Advanced Mathematical Methods

WS 2023/24

5 Parameter Estimation

Dr. Julie Schnaitmann

Department of Statistics, Econometrics and Empirical Economics

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

WIRTSCHAFTS- UND SOZIALWISSENSCHAFTLICHE FAKULTÄT

Online References

Applied Econometrics Lecture (by Prof. Grammig, available on TIMMS)

- Lecture 37: Law of Large Numbers
- Lecture 38: Central Limit Theorem

Implications of a random sample

$\{X_1, X_2, ..., X_n\}$ is a random sample if all draws of the random variable are **independently** and **identically distributed (iid)**.

Implications:

$$F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) \overset{\text{independent}}{=} F_{X_1}(x_1) \cdot F_{X_2}(x_2) \cdot ... \cdot F_{X_n}(x_n)$$
$$\overset{\text{identical}}{=} F_X(x_1) \cdot F_X(x_2) \cdot ... \cdot F_X(x_n)$$

$$f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) \overset{\text{independent}}{=} f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdot ... \cdot f_{X_n}(x_n)$$
$$\overset{\text{identical}}{=} f_X(x_1) \cdot f_X(x_2) \cdot ... \cdot f_X(x_n)$$

Implications of a random sample

$\{X_1, X_2, ..., X_n\}$ is a random sample if all draws of the random variable are **independently** and **identically distributed (iid)**.

Implications:

$$F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) \overset{\text{independent}}{=} F_{X_1}(x_1) \cdot F_{X_2}(x_2) \cdot ... \cdot F_{X_n}(x_n)$$

$$\overset{\text{identical}}{=} F_X(x_1) \cdot F_X(x_2) \cdot ... \cdot F_X(x_n)$$

$$f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) \overset{\text{independent}}{=} f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdot ... \cdot f_{X_n}(x_n)$$

$$\overset{\text{identical}}{=} f_X(x_1) \cdot f_X(x_2) \cdot ... \cdot f_X(x_n)$$

Implications of a random sample

$\{X_1, X_2, ..., X_n\}$ is a random sample if all draws of the random variable are **independently** and **identically distributed (iid)**.

Implications:

$$F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) \overset{\text{independent}}{=} F_{X_1}(x_1) \cdot F_{X_2}(x_2) \cdot ... \cdot F_{X_n}(x_n)$$

$$\overset{\text{identical}}{=} F_X(x_1) \cdot F_X(x_2) \cdot ... \cdot F_X(x_n)$$

$$f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) \overset{\text{independent}}{=} f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdot ... \cdot f_{X_n}(x_n)$$

$$\overset{\text{identical}}{=} f_X(x_1) \cdot f_X(x_2) \cdot ... \cdot f_X(x_n)$$

Asymptotic results ($n \to \infty$)

For a random sample $\{X_1, X_2, \ldots, X_n\}$ with finite $\mathbb{E}(X_i)$ and $Var(X_i)$ and an appropriately large n, following concepts apply:

1. Law of Large Numbers (LLN)
2. Central Limit Theorem (CLT)

Law of Large Numbers

Law of Large Numbers

$$\lim_{n\to\infty} P\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X)\right| > \varepsilon\right] = 0 \quad \text{for any } \epsilon > 0.$$

When the LLN holds, moments of the distribution of a population can be consistently estimated by moments of a random sample. I.e., we can write:

- $\text{plim} \frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}(X) = \mu$
- $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}(X) = \mu$

Law of Large Numbers

Law of Large Numbers

$$\lim_{n\to\infty} P\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X)\right| > \varepsilon\right] = 0 \quad \text{for any } \epsilon > 0.$$

