Statistical Inference

Chapter 1

Stochastic vectors and matrices

Random vectors

- The vector $X = (X_1, \ldots, X_p)'$ is called a random vector or p-dimensional random variable, if the components $X_1, \ldots, X_p$ are one dimensional random variables.
- The vector X is called continuous if there is a function $f(x) = f(x_1, \ldots, x_p) \geq 0$ such that

$$P(a_1 \leq X_1 \leq b_1, \ldots, a_p \leq X_p \leq b_p) = \int_{a_p}^{b_p} \ldots \int_{a_1}^{b_1} f(x_1, \ldots, x_p)\, dx_1 \ldots dx_p.$$

  The function f is called (joint) probability density function (p.d.f.) of X.

Random vectors

- The random vector X is called discrete, if X has only values in a finite or countable set $\{x_1, x_2, \ldots\} \subset \mathbb{R}^p$. The function f with

$$f(x) = \begin{cases} P(X = x) & x \in \{x_1, x_2, \ldots\} \\ 0 & \text{else} \end{cases}$$

  is called probability function or discrete p.d.f. of X.

Random vectors

Continuous random vector

Consider the 2-dimensional continuous random vector $x = (x_1, x_2)'$ with pdf

$$f(x_1, x_2) = \begin{cases} 0.8(x_1 + x_2 + x_1 x_2) & 0 \leq x_1 \leq 1 \\ & 0 \leq x_2 \leq 1 \\ 0 & \text{else.} \end{cases}$$

Random vectors

Discrete random vector

Consider the two dimensional random vector $(X, Y)'$ with

$$X = \text{proseminar grade} \qquad Y = \text{final exam grade}$$

We have the following distribution:

Y/X	1	2	3	4
1	$\frac{50}{1007}$	$\frac{40}{1007}$	$\frac{12}{1007}$	$\frac{6}{1007}$
2	$\frac{55}{1007}$	$\frac{97}{1007}$	$\frac{54}{1007}$	$\frac{10}{1007}$
3	$\frac{28}{1007}$	$\frac{100}{1007}$	$\frac{68}{1007}$	$\frac{33}{1007}$
4	$\frac{13}{1007}$	$\frac{79}{1007}$	$\frac{75}{1007}$	$\frac{36}{1007}$
5	$\frac{9}{1007}$	$\frac{44}{1007}$	$\frac{119}{1007}$	$\frac{79}{1007}$

Marginal and conditional distribution

- Let the p-dimensional random vector $X = (X_1, \ldots, X_p)'$ be partitioned into the p_1-dimensional vector X_1 and the p_2-dimensional Vector X_2, i.e. $X = (X_1', X_2')'$.
- The p_1-dimensional p.d.f. or probability function $f_{X_1}(x_1)$ of X_1 is then called marginal p.d.f. or marginal probability function of X. It is given by

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_{p_1+1} \ldots dx_p$$

for continuous random vectors, and

$$f_{X_1}(x_1) = \sum_{x_2} f(x_1, x_2)$$

for discrete random vectors.

Marginal and conditional distribution

- The conditional p.d.f. or probability function of X_1 given $X_2 = x_2$ is defined as

$$f(x_1|x_2) = \begin{cases} \dfrac{f(x_1, x_2)}{f_{X_2}(x_2)} & \text{for } f_{X_2}(x_2) > 0 \\ 0 & \text{else.} \end{cases}$$

  The marginal and conditional p.d.f.'s or probability functions for X_2 are defined in complete analogy.

Marginal and conditional distribution

Continuous random vector

Consider the 2-dimensional continuous random vector $x = (x_1, x_2)'$ with pdf

$$f(x_1, x_2) = \begin{cases} 0.8(x_1 + x_2 + x_1 x_2) & 0 \leq x_1 \leq 1 \\ & 0 \leq x_2 \leq 1 \\ 0 & \text{else.} \end{cases}$$

