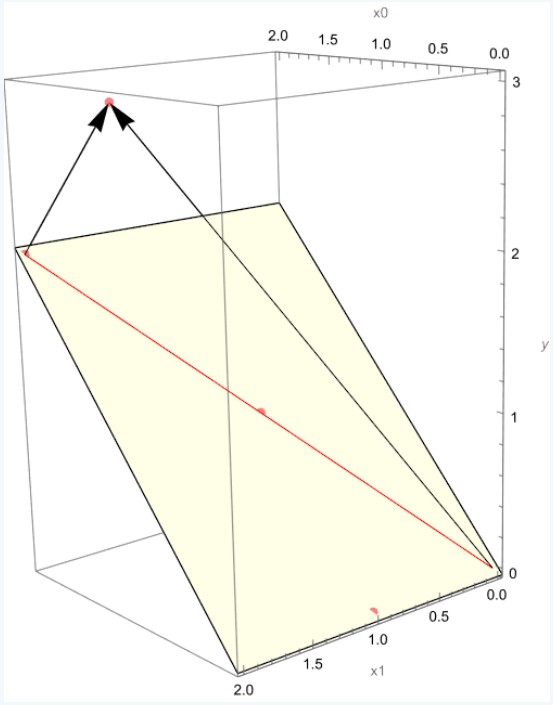
$$\boldsymbol{MX} = \begin{bmatrix} 1/2 & 0 & -1/2 \\ 0 & 0 & 0 \\ -1/2 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Obtain residuals from a projection of $\boldsymbol{y}$ on the column space of $\boldsymbol{X}$, i.e. regress $\boldsymbol{y}$ on $\boldsymbol{x}$ and predict $\boldsymbol{y} - E[\boldsymbol{y}] = \boldsymbol{y} - \hat{\boldsymbol{y}}$.

$$\boldsymbol{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}; \boldsymbol{My} = \begin{bmatrix} 1/2 & 0 & -1/2 \\ 0 & 0 & 0 \\ -1/2 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}.$$

Column space of $\mathbf{X}$ is $\mathbf{x}_0$ and $\mathbf{x}_1$.

$$\begin{bmatrix} x_0^1 = 1 & x_1^1 = 0 & y^1 = 1 \\ x_0^2 = 1 & x_1^2 = 1 & y^2 = 2 \\ x_0^3 = 1 & x_1^3 = 0 & y^3 = 3 \end{bmatrix}; \hat{\boldsymbol{y}} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}; \boldsymbol{y} - \hat{\boldsymbol{y}} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}.$$



The closest point from the vector $\boldsymbol{y}' = [1, 2, 3]$ onto the column space of $\boldsymbol{X}$, is $\hat{\boldsymbol{y}} = \boldsymbol{Xb}$, here $\hat{\boldsymbol{y}}' = [2, 2, 2]$. At this point, we can draw a line orthogonal to the column space of $\boldsymbol{X}$.

Decomposing the normal equations

The normal equations[1] in matrix form are $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{y}$. If $\boldsymbol{X}$ is partitioned into an interesting segment $\boldsymbol{X}_2$ and an uninteresting $\boldsymbol{X}_1$, normal equations are

$$\begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{y} \\ \boldsymbol{X}_2'\boldsymbol{y} \end{bmatrix}.$$

The multiplication of the two equations can be done separately

$$\begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{y} \end{bmatrix} \tag{3}$$

$$\begin{bmatrix} \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_2'\boldsymbol{y} \end{bmatrix}. \tag{4}$$

How can we find an expression for $\boldsymbol{b}_2$ that does not involve $\boldsymbol{b}_1$?

Solving for $\boldsymbol{b}_2$

Idea: Solve equation (3) for $\boldsymbol{b}_1$ in terms of $\boldsymbol{b}_2$, then substituting that solution into the equation (4).

$$\begin{aligned} \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix} &= \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{y} \end{bmatrix} \\ \boldsymbol{X}_1'\boldsymbol{X}_1\boldsymbol{b}_1 + \boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{b}_2 &= \boldsymbol{X}_1'\boldsymbol{y} \\ \boldsymbol{X}_1'\boldsymbol{X}_1\boldsymbol{b}_1 &= \boldsymbol{X}_1'\boldsymbol{y} - \boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{b}_2 \\ \boldsymbol{b}_1 &= (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y} - (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{b}_2 \\ &= (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'(\boldsymbol{y} - \boldsymbol{X}_2\boldsymbol{b}_2) \end{aligned}$$

Multiplying out equation (4) gives

$$\begin{aligned} \begin{bmatrix} \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix} &= \begin{bmatrix} \boldsymbol{X}_2'\boldsymbol{y} \end{bmatrix} \\ \boldsymbol{X}_2'\boldsymbol{X}_1\boldsymbol{b}_1 + \boldsymbol{X}_2'\boldsymbol{X}_2\boldsymbol{b}_2 &= \boldsymbol{X}_2'\boldsymbol{y} \end{aligned}$$

Plugging in the solution for $\boldsymbol{b}_1$ gives

$$\boldsymbol{X}_2'\boldsymbol{X}_1\Big((\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'(\boldsymbol{y} - \boldsymbol{X}_2\boldsymbol{b}_2)\Big) + \boldsymbol{X}_2'\boldsymbol{X}_2\boldsymbol{b}_2 = \boldsymbol{X}_2'\boldsymbol{y}.$$

[1] It is called a normal equation because $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}$ is normal to the range of $\boldsymbol{X}$.

$$\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'(\boldsymbol{y}-\boldsymbol{X}_2\boldsymbol{b}_2)+\boldsymbol{X}_2'\boldsymbol{X}_2\boldsymbol{b}_2=\boldsymbol{X}_2'\boldsymbol{y}.$$

The middle part of the first term is $\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$. This is the projection matrix $\boldsymbol{P}_{X_1}$ from a regression of $\boldsymbol{y}$ on $\boldsymbol{X}_1$.

$$\boldsymbol{X}_2'\boldsymbol{P}_{X_1}\boldsymbol{y}-\boldsymbol{X}_2'\boldsymbol{P}_{X_1}\boldsymbol{X}_2\boldsymbol{b}_2+\boldsymbol{X}_2'\boldsymbol{X}_2\boldsymbol{b}_2=\boldsymbol{X}_2'\boldsymbol{y}.$$

We can multiply by an identity matrix $\boldsymbol{I}$ without changing anything

$$\begin{aligned}
&\boldsymbol{X}_2'\boldsymbol{P}_{X_1}\boldsymbol{y}-\boldsymbol{X}_2'\boldsymbol{P}_{X_1}\boldsymbol{X}_2\boldsymbol{b}_2+\boldsymbol{X}_2'\boldsymbol{I}\boldsymbol{X}_2\boldsymbol{b}_2=\boldsymbol{X}_2'\boldsymbol{I}\boldsymbol{y}.\\
&\boldsymbol{X}_2'\boldsymbol{I}\boldsymbol{y}-\boldsymbol{X}_2'\boldsymbol{P}_{X_1}\boldsymbol{y}=\boldsymbol{X}_2'\boldsymbol{I}\boldsymbol{X}_2\boldsymbol{b}_2-\boldsymbol{X}_2'\boldsymbol{P}_{X_1}\boldsymbol{X}_2\boldsymbol{b}_2.\\
&\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{P}_{X_1})\boldsymbol{y}=\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{P}_{X_1})\boldsymbol{X}_2\boldsymbol{b}_2.
\end{aligned}$$

Now $(\boldsymbol{I}-\boldsymbol{P}_{X_1})$ is the residual maker matrix $\boldsymbol{M}_{X_1}$

$$\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{y}=\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{X}_2\boldsymbol{b}_2.$$