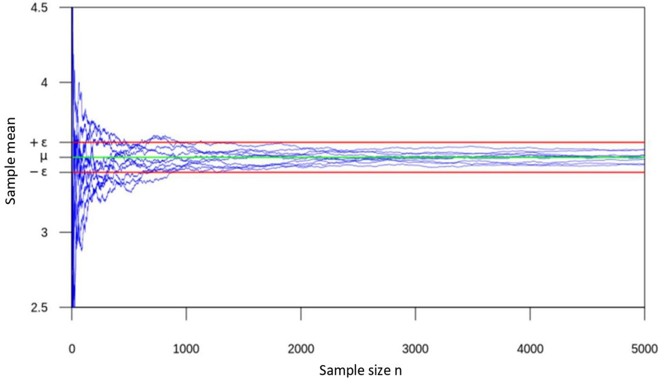
When the LLN holds, moments of the distribution of a population can be consistently estimated by moments of a random sample. I.e., we can write:

- $\text{plim} \frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}(X) = \mu$
- $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}(X) = \mu$

LLN and convergence in probability

LLN and convergence in probability

Illustration of convergence in probability using LLN:



$\Rightarrow$ A sequence $\{X_i\}$ converges in probability to a constant μ

Law of Large Numbers

Central Limit Theorem

Properties of the sample average

Central Limit Theorem

Properties of the sample average

Central Limit Theorem

Central Limit Theorem

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X)\right] \overset{a}{\sim} \mathcal{N}(0, Var(X))$$

E.g. if $\{X_i\}$ is **iid** with $\mathbb{E}(X_i) = \mu$ and $Var(X_i) = \sigma^2$ and $\bar{z}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu$ then,

$$\sqrt{n}\frac{(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} z \sim \mathcal{N}(0, 1)$$

$$\text{or} \quad \bar{X}_n - \mu \overset{a}{\sim} \mathcal{N}(0, \frac{\sigma^2}{n})$$

$$\text{or} \quad \bar{X}_n \overset{a}{\sim} \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

Parameter estimation

Underlying principle:

1. Assumption about the distribution of a random variable in the population: $\mathbb{E}(X) = \mu < \infty$, $Var(X) = \sigma^2 < \infty$
2. Draw of a random sample
3. Use estimation functions to estimate the unknown parameters of the distribution: μ, σ^2

Estimation functions (short: estimators) are measurable functions of random variables $\Rightarrow \widehat{\theta}_n = \widehat{\theta}_n(X_1, X_2, ..., X_n)$

Parameter estimation

Underlying principle:

1. Assumption about the distribution of a random variable in the population: $\mathbb{E}(X) = \mu < \infty$, $Var(X) = \sigma^2 < \infty$

2. Draw of a random sample

3. Use estimation functions to estimate the unknown parameters of the distribution: μ, σ^2

   Estimation functions (short: estimators) are measurable functions of random variables $\Rightarrow \widehat{\theta}_n = \widehat{\theta}_n(X_1, X_2, ..., X_n)$

Parameter estimation

Underlying principle:

1. Assumption about the distribution of a random variable in the population: $\mathbb{E}(X) = \mu < \infty$, $Var(X) = \sigma^2 < \infty$

2. Draw of a random sample

3. Use estimation functions to estimate the unknown parameters of the distribution: μ, σ^2

   Estimation functions (short: estimators) are measurable functions of random variables $\Rightarrow \widehat{\boldsymbol{\theta}}_n = \widehat{\theta}_n(X_1, X_2, ..., X_n)$

Quality of estimators

Finite sample properties of estimators:

- Bias
- Variance
- Efficiency (MSE)

Asymptotic concepts ($n \to \infty$):

- Consistency
- Asymptotic normality

Quality of estimators

Finite sample properties of estimators:

- Bias
- Variance
- Efficiency (MSE)

Asymptotic concepts ($n \to \infty$):

- Consistency
- Asymptotic normality

Quality of estimators

Bias of $\widehat{\theta}_n$:

- If $\mathbb{E}(\widehat{\theta}_n) = \theta \Rightarrow$ unbiased estimator
- If $\mathbb{E}(\widehat{\theta}_n) \neq \theta \Rightarrow$ biased estimator

Quality of estimators

Example

Consider a random variable X that follows a known distribution with $\mathbb{E}(X) = \mu$ and $Var(X) = \sigma^2$.