Marginal and conditional distribution

Discrete random vector

Y/X	1	2	3	4
1	$\frac{50}{1007}$	$\frac{40}{1007}$	$\frac{12}{1007}$	$\frac{6}{1007}$
2	$\frac{55}{1007}$	$\frac{97}{1007}$	$\frac{54}{1007}$	$\frac{10}{1007}$
3	$\frac{28}{1007}$	$\frac{100}{1007}$	$\frac{68}{1007}$	$\frac{33}{1007}$
4	$\frac{13}{1007}$	$\frac{79}{1007}$	$\frac{75}{1007}$	$\frac{36}{1007}$
5	$\frac{9}{1007}$	$\frac{44}{1007}$	$\frac{119}{1007}$	$\frac{79}{1007}$

Expectation or mean vector

Let $X = (X_1, \ldots, X_p)'$ be a p-dimensional random vector. Then

$$E(X) = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)' = (E(X_1), \ldots, E(X_p))'$$

is called mean vector of X.

Example

Covariance and correlation matrix

The covariance matrix $Cov(X) = \mathbf{\Sigma}$ of a p-dimensional random vector X is defined as

$$Cov(X) = \mathbf{\Sigma} \quad = \quad E(X - \mu)(X - \mu)' = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{pmatrix},$$

where $\sigma_{ij} = Cov(X_i, X_j)$, $i \neq j$, is the covariance between X_i and X_j, and $\sigma_{ii} = \sigma_i^2 = Var(X_i)$ is the variance of X_i.

Covariance and correlation matrix

The correlation matrix $\boldsymbol{R}$ of X is defined as

$$\boldsymbol{R} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \vdots & & & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix},$$

where

$$\rho_{ij} = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i) \cdot \text{Var}(X_j)}}.$$

Covariance and correlation matrix

Example

Properties of expectations and covariance matrices

Let X and Y be random vectors and $\boldsymbol{A}$, $\boldsymbol{B}$, a, b matrices and vectors.

1. $E(X + Y) = E(X) + E(Y)$
2. $E(\boldsymbol{A}X + b) = \boldsymbol{A} \cdot E(X) + b$
3. $Cov(X) = E(XX') - \mu\mu'$
4. $Var(a'X) = a'Cov(X)a = \sum_{i=1}^{k} \sum_{j=1}^{k} a_i a_j \sigma_{ij}$
5. The covariance matrix is symmetric and positive semi definite.
6. $Cov(\boldsymbol{A}X + b) = \boldsymbol{A}Cov(X)\boldsymbol{A}'$

Multivariate Normal Distribution

- A continuous p-dimensional random vector $X = (X_1, X_2, \ldots, X_p)'$ is said to have a multivariate normal (or Gaussian) distribution if it has p.d.f.

$$f(x) = (2\pi)^{-\frac{p}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(x-\mu)'\mathbf{\Sigma}^{-1}(x-\mu)\right]$$

  with $\mu \in \mathbb{R}^p$ and positive definite $(p \times p)$-matrix $\mathbf{\Sigma}$.
- It can be shown that $E(X) = \mu$ and $\text{Cov}(X) = \mathbf{\Sigma}$.
- We write

$$X \sim N_p(\mu, \mathbf{\Sigma}),$$

  The special case $\mu = \mathbf{0}$ and $\mathbf{\Sigma} = \boldsymbol{I}$ is called the (multivariate) standard normal distribution.

Multivariate Normal Distribution

Example

$X = (X_1, X_2, X_3, X_4)' \sim N(\mu, \mathbf{\Sigma})$ with

$$\mu = (1, 2, 3, 4)' \qquad \mathbf{\Sigma} = \begin{pmatrix} 3 & 1 & 0 & 1 \\ 1 & 4 & 2 & 0 \\ 0 & 2 & 5 & 2 \\ 1 & 0 & 2 & 4 \end{pmatrix}$$

Marginal and conditional distributions

- Let the multivariate normal random variable $X \sim N(\mu, \mathbf{\Sigma})$ be partitioned into the subvectors $Y = (X_1, \ldots, X_r)'$ and $Z = (X_{r+1}, \ldots, X_p)'$, i.e.