Solving for $\boldsymbol{b}_2$ gives

$$\boldsymbol{b}_2 = (\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{y}.$$

$$\boldsymbol{b}_2 = (\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{y}.$$

The residualizer matrix is symmetric and idempotent, such that $\boldsymbol{M}_{X_1}=\boldsymbol{M}_{X_1}\boldsymbol{M}_{X_1}=\boldsymbol{M}_{X_1}'\boldsymbol{M}_{X_1}$.

$$\begin{aligned}
\boldsymbol{b}_2 &= (\boldsymbol{X}_2'\boldsymbol{M}_{X_1}'\boldsymbol{M}_{X_1}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_{X_1}'\boldsymbol{M}_{X_1}\boldsymbol{y}\\
&= \Big((\boldsymbol{M}_{X_1}\boldsymbol{X}_2)'(\boldsymbol{M}_{X_1}\boldsymbol{X}_2)\Big)^{-1}(\boldsymbol{M}_{X_1}\boldsymbol{X}_2)'(\boldsymbol{M}_{X_1}\boldsymbol{y})\\
&= (\tilde{\boldsymbol{X}}_2'\tilde{\boldsymbol{X}}_2)^{-1}\tilde{\boldsymbol{X}}_2'\tilde{\boldsymbol{y}}.
\end{aligned}$$

This is the OLS solution for $\boldsymbol{b}_2$, with $\tilde{\boldsymbol{X}}_2$ instead of $\boldsymbol{X}$ and $\tilde{\boldsymbol{y}}$ instead of $\boldsymbol{y}$.

- $\tilde{\boldsymbol{X}}_2$ are residuals from a regression of $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$
- $\tilde{\boldsymbol{y}}$ are residuals from a regression of $\boldsymbol{y}$ on $\boldsymbol{X}_1$

The solution of the regression coefficients $\boldsymbol{b}_2$ in a regression that includes other regressors $\boldsymbol{X}_1$ is the same as first regressing all of $\boldsymbol{X}_2$ and $\boldsymbol{y}$ on $\boldsymbol{X}_1$, then regressing the residuals

from the $\boldsymbol{y}$ regression on the residuals from the $\boldsymbol{X}_2$ regression.

6 The Maximum Likelihood Estimator

6.1 From Probability to Likelihood

The Likelihood Principle

Suppose you have three credit cards. You forgot, which has money on it or not. Thus, the number credit cards with money, call it θ, might be 0, 1, 2, or 3. You can try your cards 4 times at random to check if you can make a payment.

The checks are random variables y_1, y_2, y_3, and y_4. They are

$$y_i = \begin{cases} 1, & \text{if the } i\text{th card has money on it,} \\ 0, & \text{otherwise.} \end{cases}$$

Since you chose y_i's uniformly, they are i.i.d. and $y_i \sim Bernoulli(\theta/3)$. After checking, we find $y_1 = 1, y_2 = 0, y_3 = 1, y_4 = 1$. We observe 3 cards with money and 1 without.

The number credit cards with money could still be 0, 1, 2, or 3.
Which is most likely?

From Probability to Likelihood

You could test for the true θ_0 in many samples. Conversely, you can check each possible value of θ to find the probability of observing the sample $(y_1 = 1, y_2 = 0, y_3 = 1, y_4 = 1)$.

Since $y_i \sim Bernoulli(\theta/3)$, we have

$$Prob(y_i = y) = \begin{cases} \theta/3, & \text{for } y = 1, \\ 1 - \theta/3, & \text{for } y = 0. \end{cases}$$

Since y_i's are independent, the joint PMF of y_1, y_2, y_3, and y_4 can be written as

$$Prob(y_1 = y, y_2 = y, y_3 = y, y_4 = y|\theta) =$$
$$Prob(y_1)Prob(y_2)Prob(y_3)Prob(y_4).$$

This depends on θ, and is called **likelihood function**:

$$L(\theta|y_i) = Prob(y_1 = 1, y_2 = 0, y_3 = 1, y_4 = 1, \theta) =$$
$$\theta/3(1 - \theta/3)\theta/3\theta/3 = (\theta/3)^3(1 - \theta/3).$$

Trial	1	2	3	4
θ	0	1	2	3
$Prob(\cdot)$	0.0000	0.0247	0.0988	0.0000

Values of the Likelihood $L(\theta|y_i)$ for different θ

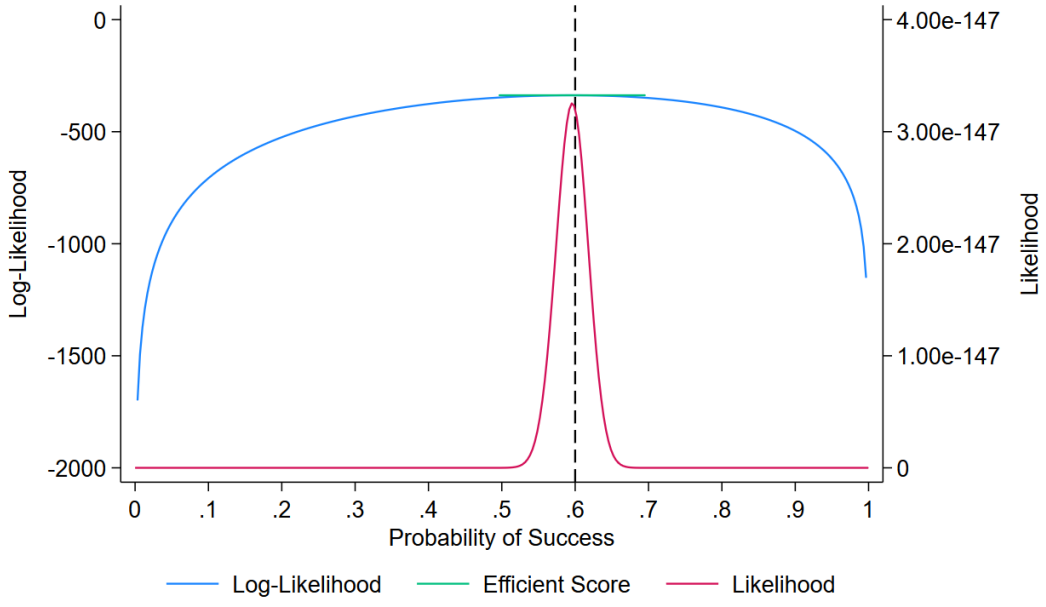
The probability of the observed sample for $\theta = 0$ and $\theta = 3$ is zero. This makes sense because our sample included both cards with and without money. The observed data is most likely to occur for $\theta = 2$.

Likelihood principle: choose θ that maximizes the likelihood of observing the actual sample to get an estimator for θ_0.

The likelihood is the probability from

- probability mass function if discrete
- probability distribution function if continuous

From Likelihood to Log-Likelihood



- The **likelihood function** $L_N(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})$ is the joint probability mass function or density $f(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{\theta})$, viewed as a function of vector $\boldsymbol{\theta}$ given the data $(\boldsymbol{y}, \boldsymbol{X})$.
- Maximizing $L_N(\boldsymbol{\theta})$ is equivalent to maximizing the **log-likelihood function** $\mathcal{L}_N(\boldsymbol{\theta}) = \ln L_N(\boldsymbol{\theta})$. Because taking the logarithm is a monotonic transformation. A maximum

Model	Range of y	Density $f(y)$	Common Parametrization
Bernoulli	0 or 1	$p^y(1-p)^{1-y}$	$p = \frac{e^{-\mathbf{x}'\beta}}{1+e^{-\mathbf{x}'\beta}}$
Poisson	$0, 1, 2, \ldots$	$e^{-\lambda}\lambda^y/y!$	$\lambda = e^{\mathbf{x}'\beta}$
Exponential	$(0, \infty)$	$\lambda e^{-\lambda y}$	$\lambda = e^{\mathbf{x}'\beta}$ or $1/\lambda = e^{\mathbf{x}'\beta}$
Normal	$(-\infty, \infty)$	$(2\pi\sigma^2)^{-1/2}e^{-(y-\mu)^2/2\sigma^2}$	$\mu = \boldsymbol{x}'\beta, \sigma^2 = \sigma^2$

for $L_N(\boldsymbol{\theta})$ corresponds with a maximum for $\mathcal{L}_N(\boldsymbol{\theta})$.