We want to estimate the parameter $\mu = \mathbb{E}(X)$ and can draw a random sample $\{X_1, \ldots, X_n\}$.

We have four different candidate estimation functions (estimators) for the parameter μ:

$(i) \quad \widehat{\theta}_n^{(a)} = X_3$

$(ii) \quad \widehat{\theta}_n^{(b)} = \frac{X_1 + X_n - 1}{2}$

$(iii) \quad \widehat{\theta}_n^{(c)} = \frac{1}{n} \sum_{i=1}^{n} X_i$

$(iv) \quad \widehat{\theta}_n^{(d)} = \frac{1}{n-1} \sum_{i=1}^{n} X_i$

Quality of estimators

Example

Quality of estimators

Variance of $\widehat{\theta}_n$:

Preferable: small variance of the estimation function $\Rightarrow$ small $Var[\widehat{\theta}_n(X_1, X_2, ..., X_n)]$

Quality of estimators

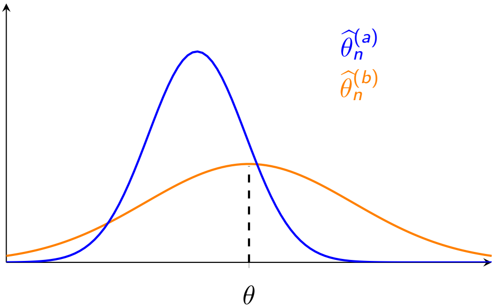
Example

Quality of estimators

Quality of estimators

Mean Squared Error (MSE) of $\widehat{\theta}_n$:

- Preferable: an unbiased estimator with a small variance
  $\Rightarrow$ Efficiency
- $\text{MSE}(\widehat{\theta}_n) \equiv \mathbb{E}[(\widehat{\theta}_n - \theta)^2] = Var(\widehat{\theta}_n) + Bias(\widehat{\theta}_n)^2$
- Trade-off: variance vs. bias of an estimator



Quality of estimators

Mean Squared Error (MSE)

Quality of estimators

Consistency

Quality of estimators

Consistency of $\widehat{\theta}_n$:

- The bigger the sample size gets, the smaller the sampling error $|\widehat{\theta}_n - \theta|$ should become
- Formally $\widehat{\theta}_n$ is consistent if

$$\lim_{n\to\infty} P(|\widehat{\theta}_n - \theta| > \varepsilon) = 0$$

- Short-hand notation: $\widehat{\theta}_n \xrightarrow{p} \theta$ or $\operatorname*{plim}_{n\to\infty} \widehat{\theta}_n = \theta$

Quality of estimators

Example

Quality of estimators

Convergence in distribution and asymptotic normality of $\widehat{\theta}_n$

Convergence in distribution: $z_n \xrightarrow{d} z$

If the c.d.f. of z_n converges to the c.d.f. of z at each point of continuity, then z_n converges in distribution to z.

Asymptotic normality of $\widehat{\theta}_n$:

The distribution of an estimator $\widehat{\theta}_n$ converges to the distribution of another random variable z, which is normally distributed.
Generally:

$z_n \xrightarrow{d} z \sim \mathcal{N}(0, 1)$

Applied to the Central Limit Theorem:

$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} z \sim \mathcal{N}(0, Var(\widehat{\theta}_n))$

Quality of estimators

Convergence in distribution and asymptotic normality of $\widehat{\theta}_n$

Convergence in distribution: $z_n \xrightarrow{d} z$

If the c.d.f. of z_n converges to the c.d.f. of z at each point of continuity, then z_n converges in distribution to z.

Asymptotic normality of $\widehat{\theta}_n$:

The distribution of an estimator $\widehat{\theta}_n$ converges to the distribution of another random variable z, which is normally distributed.
Generally:

$$z_n \xrightarrow{d} z \sim \mathcal{N}(0,1)$$

Applied to the Central Limit Theorem:

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} z \sim \mathcal{N}(0, Var(\widehat{\theta}_n))$$

Quality of estimators

Convergence in distribution and asymptotic normality of $\widehat{\theta}_n$

Convergence in distribution: $z_n \xrightarrow{d} z$

If the c.d.f. of z_n converges to the c.d.f. of z at each point of continuity, then z_n converges in distribution to z.