$$X = \begin{pmatrix} Y \\ Z \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_Y \\ \mu_Z \end{pmatrix}, \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_Y & \mathbf{\Sigma}_{YZ} \\ \mathbf{\Sigma}_{ZY} & \mathbf{\Sigma}_Z \end{pmatrix}.$$

- Then Y has an r-dimensional normal distribution $Y \sim N(\mu_Y, \mathbf{\Sigma}_Y)$.
- The conditional distribution of Y given Z is again multivariate normal with mean

$$\mu_{Y|Z} = \mu_Y + \mathbf{\Sigma}_{YZ} \cdot \mathbf{\Sigma}_Z^{-1}(Z - \mu_Z)$$

and covariance matrix

$$\mathbf{\Sigma}_{Y|Z} = \mathbf{\Sigma}_Y - \mathbf{\Sigma}_{YZ}\mathbf{\Sigma}_Z^{-1}\mathbf{\Sigma}_{ZY}.$$

Marginal and conditional distributions

- Furthermore, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are independent if and only if $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are uncorrelated, i.e. if $\boldsymbol{\Sigma}_{YZ} = \boldsymbol{\Sigma}_{ZY} = \mathbf{0}$.
- The equivalence is generally not true for non-normal random vectors: If $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are independent they are also uncorrelated, but in general $\boldsymbol{\Sigma}_{ZY} = \mathbf{0}$ does not imply independence.

Marginal and conditional distributions

Example

$Y = (X_1, X_2)', Z = (Z_1, Z_2)'$

Linear transformations

Assume $X \sim N_p(\mu, \boldsymbol{\Sigma})$ is multivariate normal. Then the linear transformation

$$Y = \boldsymbol{D}X + d$$

with the $m \times p$ matrix $\boldsymbol{D}$ and the $m \times 1$ vector d is again multivariate normal

$$Y \sim N_m(\boldsymbol{D}\mu + d, \boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}').$$

Statistical Inference

Chapter 2

Classical Inference

Classical Inference

Situation

Situation

We consider a **random variable** X (discrete or continuous), whose **distribution** $\mathbf{P}_\theta$ depends on an unknown **parameter** θ. Two main goals:

- Estimate the unknown parameter θ using statistical inference.
- Assess the uncertainty of the estimate.

Bernoulli Distribution / Binomial Distribution

Let X be binary with possible values 0 or 1 and **probability function**

- $f(1) = \mathrm{P}(X = 1) = \pi$,
- $f(0) = \mathrm{P}(X = 0) = 1 - \pi$

or more compactly

$$f(x) = \begin{cases} \pi^x (1-\pi)^{1-x} & x \in \{0, 1\} \\ 0 & \text{else.} \end{cases}$$

Here $\theta = \pi = \mathrm{P}(X{=}1) = \mathrm{E}(X)$.

Connection to the Binomial distribution

Let $X_1, X_2, \ldots, X_n$ be i.i.d. **Bernoulli-distributed** random variables with $P(X_i=1) = \pi$, $P(X_i=0) = 1-\pi$;
Then

$$\begin{aligned} X &= X_1 + \cdots + X_n \\ &= \textit{'frequency with which an event E occurs'} \\ &\sim B(n, \pi) \end{aligned}$$

The probability function is given by

$$f(x) = P(X = x) = \begin{cases} \binom{n}{x} \pi^x (1-\pi)^{n-x} & x = 0, 1, \ldots, n \\ 0 & \text{else.} \end{cases}$$

Some properties of the Binomial distribution

a) Probability generating function

$G_x(t) = \mathrm{E}(t^x) = (1 - \pi + t\pi)^n$

b) Expected Value and Variance

$\mathrm{E}(X) = n\pi$

$\mathrm{Var}(X) = n\pi(1 - \pi)$

c) Sum

Let X, Y be independent with $X \sim B(n,\pi)$ and $Y \sim B(m,\pi)$. Then

$$G_{x+y}(t) = G_x(t)G_y(t) = (1 - \pi + t\pi)^{n+m}$$

This is the probability generating function of the $B(n+m, \pi)$ distribution,

i.e. $x+y \sim B(n+m, \pi)$.