6.2 The Econometric Model

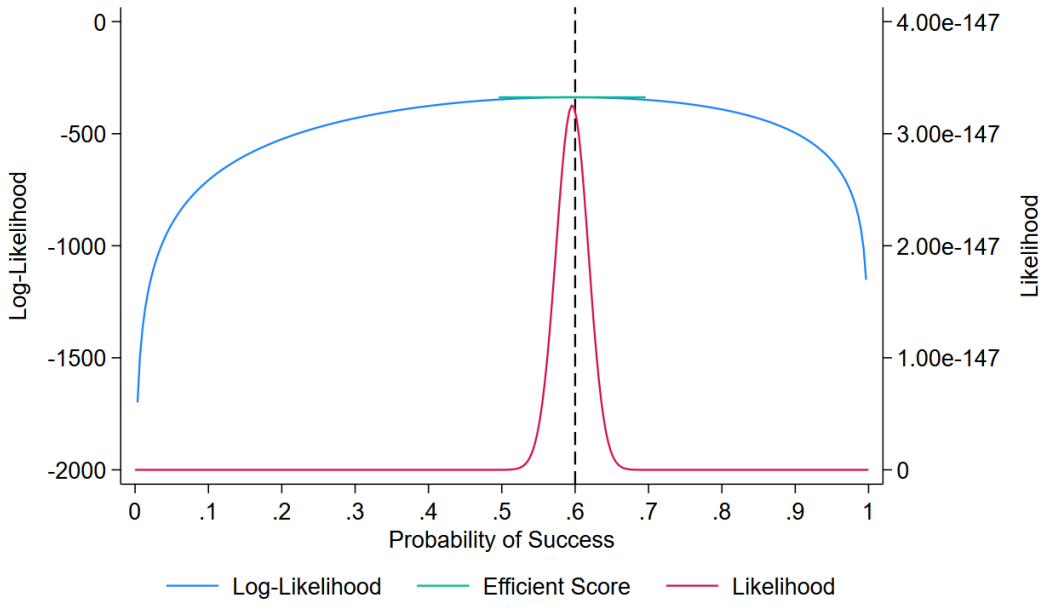
Specification of a Likelihood Function

The conditional likelihood $L_N(\boldsymbol{\theta}) = f(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{\theta})/f(\boldsymbol{X}|\boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})$ does not require the specification of the marginal distribution of $\boldsymbol{X}$.

For observations (y_i, x_i) independent over i and distributed with $f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})$,

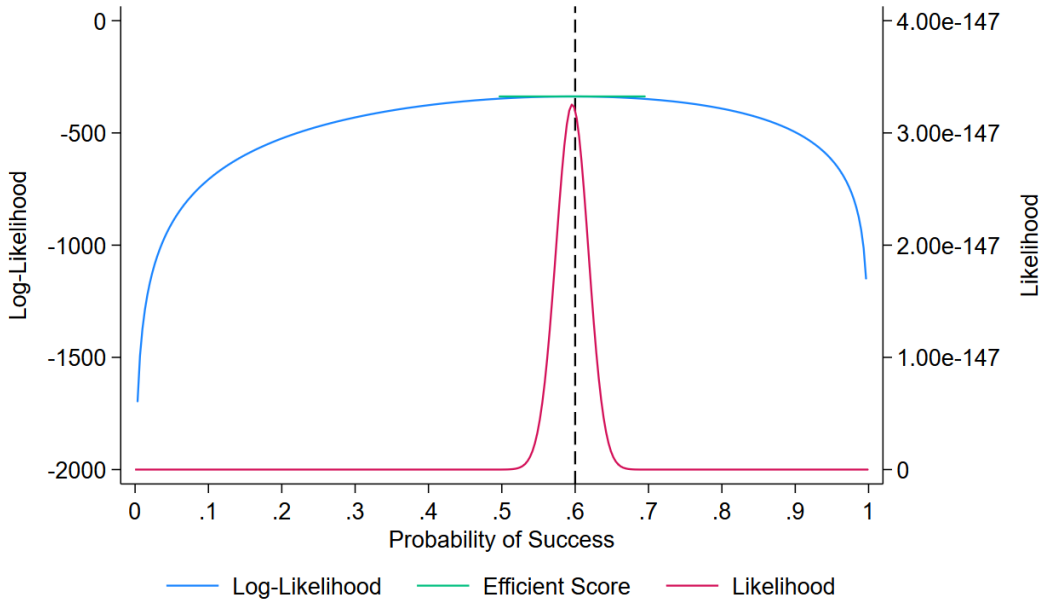
- the joint density is
$$f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = \Pi_{i=1}^N f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}),$$
- the log-likelihood function divided by N is
$$\frac{1}{N}\mathcal{L}_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^N \ln f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}).$$

Maximum Likelihood Estimator



The **maximum likelihood estimator** (MLE) is the estimator that maximizes the (conditional) log-likelihood function $\mathcal{L}_N(\theta)$.
The MLE is the local maximum that solves the first-order conditions

$$\frac{1}{N}\frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \ln f(y_i|\boldsymbol{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$



This estimator is an extremum estimator on based on the conditional density of y given $\boldsymbol{x}$. The gradient vector $\frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is called the **score vector**, as it sums the first derivatives of the log density, and when evaluated at $\boldsymbol{\theta}_0$ it is called the **efficient score**.

How Were the Data Generated?

Simple Random Sampling

$\{x_{i1},\ldots,x_{iK},y_i\}_{i=1}^{N}$ i.i.d. (independent and identically distributed)

This assumption means that

- observation i has no information content for observation $j \neq i$
- all observations i come from the same distribution

This assumption is guaranteed by simple random sampling provided there is no systematic non-response or truncation.

I.i.d. data simplify the maximization as the joint density of the two variables is simply the product of the two marginal densities.

For example with a normal joint pdf with two observations

$$f\left(y_1, y_2\right) = f_{Y_1}\left(y_1\right) f_{Y_2}\left(y_2\right) = \frac{1}{2\pi\sigma^2} e^{-\frac{\left[(y_1-\mu)^2+(y_2-\mu)^2\right]}{2\sigma^2}}.$$

With dependent observations we would have to maximize the following likelihood function, where ρ is the correlation:

$$\frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} e^{-\frac{\left[(y_1-\mu)^2+(y_2-\mu)^2-(y_1-\mu)(y_2-\mu)\right]}{2\sigma^2(1-\rho^2)}}.$$

The Score has Expected Value Zero

Likelihood Equation:

$$E_f\left[\boldsymbol{g}(\boldsymbol{\theta})\right] = E_f\left[\frac{\partial \ln f(y|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \int \frac{\partial \ln f(y|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(y|\boldsymbol{x},\boldsymbol{\theta}) dy = \mathbf{0}.$$

Example

$$\int f(y|\boldsymbol{\theta})dy = 1. \quad \frac{\partial}{\partial \boldsymbol{\theta}} \int f(y|\boldsymbol{\theta})dy = 0.$$

$$\int \frac{\partial f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} dy = 0.$$

$$\partial \ln f(y|\boldsymbol{\theta})/\partial \boldsymbol{\theta} = [\partial f(y|\boldsymbol{\theta})/\partial \boldsymbol{\theta}]/[f(y|\boldsymbol{\theta})]$$

$$\frac{\partial f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial \ln f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(y|\boldsymbol{\theta}).$$

$$\int \frac{\partial \ln f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(y|\boldsymbol{\theta}) dy = 0.$$

Fisher Information

The information matrix is the expectation of the outer product of the score vector,

$$\mathcal{I} = E_f\left[\frac{\partial \ln f(y|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f(y|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right].$$

The Fisher information $\mathcal{I}$ is equals the variance of the score, since $\frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ has mean zero.