Asymptotic normality of $\widehat{\theta}_n$:

The distribution of an estimator $\widehat{\theta}_n$ converges to the distribution of another random variable z, which is normally distributed.
Generally:

$$z_n \xrightarrow{d} z \sim \mathcal{N}(0, 1)$$

Applied to the Central Limit Theorem:

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} z \sim \mathcal{N}(0, Var(\widehat{\theta}_n))$$

Estimation techniques

1. Method of Moments
2. Maximum Likelihood Method
3. Ordinary Least Squares (OLS)

Method of Moments

Idea:

- Use the known relationship between parameters and theoretical moments
- Replace theoretical by empirical moments
- Method of Moment estimators are consistent

Method of Moments

Example

Method of Moments

Example

Method of Moments

Exponential distribution $f_X(x;\lambda) = \lambda e^{-\lambda x}$:

- theoretical moments:
  1. $\mathbb{E}(X) = \frac{1}{\lambda} \Longrightarrow \lambda = \frac{1}{\mathbb{E}(X)}$
  2. $Var(X) = \frac{1}{\lambda^2} \Longrightarrow \lambda = \frac{1}{\sqrt{Var(X)}} = \frac{1}{\sqrt{\mathbb{E}(X^2) - \mathbb{E}(X)^2}}$
- replace by empirical moments:
  1. $\widehat{\lambda}_1 = \frac{1}{\frac{1}{n}\sum_{i=1}^n X_i}$
  2. $\widehat{\lambda}_2 = \frac{1}{\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2 - (\frac{1}{n}\sum_{i=1}^n X_i)^2}}$

Maximum Likelihood Method

Idea:

- Choose $\widehat{\theta}_n$ such that the likelihood function $\mathcal{L}(\tilde{\theta})$ is maximized
- The likelihood function $\mathcal{L}(\tilde{\theta})$ is the joint density function of $X_1, ..., X_n$ with parameters $\tilde{\theta}$

$$\mathcal{L}(\tilde{\theta}) = f_{X_1, X_2, ..., X_n}(x_1, x_2, ...x_n; \tilde{\theta})$$

Intuition:

1. We decide on the distribution of $X \sim D(\theta)$ in the population
2. Draw a random sample of X
3. Keep the sample fixed and choose the parameter such that $\mathcal{L}(\tilde{\theta})$ is maximized $\Rightarrow$ "Maximize the probability to observe the realization of our sample"

Maximum Likelihood Method

Idea:

- Choose $\widehat{\theta}_n$ such that the likelihood function $\mathcal{L}(\tilde{\theta})$ is maximized
- The likelihood function $\mathcal{L}(\tilde{\theta})$ is the joint density function of $X_1, ..., X_n$ with parameters $\tilde{\theta}$

$$\mathcal{L}(\tilde{\theta}) = f_{X_1, X_2, ..., X_n}(x_1, x_2, ... x_n; \tilde{\theta})$$

Intuition:

1. We decide on the distribution of $X \sim D(\theta)$ in the population
2. Draw a random sample of X
3. Keep the sample fixed and choose the parameter such that $\mathcal{L}(\tilde{\theta})$ is maximized $\Rightarrow$ "Maximize the probability to observe the realization of our sample"

Maximum Likelihood Method

Recipe:

1. Set up the likelihood function and exploit the **iid** sample implications
2. Take the *ln* of the likelihood function
3. Maximize the log-likelihood function w.r.t. the parameter $\tilde{\theta}$

Maximum Likelihood Method

Recipe:

1. Set up the likelihood function and exploit the **iid** sample implications
2. Take the *ln* of the likelihood function
3. Maximize the log-likelihood function w.r.t. the parameter $\tilde{\theta}$