Proof a)

$$
\begin{aligned}
G_X(t) = \mathsf{E}(t^X) &= \sum_{x=0}^{n} t^x \binom{n}{x} \pi^x (1-\pi)^{n-x} \\
&= \sum_{x=0}^{n} \binom{n}{x} (t\pi)^x (1-\pi)^{n-x} \\
&= (1 - \pi + t\pi)^n
\end{aligned}
$$

Use here

$$
(z+y)^n = \sum_{x=0}^{n} \binom{n}{x} y^x z^{n-x}
$$

Proof b)

$$G_X(t) = (1 - \pi + t\pi)^n$$

$$G'_x(t) = n\pi(1 - \pi + t\pi)^{n-1}$$

$$G''_x(t) = n\pi^2(n-1)(1 - \pi + t\pi)^{n-2}$$

$$\mathrm{E}(X) = G'_x(1) = n\pi(1 - \pi + \pi)^{n-1} = n\pi$$

$$\begin{aligned}
\mathrm{Var}(X) &= G''_x(1) + G'_x(1) - (G'_x(1))^2 \\
&= n\pi^2(n-1) + n\pi - n^2\pi^2 \\
&= n\pi(\pi(n-1) + 1 - n\pi) \\
&= n\pi(1 - \pi)
\end{aligned}$$

Proof c)

- Let X, Y be independent with $X \sim B(n, \pi)$ and $Y \sim B(m, \pi)$.
- Then
$$G_{X+Y}(t) = G_X(t)G_Y(t) = (1 - \pi + t\pi)^{n+m}.$$
- This is the probability generating function of the $B(n + m, \pi)$ distribution, i.e. $X + Y \sim B(n + m, \pi)$.

Poisson distribution

Consider the **Poisson distributed** random variable X with probability function

$$P_\lambda(X = x) = f_\lambda(x) = \begin{cases} \frac{\lambda^x}{x!} \exp(-\lambda) & x = 0, 1, 2, \ldots \\ 0 & \text{else} \end{cases}$$

We have $\mathrm{E}(X) = \lambda$, $\mathrm{Var}(X) = \lambda$ (see exercise). The goal is to estimate λ, *i.e.* $\theta = \lambda$.

Requirement for X being Poisson distributed

- Two events can not happen at the same time.
- The probability that an event happens within a small time interval of length Δt is approximately $\lambda \Delta t$.
- The probability that a certain number of events happen within a time interval depends on the length of the interval but not on the specific location at the time axis.
- The number of events in two disjunct time intervals are independent.

Examples

Some examples for Poisson distributions:

- *Radioactive decay*
  The duration for an atom nucleus to decay is **Exponentially distributed**. The radioactive half life is the period after that half of the atom nuclei are decayed on average. The number of decays per time unit is **Poisson distributed**.

- *Rice grains*
  Suppose we distribute a number N of rice grains randomly to a number of squares. Then the number of rice grains at the squares is (approximately) **Poisson distributed** with $\lambda = \frac{N}{n}$.

Normal distribution

Let $X \sim N(\mu,\sigma^2)$ with parameter μ and σ^2.

If σ^2 is known then

$$\theta = \mu,$$

otherwise we have

$$\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}.$$

Arbitrary distribution

Let X be an **arbitrary distribution** with existing mean $\mathrm{E}(X) = \mu$ and variance $\mathrm{Var}(X) = \sigma^2$.

The goal is to estimate μ (and possibly σ^2), i.e. $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$.

Special case: population

Suppose we are given a population of size N with elements $y_1, y_2, \ldots, y_N$. Aim is to estimate the population mean and variance

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i,$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2.$$

Then $X =$ '*randomly sampled element of the population*' is a random variable with

$$\mathrm{E}(X) = \frac{1}{N}(y_1 + \cdots + y_N) = \bar{y} = \mu$$

and

$$\mathrm{Var}(X) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2 = \sigma^2.$$

Random sample

In order to estimate the unknown parameter θ, we draw an i.i.d. (independent and identically distributed) random sample $X_1, X_2, \ldots, X_n$, *i.e.* the X_i, *i=1,. . . , n*, are independent and identically distributed as X.