- Large values of $\mathcal{I}$ mean that small changes in $\boldsymbol{\theta}$ lead to large changes in the log-likelihood
  $\rightarrow \mathcal{L}_N(\boldsymbol{\theta})$ contains considerable information about $\boldsymbol{\theta}$,
- Small values of $\mathcal{I}$ mean that the maximum is shallow and there are many nearby values of $\boldsymbol{\theta}$ with a similar log-likelihood.

Information Matrix Equality

The Fisher information $\mathcal{I}$ is equals the expectation of the Hessian $\boldsymbol{H}$:

$$-E_f\left[\boldsymbol{H}(\boldsymbol{\theta})\right] = -E_f\left[\frac{\partial^2 \ln f(y|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right] = E_f\left[\frac{\partial \ln f(y|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial \ln f(y|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right].$$

Example

For vector moment function, e.g., $\boldsymbol{m}(y,\boldsymbol{\theta}) = \dfrac{\partial \ln f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ with $E[\boldsymbol{m}(y,\boldsymbol{\theta})] = 0$,

$$\int \boldsymbol{m}(y,\boldsymbol{\theta}) f(y|\boldsymbol{\theta}) dy = 0.$$

$$\int \left(\frac{\partial \boldsymbol{m}(y,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} f(y|\boldsymbol{\theta}) + \boldsymbol{m}(y,\boldsymbol{\theta})\frac{\partial f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right) dy = 0.$$

$$\int \left(\frac{\partial \boldsymbol{m}(y,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} f(y|\boldsymbol{\theta}) + \boldsymbol{m}(y,\boldsymbol{\theta})\frac{\partial \ln f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} f(y|\boldsymbol{\theta})\right) dy = 0.$$

$$E\left[\frac{\partial \boldsymbol{m}(y,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right] = -E\left[\boldsymbol{m}(y,\boldsymbol{\theta})\frac{\partial \ln f(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right] = 0.$$

The Information Matrix in Practice

The variance of the sum of random score vector is:

Information matrix equality:

$$\text{Var}\left[\sum_{i=1}^{n} \boldsymbol{g}_i(\boldsymbol{\theta})\right] = \text{Var}\left[\boldsymbol{g}(\boldsymbol{\theta})\right] = -E_f\left[\boldsymbol{H}(\boldsymbol{\theta})\right] = -E\left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right].$$

After taking the expected value, $\widehat{\boldsymbol{\theta}}$ is substituted for $\boldsymbol{\theta}$. Problem: Taking the expected value of the second derivative matrix is frequently infeasible.

There exist two alternatives which are asymptotically equivalent:

- Ignore the expected value operator:

$$\widehat{\boldsymbol{I}}(\widehat{\boldsymbol{\theta}}) = -\frac{\partial^2 \ln L}{\partial \widehat{\boldsymbol{\theta}} \partial \widehat{\boldsymbol{\theta}}'}.$$

- Berndt-Hall-Hall-Hausman (BHHH) algorithm
  Never take a second derivative and sum over the outer product of the scores: (first derivatives per observation):

$$\check{\boldsymbol{I}}(\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \widehat{\boldsymbol{g}}_i \widehat{\boldsymbol{g}}_i' = \sum_{i=1}^{n} \left(\frac{\partial \ln f\left(y_i, \widehat{\boldsymbol{\theta}}\right)}{\partial \widehat{\boldsymbol{\theta}}} \right) \left(\frac{\partial \ln f\left(y_i, \widehat{\boldsymbol{\theta}}\right)}{\partial \widehat{\boldsymbol{\theta}}} \right)'.$$

6.3 Properties of the Maximum Likelihood Estimator

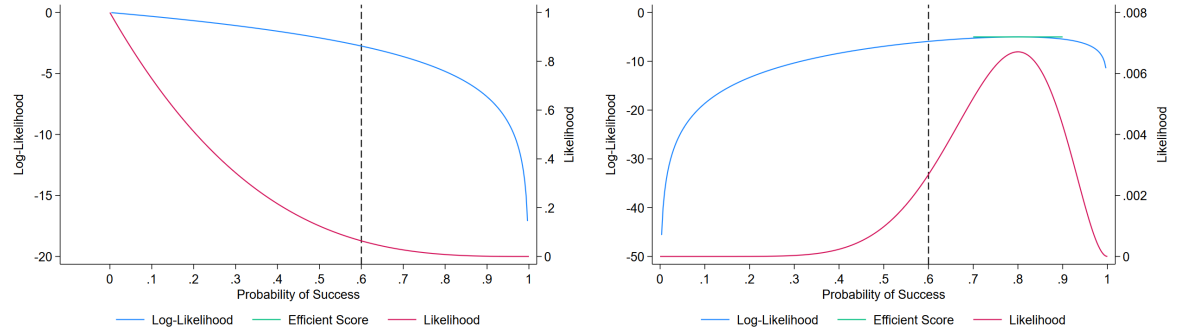
Properties of the MLE

- *Small sample properties of* $\hat{\boldsymbol{\theta}}$
  - may be biased
  - may have unknown distribution
  - variance may be biased, even towards zero
- *Large sample properties of* $\hat{\boldsymbol{\theta}}$
  - consistent
  - approx. normal
  - asymptotically efficient
  - invariant

Consistency

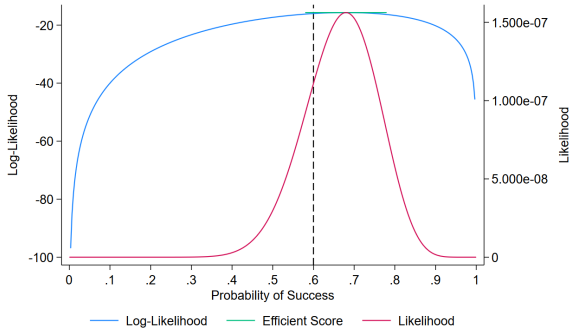
Law of Large Numbers

As N increases, the distribution of $\hat{\boldsymbol{\theta}}$ becomes more tightly centered around $\boldsymbol{\theta}$.

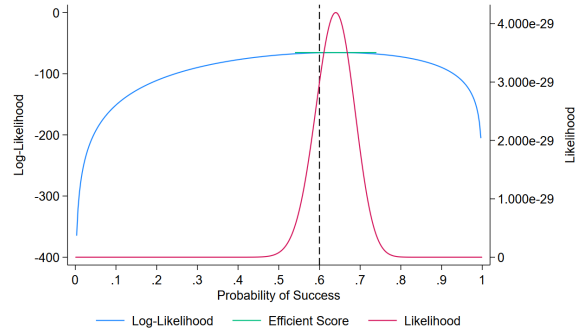


(a) N=3

(b) N=10



(c) N=25



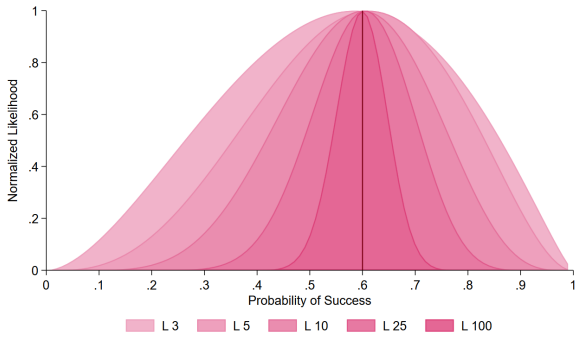
(d) N=100

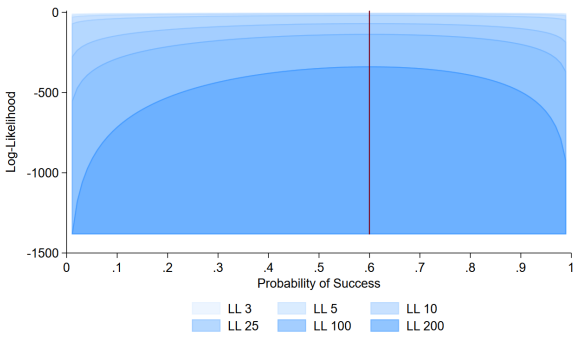
Likelihood Inequality

$$E[(1/N)\ln L(\hat{\boldsymbol{\theta}})] \geq E[(1/N)\ln L(\boldsymbol{\theta})].$$

The expected value of the log-likelihood is maximized at the true value of the parameters.