ML - 1. Set up the likelihood function

Use the **iid** property of a random sample:

$$\mathcal{L}(\tilde{\boldsymbol{\theta}}) = f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n; \tilde{\boldsymbol{\theta}})$$

$$\overset{\text{independent}}{=} f_{X_1}(x_1; \tilde{\boldsymbol{\theta}}) \cdot f_{X_2}(x_2; \tilde{\boldsymbol{\theta}}) \cdot \ldots \cdot f_{X_n}(x_n; \tilde{\boldsymbol{\theta}})$$

$$\overset{\text{identical}}{=} f_X(x_1; \tilde{\boldsymbol{\theta}}) \cdot f_X(x_2; \tilde{\boldsymbol{\theta}}) \cdot \ldots \cdot f_X(x_n; \tilde{\boldsymbol{\theta}})$$

$$= \prod_{i=1}^{n} f_X(x_i; \tilde{\boldsymbol{\theta}})$$

ML - 2. Log-Transformation

Logarithmic rules

1. $ln(a \cdot b) \Leftrightarrow ln(a) + ln(b)$
2. $ln(\frac{a}{b}) \Leftrightarrow ln(a) - ln(b)$
3. $ln(a)^b \Leftrightarrow \ b \cdot ln(a)$
4. $ln(e^x) \Leftrightarrow e^{ln(x)} \Leftrightarrow x$

$$\mathcal{L}(\tilde{\boldsymbol{\theta}}) = \prod_{i=1}^{n} f_X(x_i; \tilde{\boldsymbol{\theta}}) \qquad | \, ln$$

$$ln\mathcal{L}(\tilde{\boldsymbol{\theta}}) = \sum_{i=1}^{n} ln \, f_X(x_i; \tilde{\boldsymbol{\theta}})$$

ML - 3. Log-likelihood maximization

Maximize the log-likelihood function:

$$\widehat{\boldsymbol{\theta}}_{ML} = \underset{\tilde{\boldsymbol{\theta}}}{\operatorname{argmax}}\, ln\mathcal{L}(\tilde{\boldsymbol{\theta}}) = \sum_{i=1}^{n} ln\, f_X(x_i; \tilde{\boldsymbol{\theta}})$$

Determine $\widehat{\boldsymbol{\theta}}_{ML}$ via the F.O.C. $\frac{\partial ln\mathcal{L}(\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}} \overset{!}{=} 0$:

$$\sum_{i=1}^{n} \frac{\partial ln\, f_X(x_i; \widehat{\boldsymbol{\theta}})}{\partial \tilde{\theta}_1} \overset{!}{=} 0$$

$$\vdots$$

$$\sum_{i=1}^{n} \frac{\partial ln\, f_X(x_i; \widehat{\boldsymbol{\theta}})}{\partial \tilde{\theta}_K} \overset{!}{=} 0$$

Maximum Likelihood Method

Example

Maximum Likelihood Method

Example

Maximum Likelihood Method

Example

Maximum Likelihood Method

Example

Properties of the ML estimator

If the likelihood function is correctly specified, the ML estimator has following properties:

- Consistency
- Asymptotic efficiency (for large n this estimator has the smallest MSE)
- Asymptotic normality

Conditional Maximum Likelihood (CML)

Procedure:

1. Specification of the conditional distribution $Y|X = x$
2. Specification of conditional moments (functions of X)
3. Insert conditional moments into the likelihood function
4. Maximize the (log-) likelihood function w.r.t. the unknown parameter