To estimate θ, we use a so-called **statistic**

$$\hat{\theta} = T(X_1, \ldots, X_N),$$

which is an appropriate function of the sample.

Example

Bernoulli distribution

$$\begin{aligned}\hat{\pi} &= \textit{'relative frequency of ones'} \\ &= \frac{1}{n}(X_1 + \cdots + X_n) \\ &= \bar{X},\end{aligned}$$

i.e.

$$T(X_1, \ldots, X_n) = \bar{X}.$$

Example

Poisson distribution

$$\hat{\lambda} = \bar{X}$$

or (because λ=Var(X))

$$\hat{\lambda} = \tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

or

$$\hat{\lambda} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Example

Normal distribution

$$\hat{\mu} = \bar{X}$$

$$\hat{\sigma}^2 = \tilde{S}^2 \quad or \quad \hat{\sigma}^2 = S^2$$

Arbitrary distribution

$$\hat{\mu} = \bar{X}$$

$$\hat{\sigma}^2 = \tilde{S}^2 \quad or \quad \hat{\sigma}^2 = S^2$$

Example

Gambler's ruin

- Gambler plays a series of games against the casino,
- Bet 1€,
- Probability of winning the game $p = 0.52$,
- Gambler's stake 5€, casino's stake 50€.

Goal is to find the probability π of ruin!
Here we estimate the probability via simulation.
Simulate n game series and define

$$X_i = \begin{cases} 1 & \text{gambler is ruined in } i\text{-th series} \\ 0 & \text{gambler is not ruined.} \end{cases}$$

Example

Gambler's ruin (continued)

We have

$$X_i \sim \mathrm{B}(1, \pi), \quad \mathrm{P}(X_i = 1) = \pi, \quad \mathrm{P}(X_i = 0) = 1 - \pi.$$

Estimate π through

$$\hat{\pi} = \frac{1}{n}(X_1 + \cdots + X_n) = \bar{X}.$$

Question: How large should we choose n for the estimate to be precise enough?

Classical Inference

Methods of evaluating estimators

Biased and unbiased estimators

Consider the expected value $\mathrm{E}(\hat{\theta})$ of the estimator $\hat{\theta}$ for θ. The **estimator $\hat{\theta}$ is unbiased for θ** if

$$E(\hat{\theta}) = \theta$$

for all possible parameters θ. Otherwise the **estimator is biased** with

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

Frequency interpretation

If we repeat the estimation method several times with realized estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ then