Figure 15: $\hat{\theta}$, Likelihood and Log-Likelihood as $n \to \infty$. True $\theta = 0.6$.





$$\lim_{n\to\infty} P(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}| > \epsilon) = \mathbf{0}. \qquad \lim_{n\to\infty} E[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}.$$

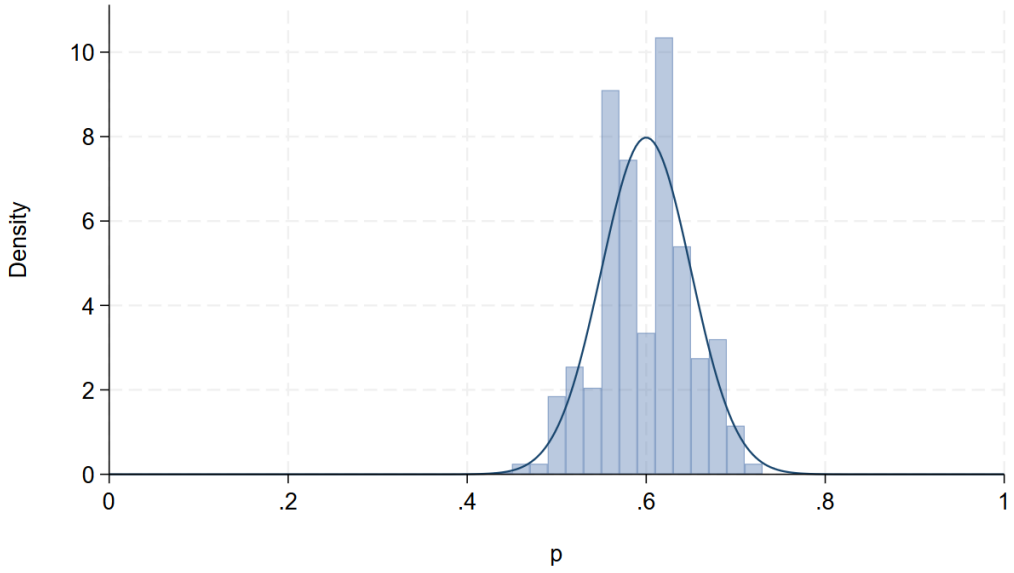
Approximate Normality

Central Limit Theorem

As N becomes large,

$$\hat{\boldsymbol{\theta}} \overset{\text{a}}{\sim} N\left[\boldsymbol{\theta}, -\left(E\left[\frac{\partial^2 \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]\right)^{-1}\right].$$

Figure 16: Sampling distribution of $\hat{\boldsymbol{\theta}}$ drawn from Bernoulli distribution and normal distribution at $N = 100$. True $\boldsymbol{\theta} = 0.6$.



Efficiency

The precision of the estimate $\hat{\boldsymbol{\theta}}$ is limited by the Fisher information $\mathcal{I}$ of the likelihood.

$$\text{Var}\left(\hat{\boldsymbol{\theta}}\right) \geq \frac{1}{\mathcal{I}\left(\boldsymbol{\theta}\right)}.$$

For large samples, this is the so-called Cramér-Rao lower bound for the variance matrix of consistent asymptotically normal estimators with convergence to normality of $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ uniform in compact intervals of $\boldsymbol{\theta}_0$.

Under the strong assumption of correct specification of the conditional density, the MLE has the **smallest asymptotic variance** among root-N consistent estimators.

Example

Since the MLE is unbiased,

$$\mathrm{E}\left[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \;\middle|\; \boldsymbol{\theta}\right] = \int \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) f(y;\boldsymbol{\theta})\, dy = 0 \text{ regardless of the value of } \boldsymbol{\theta}.$$

This expression is zero independent of $\boldsymbol{\theta}$, so its partial derivative with respect to $\boldsymbol{\theta}$ must also be zero. By the product rule, this partial derivative is also equal to

$$0 = \frac{\partial}{\partial \boldsymbol{\theta}} \int \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) f(y;\boldsymbol{\theta})\, dy = \int \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \frac{\partial f}{\partial \boldsymbol{\theta}}\, dy - \int f\, dy.$$

For each $\boldsymbol{\theta}$, the likelihood function is a probability density function, and therefore $\int f\, dy = 1$. By using the chain rule on the partial derivative of $\ln f$ and then dividing and multiplying by $f(y;\boldsymbol{\theta})$, one can verify that

$$\frac{\partial f}{\partial \boldsymbol{\theta}} = f \frac{\partial \ln f}{\partial \boldsymbol{\theta}}.$$

Using these two facts, we get

$$\int \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) f \frac{\partial \ln f}{\partial \boldsymbol{\theta}}\, dy = 1.$$

Factoring the integrand gives $\int \left(\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\sqrt{f}\right)\left(\sqrt{f}\,\frac{\partial \ln f}{\partial \boldsymbol{\theta}}\right)\, dy = 1.$
Squaring the expression in the integral, the Cauchy-Schwarz inequality yields

$$1 = \left(\int \left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\sqrt{f}\right] \cdot \left[\sqrt{f}\,\frac{\partial \ln f}{\partial \boldsymbol{\theta}}\right]\, dy\right)^2 \leq \left[\int \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^2 f\, dy\right] \cdot \left[\int \left(\frac{\partial \ln f}{\partial \boldsymbol{\theta}}\right)^2 f\, dy\right].$$

The first factor is the expected mean-squared error (the variance) of the estimator $\hat{\boldsymbol{\theta}}$, the second factor is the Fisher Information.

Invariance

The MLE of $\boldsymbol{\gamma} = \boldsymbol{c}(\boldsymbol{\theta})$ is $\widehat{\boldsymbol{\theta}} = \boldsymbol{c}(\widehat{\boldsymbol{\theta}})$ if $\boldsymbol{c}(\boldsymbol{\theta})$ is a continuous and continuous differentiable function.

- This simplifies the log-likelihood,
- This allows a function of $\widehat{\boldsymbol{\theta}}$ to serve as MLE if it is desired to analyze the function of an MLE.

Example

Suppose that the normal log-likelihood is parameterized in terms of the precision parameter, $\theta^2 = 1/\sigma^2$. The log-likelihood becomes

$$\ln L(\mu, \sigma^2) = -(N/2)\ln(2\pi) + (N/2)\ln\theta^2 - \frac{\theta^2}{2}\sum_{i=1}^{N}(y_i - \mu)^2.$$

The MLE for μ is $\bar{x}$. But the likelihood equation for θ^2 is now

$$\frac{\partial \ln L(\mu, \theta^2)}{\partial \theta^2} = 1/2\left[N/\theta^2 - \sum_{i=1}^{N}(y_i - \mu)^2\right] = 0,$$

which has solution $\hat{\theta}^2 = N/\sum_{i=1}^{N}(y_i - \mu)^2 = 1/\hat{\sigma}^2$.