The marginal, joint and conditional distribution

Relationship of the marginal, joint, and conditional distribution:

$$f_{X_1|X_2} = \frac{f_{X_1,X_2}}{f_{X_2}}$$

$$\Leftrightarrow \quad f_{X_1,X_2} = f_{X_1|X_2} \cdot f_{X_2}$$

$$\Leftrightarrow \quad f_{X_2} = \frac{f_{X_1,X_2}}{f_{X_1|X_2}}$$

The marginal, joint and conditional distribution

Example: Exploiting this relationship allows us to write the joint density $f_{X_1,\ldots,X_5}$ as product of four conditional and one marginal density:

$$f_{X_1,X_2} = f_{X_2|X_1} \cdot f_{X_1} \tag{1}$$

$$\underline{f_{X_1,X_2,X_3}} = \underline{f_{X_3|X_1,X_2} \cdot f_{X_1,X_2}} \tag{2}$$

$$f_{X_1,X_2,X_3,X_4} = f_{X_4|X_1,X_2,X_3} \cdot \underline{f_{X_1,X_2,X_3}} \tag{3}$$

$$f_{X_1,X_2,X_3,X_4,X_5} = f_{X_5|X_1,X_2,X_3,X_4} \cdot f_{X_1,X_2,X_3,X_4} \tag{4}$$

Plugging in (1)-(4) $f_{X_1,\ldots,X_5}$ can be expressed as:

$$= f_{X_5|X_1,X_2,X_3,X_4} \cdot f_{X_4|X_1,X_2,X_3} \cdot f_{X_3|X_1,X_2} \cdot f_{X_2|X_1} \cdot f_{X_1}$$

Conditional Maximum Likelihood (CML)

- A binary response model $\rightarrow$ outcome Y is either 0 or 1
- $Y|X = x \sim Be(p(x))$
- specify a model for the probability of succes $p(x)$. Some examples:
  (a) linear probability model

$$p(x) = E[Y|X = x] = \beta_0 + \beta_1 x$$

  (b) nonlinear probability model $\rightarrow$ probability for a certain event doesn't change linear in x - here: Probit model

$$p(x) = E[Y|X = x] = \underbrace{F(\beta_0 + \beta_1 x) = \int_{-\infty}^{\beta_0+\beta_1 x} \frac{1}{\sqrt{2\pi}} e^{\frac{z^2}{2}} dz}_{c.d.f.\ of\ a\ standard\ normal\ distribution}$$

Conditional Maximum Likelihood (CML)

- A binary response model $\rightarrow$ outcome Y is either 0 or 1
- $Y|X = x \sim Be(p(x))$
- specify a model for the probability of succes $p(x)$. Some examples:
  (a) linear probability model

$$p(x) = E[Y|X = x] = \beta_0 + \beta_1 x$$

  (b) nonlinear probability model $\rightarrow$ probability for a certain event doesn't change linear in x - here: Probit model

$$p(x) = E[Y|X = x] = \underbrace{F(\beta_0 + \beta_1 x) = \int_{-\infty}^{\beta_0 + \beta_1 x} \frac{1}{\sqrt{2\pi}} e^{\frac{z^2}{2}} dz}_{c.d.f.\ of\ a\ standard\ normal\ distribution}$$

Conditional Maximum Likelihood (CML)

- A binary response model $\rightarrow$ outcome Y is either 0 or 1
- $Y|X = x \sim Be(p(x))$
- specify a model for the probability of succes $p(x)$. Some examples:
  (a) linear probability model

$$p(x) = E[Y|X = x] = \beta_0 + \beta_1 x$$

  (b) nonlinear probability model $\rightarrow$ probability for a certain event doesn't change linear in x - here: Probit model

$$p(x) = E[Y|X = x] = \underbrace{F(\beta_0 + \beta_1 x) = \int_{-\infty}^{\beta_0+\beta_1 x} \frac{1}{\sqrt{2\pi}} e^{\frac{z^2}{2}} dz}_{c.d.f.\ of\ a\ standard\ normal\ distribution}$$

Conditional Maximum Likelihood (CML)