$$\frac{1}{m}(\hat{\theta}_1 + \cdots + \hat{\theta}_m) \approx \theta.$$

Frequency interpretation

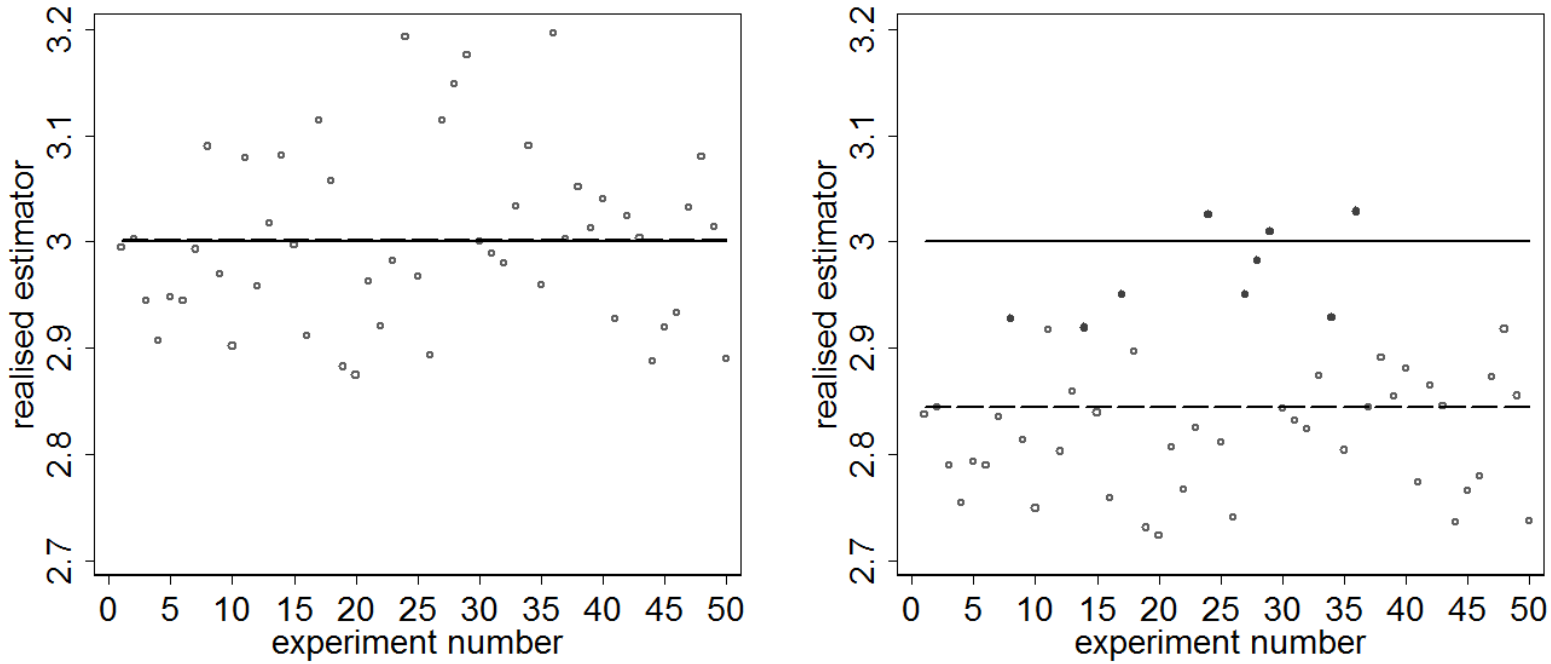


Figure: *Illustration of an unbiased (left) and a biased (right) estimation method.*

Example

Arbitrary distribution

We have

$$\begin{aligned}
\mathrm{E}(\bar{X}) &= \mathrm{E}\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) \\
&= \frac{1}{n}(\mathrm{E}(X_1) + \cdots + \mathrm{E}(X_n)) \\
&= \frac{1}{n}(\mu + \cdots + \mu) \\
&= \frac{1}{n} n\mu \\
&= \mu,
\end{aligned}$$

i.e. $\hat{\mu} = \bar{X}$ is unbiased for the expected value of a distribution.

Example

Arbitrary distribution (continued)

For the variance we have

$$\mathrm{E}(S^2) = \mathrm{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \sigma^2,$$

i.e. $\hat{\sigma}^2 = S^2$ is unbiased for σ^2 and $\hat{\sigma}^2 = \tilde{S}^2$ is biased for σ^2 as

$$\begin{aligned}
\mathrm{E}(\tilde{S}^2) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) \\
&= \mathrm{E}\left(\frac{n-1}{n(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) \\
&= \mathrm{E}\left(\frac{n-1}{n}S^2\right) = \frac{n-1}{n}\sigma^2 \neq \sigma^2.
\end{aligned}$$

Proof $\mathrm{E}(S^2)$

We have

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \mu)^2 &= \sum_{i=1}^{n}\Big((x_i - \bar{x}) + (\bar{x} - \mu)\Big)^2 \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - \mu) + \sum_{i=1}^{n}(\bar{x} - \mu)^2 \\
&= (n-1)s^2 + n(\bar{x} - \mu)^2
\end{aligned}$$

Proof continued

Note

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - \mu) &= (\bar{x} - \mu)\sum_{i=1}^{n}(x_i - \bar{x}) \\
&= (\bar{x} - \mu)\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\bar{x}\right) \\
&= (\bar{x} - \mu)\left(\sum_{i=1}^{n} x_i - n\bar{x}\right) \\
&= (\bar{x} - \mu)\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i\right) \\
&= 0
\end{aligned}$$