The MLE is also equivariant with respect to certain **transformations of the data**.

If $y = c(x)$ where c is one to one and does not depend on the parameters to be estimated, then the density functions satisfy

$$f_Y(y) = \frac{f_X(x)}{|c'(x)|},$$

and hence the likelihood functions for x and y differ only by a factor that does not depend on the model parameters.

Example

The MLE parameters of the log-normal distribution are the same as those of the normal distribution fitted to the logarithm of the data.

7 The Generalized Method of Moments

7.1 How to choose from too many restrictions?

Minimize the quadratic form

The overidentified GMM estimator $\hat{\theta}_{GMM}(W_n)$ for K parameters in θ identified by $L > K$ moment conditions is a function of the weighting matrix W_n for a sample of $i = 1, ..., n$ observations:

$$\hat{\theta}_{GMM}(W_n) = \underset{\theta}{\text{argmin}} q_n(\theta),$$

where the quadratic form $q_n(\theta)$ is the criterion function and is given as a function of the sample moments $\bar{m}_n(\theta)$

$$q_n(\theta) = \bar{m}_n(\theta)' W \bar{m}_n(\theta).$$

The sample moments are a function

$$\bar{m}_n(\theta) = 1/n \sum_{i=1}^{N} m(X_i, Z_i, \theta_0)$$

of the model variables X_i, the instruments Z_i, and the true parameters θ_0.

What are the properties of the quadratic form

$$\underset{1\times 1}{q_n(\theta)} = \underset{1\times L}{\bar{m}_n(\theta)'} \underset{L\times L}{W} \underset{L\times 1}{\bar{m}_n(\theta)}.$$

Quadratic form criterion function $q_n(\theta) \geq 0$ is a scalar!

Weighting matrix W is symmetric (and positive definite that is $x'Wx > 0$ for all non-zero x)!

7.2 Get the sampling error (at least approximately)

Get an approximate deviation from the true θ_0

First order Taylor expansion of sample moments $\bar{m}_n(\hat{\theta}_{GMM})$ around $\bar{m}_n(\theta_0)$ at true parameters gives:

$$\bar{m}_n(\hat{\theta}_{GMM}) \approx \bar{m}_n(\theta_0) + \bar{G}_n(\bar{\theta})(\hat{\theta}_{GMM} - \theta_0),$$

where $\bar{G}_n(\bar{\theta}) = \frac{\partial \bar{m}_n(\bar{\theta})}{\partial \bar{\theta}'}$ and $\bar{\theta}$ is a point between $\hat{\theta}_{GMM}$ and θ_0.

Check the dimensions

First order Taylor expansion of sample moments $\bar{m}_n(\hat{\theta}_{GMM})$ around $\bar{m}_n(\theta_0)$ at true parameters gives:

$$\underset{L\times 1}{\bar{m}_n(\hat{\theta}_{GMM})} \approx \underset{L\times 1}{\bar{m}_n(\theta_0)} + \underset{L\times K}{\bar{G}_n(\bar{\theta})}\underset{K\times 1}{(\hat{\theta}_{GMM} - \theta_0)},$$

where $\bar{G}_n(\bar{\theta}) = \underset{1\times K}{\frac{\partial \bar{m}_n(\bar{\theta})}{\partial \underset{L\times 1}{\bar{\theta}'}}}$ and $\bar{\theta}$ is a point between $\hat{\theta}_{GMM}$ and θ_0, because of the

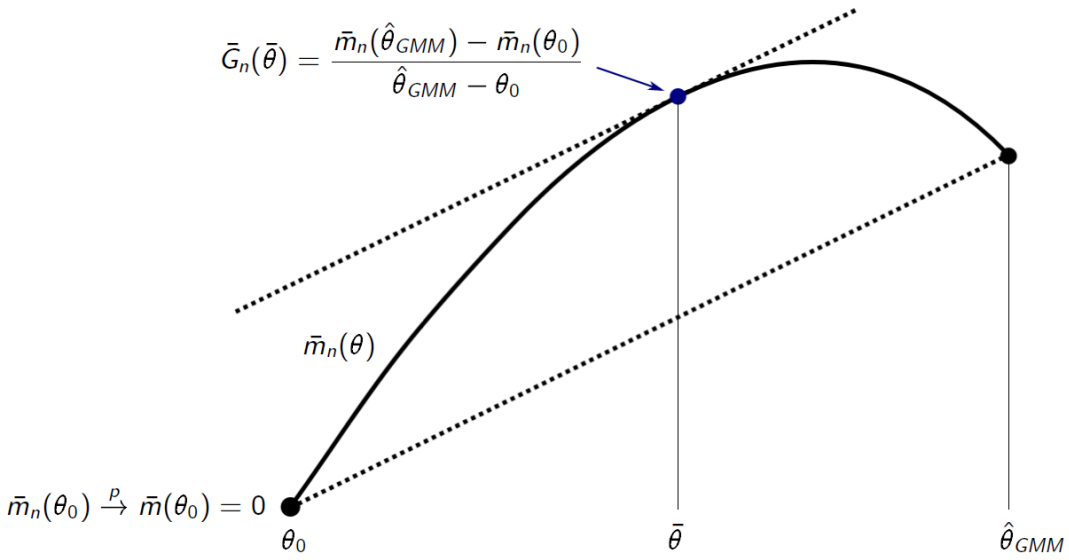
Mean value theorem...

Approximation introduced $\bar{\theta}$

...where $\bar{G}_n(\bar{\theta}) = \frac{\partial \bar{m}_n(\bar{\theta})}{\partial \bar{\theta}'}$ and $\bar{\theta}$ is a point between $\hat{\theta}_{GMM}$ and θ_0.

Mean value theorem

$$\bar{G}_n(\bar{\theta}) = \frac{\bar{m}_n(\hat{\theta}_{GMM}) - \bar{m}_n(\theta_0)}{\hat{\theta}_{GMM} - \theta_0} \text{ for } \theta_0 < \bar{\theta} < \hat{\theta}_{GMM}.$$



Do the minimization

To minimize the quadratic form criterion, we take the first derivative of

$$q_n(\theta) = \bar{m}_n(\theta)' W \bar{m}_n(\theta)$$

$$\frac{\partial q_n(\hat{\theta}_{GMM})}{\partial \hat{\theta}_{GMM}} = 2\bar{G}_n(\hat{\theta}_{GMM})' W_n \bar{m}_n(\hat{\theta}_{GMM}) = 0.$$

Express as much as possible asymptotically

$$\frac{\partial q_n(\hat{\theta}_{GMM})}{\partial \hat{\theta}_{GMM}} = 2\bar{G}_n(\hat{\theta}_{GMM})' W_n \bar{m}_n(\hat{\theta}_{GMM}) = 0,$$

Plug in the approximation from before

$$\bar{m}_n(\hat{\theta}_{GMM}) \approx \bar{m}_n(\theta_0) + \bar{G}_n(\bar{\theta})(\hat{\theta}_{GMM} - \theta_0)$$

to obtain

$$\bar{G}_n(\hat{\theta}_{GMM})' W_n \bar{m}_n(\theta_0) + \bar{G}_n(\hat{\theta}_{GMM})' W_n \bar{G}_n(\bar{\theta})(\hat{\theta}_{GMM} - \theta_0) \approx 0$$

which we rearrange to get the very useful

$$\hat{\theta}_{GMM} \approx \theta_0 - (\bar{G}_n(\hat{\theta}_{GMM})' W_n \bar{G}_n(\bar{\theta}))^{-1} \bar{G}_n(\hat{\theta}_{GMM})' W_n \bar{m}_n(\theta_0).$$

So the estimate $\hat{\theta}_{GMM}$ is approximately the true parameter θ_0 plus a sampling error that depends on the sample moment $\bar{m}_n(\theta_0)$.