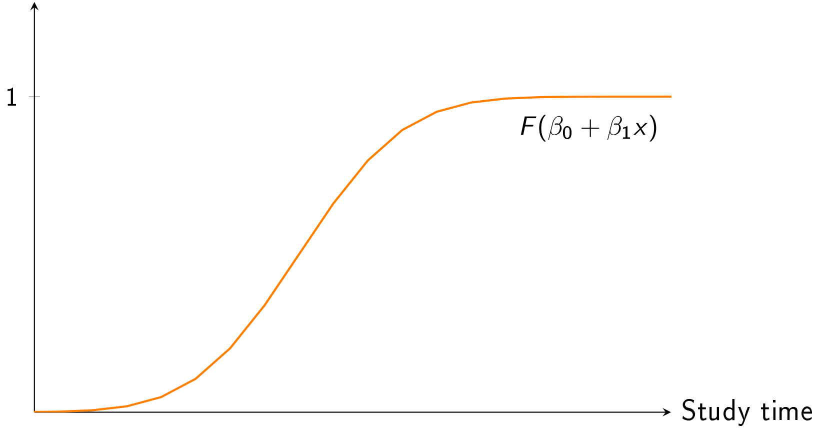
- Pass or fail in an exam given the amount of study time
- X: Study time
  Y = 1: Pass
  Y = 0: Fail
- Probability to pass the exam doesn't change linearly with increasing study time
- $P(Y = 1|X) = p(x) \rightarrow$ probability Pass

  $P(Y = 0|X) = 1 - P(Y = 1|X) = 1 - p(x) \rightarrow$ probability Fail

  $\Rightarrow$ Bernoulli distribution

Conditional Maximum Likelihood (CML)

Probability to pass the exam as a function of study time:



Conditional Maximum Likelihood (CML)

Estimating the parameters β_0 and β_1:

1. Set up conditional likelihood function using a random sample

$$\mathcal{L}(\tilde{\beta}_0, \tilde{\beta}_1) = \prod_{i=1}^{n} \underbrace{p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}}_{\text{Bernoulli prob fct}}$$

2. Plug in expression for $p(x_i)$ and caluclate log-likelihood
3. Maximize the log-likelihood w.r.t. β_0 and β_1

$$\frac{\partial ln\mathcal{L}}{\partial \tilde{\beta}_0} \overset{!}{=} 0 \qquad \frac{\partial ln\mathcal{L}}{\partial \tilde{\beta}_1} \overset{!}{=} 0$$

4. Solve the F.O.C. for $\widehat{\beta}_0$ and $\widehat{\beta}_1$

Conditional Maximum Likelihood (CML)

Example

Conditional Maximum Likelihood (CML)

Example

Conditional Maximum Likelihood (CML)

Example

Ordinary Least Squares (OLS)

Basic idea:

- Explain a variable Y (dependent variable) as linear combination of other variables $X_1, X_2, ..., X_K$ (explanatory variables/regressors) and the residual

$$Y = \tilde{\beta}_1 X_1 + \tilde{\beta}_2 X_2 + ... + \tilde{\beta}_K X_K + \tilde{u}$$

- While Y and X are directly observable, $\tilde{u}$ is not $\Rightarrow$ $\tilde{u}$ depends on how we choose $\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_K$

$$\tilde{u}_i = \tilde{u}_i(\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_K) = y_i - \tilde{\beta}_1 X_{i1} - \tilde{\beta}_2 X_{i2} - ... - \tilde{\beta}_K X_{iK}$$

- How can $\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_K$ be chosen?

Ordinary Least Squares (OLS)

Basic idea:

- Explain a variable Y (dependent variable) as linear combination of other variables $X_1, X_2, ..., X_K$ (explanatory variables/regressors) and the residual

$$Y = \tilde{\beta}_1 X_1 + \tilde{\beta}_2 X_2 + ... + \tilde{\beta}_K X_K + \tilde{u}$$

- While Y and X are directly observable, $\tilde{u}$ is not $\Rightarrow$ $\tilde{u}$ depends on how we choose $\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_K$

$$\tilde{u}_i = \tilde{u}_i(\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_K) = y_i - \tilde{\beta}_1 X_{i1} - \tilde{\beta}_2 X_{i2} - ... - \tilde{\beta}_K X_{iK}$$

- How can $\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_K$ be chosen?