Proof continued

Applying expected values on both sides yields

$$\mathrm{E}\left(\sum_{i=1}^{n}(X_i-\mu)^2\right) = (n-1)\mathrm{E}(S^2) + n\underbrace{\mathrm{E}(\bar{X}-\mu)^2}_{\mathrm{Var}(\bar{X})}$$

Because of

$$\mathrm{E}\left(\sum_{i=1}^{n}(X_i-\mu)^2\right) = \sum_{i=1}^{n}\mathrm{E}(X_i-\mu)^2 = n\sigma^2$$

we obtain

$$n\sigma^2 = (n-1)\mathrm{E}(S^2) + n\mathrm{Var}(\bar{X}) = (n-1)\mathrm{E}(S^2) + \frac{n\sigma^2}{n}$$

and therefore

$$\mathrm{E}(S^2) = \sigma^2.$$

Remark

Note that $S = \sqrt{S^2}$ is not an unbiased estimator for σ! Indeed we have

$$\mathrm{E}(S) = \mathrm{E}(\sqrt{S^2}) < \sqrt{\sigma^2} = \sigma$$

because of Jensen's inequality *($\sqrt{\cdots}$ concave)*.

Comparison of unbiased estimators using the variance

Two (or more) unbiased estimators can be compared with the variance of the estimators. **An estimator $\hat{\theta}_1$ is at least as powerful as another estimator $\hat{\theta}_2$ if**

$$\text{Var}(\hat{\theta}_1) \leq \text{Var}(\hat{\theta}_2)$$

for all possible values of θ.