Quickly check dimensions

Useful approximation

$$\underset{K\times 1}{\hat{\theta}_{GMM}} \approx \underset{K\times 1}{\theta_0} - (\underset{K\times L}{\bar{G}_n(\hat{\theta}_{GMM})'} \underset{L\times L}{W_n} \underset{L\times K}{\bar{G}_n(\bar{\theta})})^{-1} \underset{K\times L}{\bar{G}_n(\hat{\theta}_{GMM})'} \underset{L\times L}{W_n} \underset{L\times 1}{\bar{m}_n(\theta_0)}.$$

So the estimate $\hat{\theta}_{GMM}$ is approximately the true parameter θ_0 plus a sampling error that depends on the sample moment $\bar{m}_n(\theta_0)$.

7.3 The econometric model

Three assumptions: moment conditions

GMM1: Moment Conditions and Identification

.

$$\bar{m}(\theta_a) \neq \bar{m}(\theta_0) = E[m(X_i, Z_i, \theta_0)] = 0.$$

Identification implies that the probability limit of the GMM criterion function is uniquely minimized at the true parameters.

Three assumptions: law of large numbers

GMM2: Law of Large Numbers Applies

.

$$\bar{m}_n(\theta) = 1/n \sum_{i=1}^{N} m(X_i, Z_i, \theta_0) \xrightarrow{p} E[m(X_i, Z_i, \theta_0)].$$

The data meets the conditions for a law of large numbers to apply, so that we may assume that the empirical moments converge in probability to their expectation.

Three assumptions: central limit theorem

GMM3: Central Limit Theorem Applies

.

$$\sqrt{n}\bar{m}_n(\theta) = \sqrt{n}/n \sum_{i=1}^{N} m(X_i, Z_i, \theta_0) \xrightarrow{d} N[0, \Phi].$$

The empirical moments obey a central limit theorem. This assumes that the moments have a finite asymptotic covariance matrix $E[m(X_i, Z_i, \theta_0)m(X_i, Z_i, \theta_0)'] = \Phi$.

7.4 Consistency

Recall the useful approximation of the estimator:

$$\hat{\theta}_{GMM} \approx \theta_0 - (\bar{G}_n(\hat{\theta}_{GMM})'W_n\bar{G}_n(\bar{\theta}))^{-1}\bar{G}_n(\hat{\theta}_{GMM})'W_n\bar{m}_n(\theta_0).$$

Assumption GMM2 implies that

$$\bar{m}_n(\theta_0) = 1/n \sum_{i=1}^{N} m(X_i, Z_i, \theta_0) \xrightarrow{p} E[m(X_i, Z_i, \theta_0)] = \bar{m}(\theta_0).$$

That is, the sample moment equals the population moment in probability. Assumption GMM1 implies that

$$\bar{m}(\theta_0) = 0.$$

Then

$$\bar{m}_n(\theta_0) \xrightarrow{p} \bar{m}(\theta_0) = 0,$$

such that

$$\hat{\theta}_{GMM} \xrightarrow{p} \theta_0 \text{ for } N \to \infty$$

That is, by GMM1 and GMM2 the GMM estimator is consistent.

7.5 Asymptotic normality

Recall the useful approximation of the estimator:

$$\hat{\theta}_{GMM} \approx \theta_0 - (\bar{G}_n(\hat{\theta}_{GMM})' W_n \bar{G}_n(\bar{\theta}))^{-1} \bar{G}_n(\hat{\theta}_{GMM})' W_n \bar{m}_n(\theta_0).$$

Rewrite to obtain

$$\sqrt{n}(\hat{\theta}_{GMM} - \theta_0) \approx -(\bar{G}_n(\hat{\theta}_{GMM})' W_n \bar{G}_n(\bar{\theta}))^{-1} \bar{G}_n(\hat{\theta}_{GMM})' W_n \sqrt{n} \bar{m}_n(\theta_0),$$

The right hand side has several parts for which we made assumptions on what happens when $N \to \infty$. Under the central limit theorem (GMM3)

$$\sqrt{n}\bar{m}_n(\theta_0) \xrightarrow{d} N[0, \Phi]$$

$$plim W_n = W$$

$$plim \bar{G}_n(\hat{\theta}_{GMM}) = plim \bar{G}_n(\bar{\theta}) = plim \frac{\partial m(X_i, Z_i, \theta_0)}{\partial \theta_0'} = E\left[\frac{\partial \bar{m}(\theta_0)}{\partial \theta_0'}\right] = \Gamma(\theta_0)$$

With $plim W_n = W$ and

$$plim \bar{G}_n(\hat{\theta}_{GMM}) = plim \bar{G}_n(\bar{\theta}) = \Gamma(\theta_0)$$

the expression

$$\sqrt{n}(\hat{\theta}_{GMM} - \theta_0) \approx -(\bar{G}_n(\hat{\theta}_{GMM})' W_n \bar{G}_n(\bar{\theta}))^{-1} \bar{G}_n(\hat{\theta}_{GMM})' W_n \sqrt{n}\bar{m}_n(\theta_0)$$

becomes

$$\sqrt{n}(\hat{\theta}_{GMM} - \theta_0) \approx -(\Gamma(\theta_0)' W \Gamma(\theta_0))^{-1} \Gamma(\theta_0)' W \sqrt{n}\bar{m}_n(\theta_0)$$

from which we get the variance V. So

$$\sqrt{n}(\hat{\theta}_{GMM} - \theta_0) \xrightarrow{d} N[0, V]$$

with

$$\underset{K \times K}{V} = 1/n[\Gamma(\theta_0)' W \Gamma(\theta_0)]^{-1} [\Gamma(\theta_0)' W \Phi W' \Gamma(\theta_0)][\Gamma(\theta_0)' W \Gamma(\theta_0)]^{-1}$$

That is by GMM1, GMM2, and GMM3 the GMM estimator is asymptotic normal.

7.6 Asymptotic efficiency

Which weighting matrix W gives us the smallest possible asymptotic variance of the GMM estimator $\hat{\theta}_{GMM}$.

The variance of the GMM estimator V depends on the choice of W

$$V = 1/n[\Gamma(\theta_0)' W \Gamma(\theta_0)]^{-1} [\Gamma(\theta_0)' W \Phi W' \Gamma(\theta_0)][\Gamma(\theta_0)' W \Gamma(\theta_0)]^{-1}$$