Ordinary Least Squares (OLS)

Basic idea:

- Explain a variable Y (dependent variable) as linear combination of other variables $X_1, X_2, ..., X_K$ (explanatory variables/regressors) and the residual

$$Y = \tilde{\beta}_1 X_1 + \tilde{\beta}_2 X_2 + ... + \tilde{\beta}_K X_K + \tilde{u}$$

- While Y and X are directly observable, $\tilde{u}$ is not $\Rightarrow$ $\tilde{u}$ depends on how we choose $\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_K$

$$\tilde{u}_i = \tilde{u}_i(\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_K) = y_i - \tilde{\beta}_1 X_{i1} - \tilde{\beta}_2 X_{i2} - ... - \tilde{\beta}_K X_{iK}$$

- How can $\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_K$ be chosen?

Ordinary Least Squares (OLS)

Ordinary Least Squares (OLS)

Example for $k = 2$:

Ordinary Least Squares (OLS)

Example for $k = 2$:

Ordinary Least Squares (OLS)

Example for $k = 2$:

OLS - Choosing $\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_K$

Choosing $\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_K$ either by:

1. Minimization of the least squares function:

$$\widehat{\boldsymbol{\beta}} = \underset{\tilde{\beta}}{argmin} \sum_{i=1}^{n} (y_i - \tilde{\beta}' x_i)^2$$

2. Moment restrictions: $\widehat{u} = y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_{i2} - ... - \widehat{\beta}_K x_{iK}$

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i x_{i1} \stackrel{!}{=} 0\,, \quad \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i x_{i2} \stackrel{!}{=} 0\,, \; \ldots\,, \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i x_{iK} \stackrel{!}{=} 0$$

$\Rightarrow$ Both approaches lead to the same result

$$\widehat{\beta}_{\boldsymbol{n}} = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i' \boldsymbol{x}_i\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i' y_i\right) = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{y})$$

OLS - Choosing $\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_K$

Choosing $\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_K$ either by:

1. Minimization of the least squares function:

$$\widehat{\boldsymbol{\beta}} = \underset{\tilde{\beta}}{argmin} \sum_{i=1}^{n} (y_i - \tilde{\beta}' x_i)^2$$

2. Moment restrictions: $\widehat{u} = y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_{i2} - ... - \widehat{\beta}_K x_{iK}$

$$\frac{1}{n} \sum_{i=1}^{n} \widehat{u}_i x_{i1} \overset{!}{=} 0 \,, \quad \frac{1}{n} \sum_{i=1}^{n} \widehat{u}_i x_{i2} \overset{!}{=} 0 \,, \; \ldots \,, \frac{1}{n} \sum_{i=1}^{n} \widehat{u}_i x_{iK} \overset{!}{=} 0$$

$\Rightarrow$ Both approaches lead to the same result

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} = \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x_i' x_i} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x_i'} y_i \right) = (\boldsymbol{X'X})^{-1} (\boldsymbol{X'y})$$

OLS - classical assumptions

1. Linearity
   $Y = \beta_1 + \beta_2 X_2 + ... + \beta_K X_K$
2. Strict exogeneity
   $\mathbb{E}(u|X_1, X_2, ..., X_K) = 0$
3. Conditional homoscedasticity
   $Var(u|X_1, X_2, ..., X_K) = \sigma^2$
4. Distribution assumption
   $u|X_1, X_2, ..., X_K \sim \mathcal{N}(0, \sigma^2)$

OLS and ML

Example for $k = 2$:

OLS and ML

Example for $k = 2$:

OLS Example

We want to estimate the effect of more education on wage (β_2: return to schooling)

$$wage = \beta_1 + \beta_2 education + \beta_3 experience + \beta_4 experience^2 + u$$

Estimate $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}(\boldsymbol{X'y})$

$\Rightarrow$ Under the above mentioned assumptions $\widehat{\beta}_2$ gives the estimated marginal effect (ceteris paribus) of schooling on wage