So let us minimize V to get the *optimal* weight matrix. Try from GMM3

$$\underset{n \to \infty}{plim} W_n = W = \Phi^{-1}$$

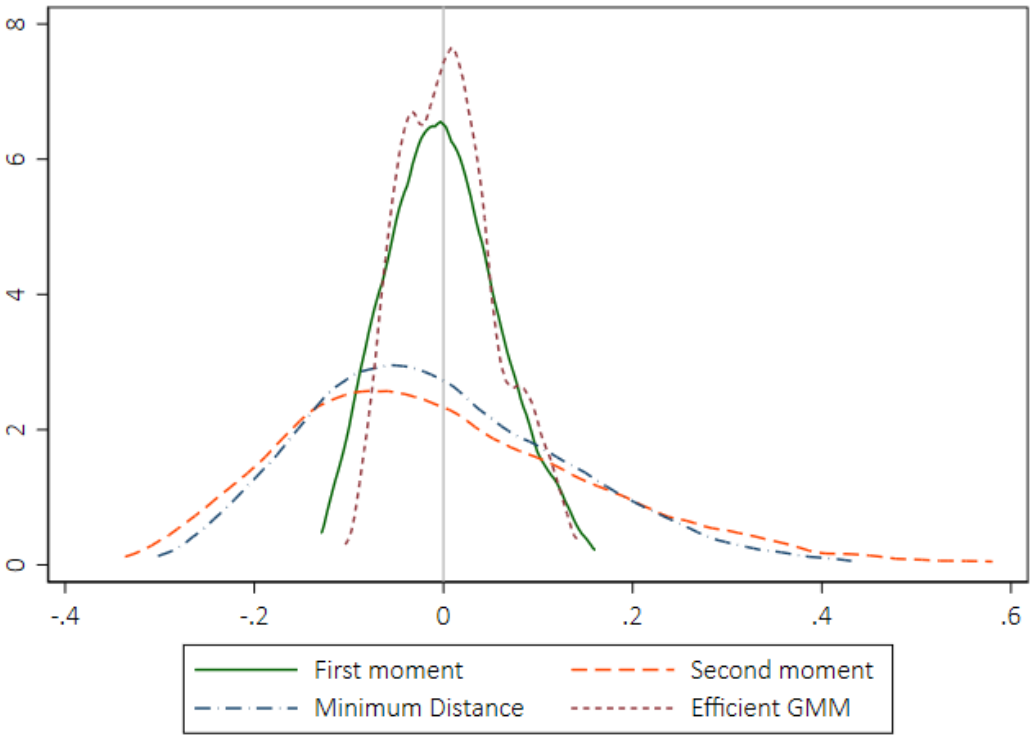
$$V_{GMM,optimal} = 1/n[\Gamma(\theta_0)' \Phi^{-1} \Gamma(\theta_0)]^{-1} [\Gamma(\theta_0)' \Phi^{-1} \Phi \Phi^{-1'} \Gamma(\theta_0)][\Gamma(\theta_0)' \Phi^{-1} \Gamma(\theta_0)]^{-1}$$

Which can be simplified to

$$V_{GMM,optimal} = 1/n[\Gamma(\theta_0)' \Phi^{-1} \Gamma(\theta_0)]^{-1}$$

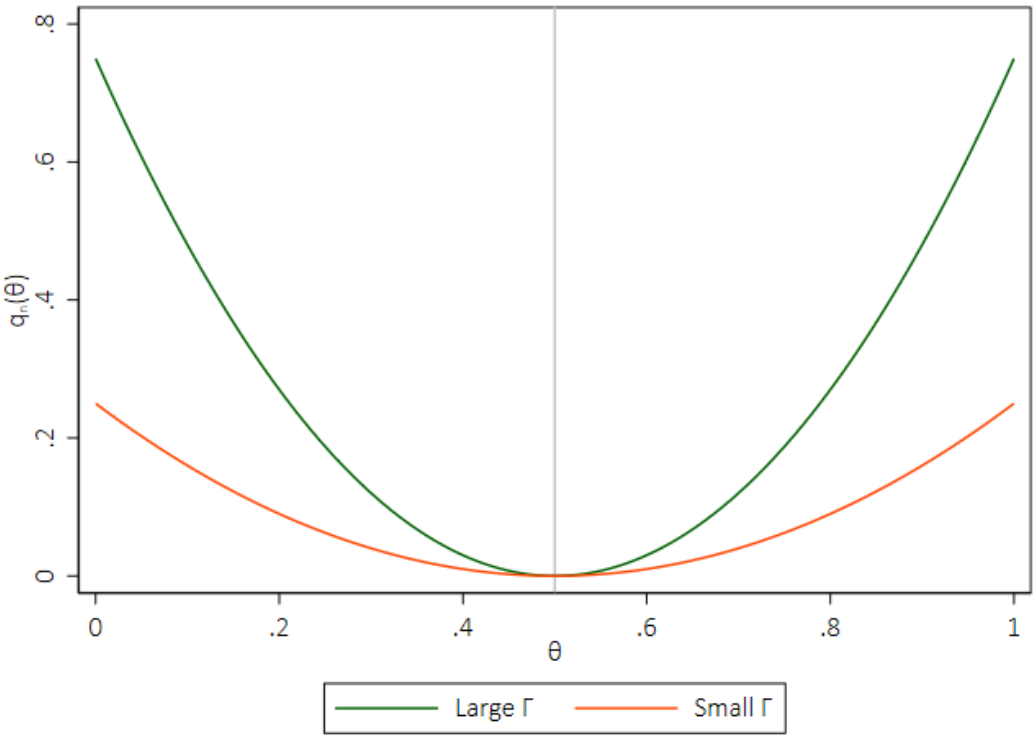
$$V_{GMM,optimal} = 1/n[\Gamma(\theta_0)' \Phi^{-1} \Gamma(\theta_0)]^{-1}$$

If Φ is small, there is little variation of this specific sample moment around zero and the moment condition is very informative about θ_0. So it is best to assign a high weight to it.



$$V_{GMM,optimal} = 1/n[\Gamma(\theta_0)'\Phi^{-1}\Gamma(\theta_0)]^{-1}$$

If Γ is large, there is a large penalty from violating the moment condition by evaluating at $\theta \neq \theta_0$. Then the moment condition is very informative about θ_0. V is inversely related to Γ.



Estimate the variance in practice

$$\hat{V}_{GMM,optimal} = 1/n[\bar{G}_n(\hat{\theta})'\Phi_n^{-1}\bar{G}_n(\hat{\theta})]^{-1}$$

Consistent estimator

$$\Phi_n = NV(\bar{m}_n(\hat{\theta}))$$

$$\bar{G}_n(\hat{\theta}) = \frac{\partial m(X_i, Z_i, \hat{\theta})}{\partial \hat{\theta}'}$$

8 Conclusion

Congratulations! If you made it through this document, you are ready to read some econometrics papers, program and develop new estimators, and analyze statistical properties. If this caught your interest, check out non-parametric and Bayesian econometrics.

References

Angrist, J. D., and J.-S. Pischke (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Cameron, A. C., and P. K. Trivedi (2005): *Microeconometrics: Methods and Applications*. Cambridge University Press, 3rd edn.

Filoso, V. (2013): "Regression Anatomy, Revealed," *The Stata Journal*, 13(1), 92–106.

Frisch, R., and F. V. Waugh (1933): "Partial Time Regressions as Compared with Individual Trends," *Econometrica*, 1(4), 387–401.

Greene, W. H. (2011): *Econometric Analysis*. Prentice Hall, 5th edn.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029–1054.

——— (2012): "Proofs for large sample properties of generalized method of moments estimators," *Journal of Econometrics*, 170(2), 325–330, Thirtieth Anniversary of Generalized Method of Moments.

Hill, R. C., W. E. Griffiths, and G. C. Lim (2010): *Principles of Econometrics*. John Wiley & Sons, 4th edn.

Kennedy, P. (2008): *A Guide to Econometrics*. Blackwell Publishing, 6th edn., In particular, Chapter 7, 8.1–8.3.

Lovell, M. C. (2008): "A Simple Proof of the FWL Theorem," *The Journal of Economic Education*, 39(1), 88–91.

Pishro-Nik, H. (2014): *Introduction to Probability, Statistics, and Random Processes*. Kappa Research LLC.

Plackett, R. L. (1972): "Studies in the History of Probability and Statistics. XXIX: The discovery of the method of least squares," *Biometrika*, 59(2), 239–251. Cited on page 63.

Rostam-Afschar, D., and R. Jessen (2014): "GRAPH3D: Stata module to draw colored, scalable, rotatable 3D plots," Statistical Software Components, Boston College Department of Economics.

Stock, J. H., and M. W. Watson (2012): *Introduction to Econometrics*. Pearson Addison-Wesley, 3rd edn.

Verbeek, M. (2012): *A Guide to Modern Econometrics*. John Wiley & Sons, 3rd edn.

Wooldridge, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

——— (2009): *Introductory Econometrics: A Modern Approach*. Cengage Learning, 4th edn. Cited on page 67.