Frequency interpretation

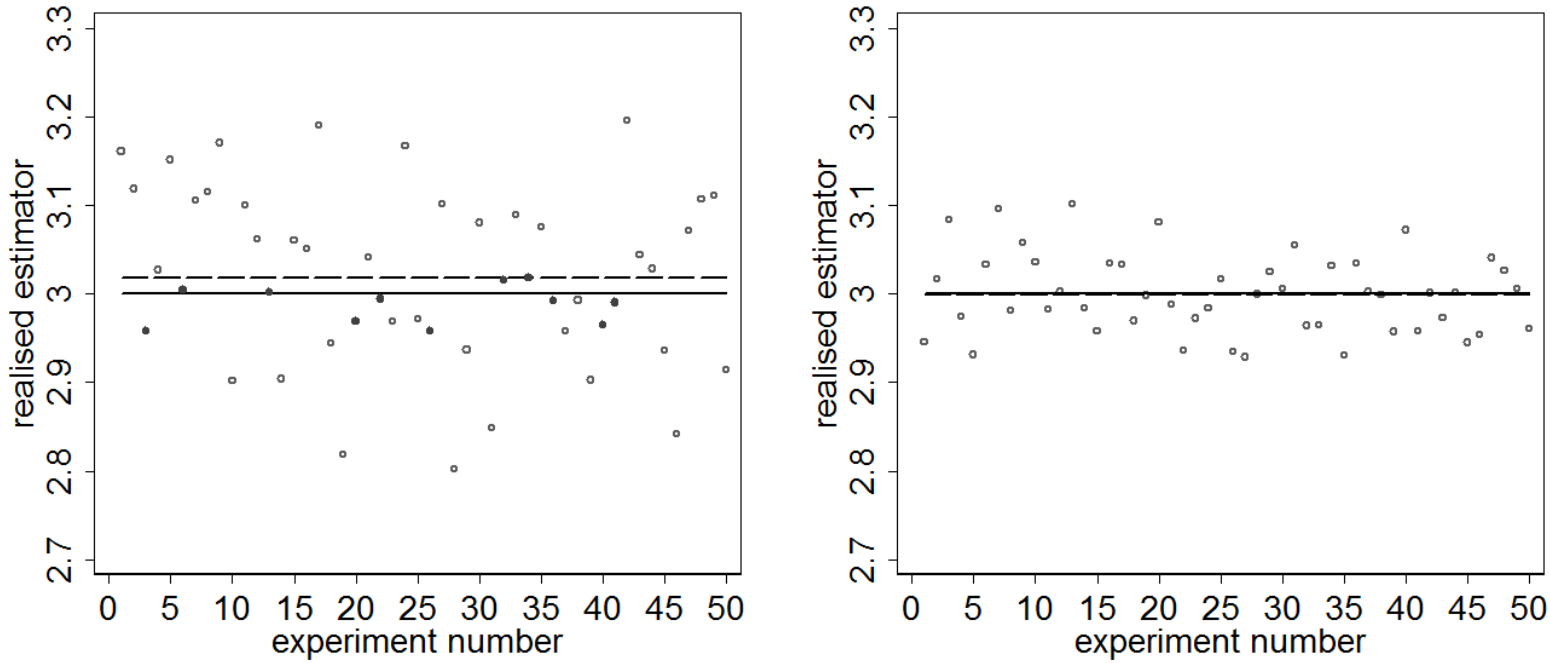


Figure: *Illustration of the variance of estimators.*

Example

Arbitrary distribution

The variance of $\hat{\mu}=\bar{X}$ is given by

$$\begin{aligned}
\text{Var}(\bar{X}) &= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} \text{Var}X_i\right) \\
&= \frac{1}{n^2} n\sigma^2 \\
&= \frac{\sigma^2}{n} = \frac{\text{Var}(X)}{n}
\end{aligned}$$

Example

Arbitrary distribution (continued)

In the Bernoulli case we have $\text{Var}(X) = \pi(1 - \pi)$ and therefore

$$\text{Var}(\hat{\pi}) = \text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n} = \frac{\pi(1-\pi)}{n}$$

Example

Arbitrary distribution (continued)

For Poisson distributed variables we obtain for $\hat{\lambda}_1 = \bar{X}$

$$\text{Var}(\hat{\lambda}_1) = \frac{\text{Var}(X)}{n} = \frac{\lambda}{n}.$$

For the competing estimator $\hat{\lambda}_2 = S^2$ one can show

$$\text{Var}(\hat{\lambda}_1) < \text{Var}(\hat{\lambda}_2).$$

Mean squared error (MSE)

The **mean squared error (MSE)** can be used to compare competing (not necessarily unbiased) estimators. It is defined as

$$\text{MSE}_\theta(\hat{\theta}) = \text{E}(\hat{\theta} - \theta)^2$$

Justification for the MSE

a) We have the decomposition

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{Bias}^2(\hat{\theta})$$

b) For arbitrary $\epsilon > 0$ we have

$$\text{P}\left(\left|\hat{\theta} - \theta\right| < \epsilon\right) > 1 - \frac{1}{\epsilon^2}\text{E}(\hat{\theta} - \theta)^2 = 1 - \frac{1}{\epsilon^2}\text{MSE}(\hat{\theta}),$$

i.e. the *'coverage probability'* becomes larger as $\text{MSE}(\hat{\theta})$ decreases!

Proof a)

$$
\begin{aligned}
\mathrm{MSE}_\theta(\hat{\theta}) &= \mathrm{E}(\hat{\theta} - \theta)^2 \\
&= \mathrm{E}\left[(\hat{\theta} - \mathrm{E}(\hat{\theta})) + (\mathrm{E}(\hat{\theta}) - \theta)\right]^2 \\
&= \mathrm{E}\left(\hat{\theta} - \mathrm{E}(\hat{\theta})\right)^2 + 2\mathrm{E}\underbrace{\left(\hat{\theta} - \mathrm{E}(\hat{\theta})\right)\underbrace{\left(\mathrm{E}(\hat{\theta}) - \theta\right)}_{\text{constant}}}_{0} + \mathrm{E}\left(\mathrm{E}(\hat{\theta}) - \theta\right)^2 \\
&= \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}^2(\hat{\theta})
\end{aligned}
$$

Proof b)

Define the random variable

$$V = \begin{cases} \epsilon^2 & \left|\hat{\theta} - \theta\right| \geq \epsilon \\ 0 & \text{else.} \end{cases}$$

Then

$$\mathrm{E}(V) = \epsilon^2 \, \mathrm{P}(\left|\hat{\theta} - \theta\right| \geq \epsilon)$$

and because of $V \leq (\hat{\theta} - \theta)^2$

$$\mathrm{E}(V) \leq \mathrm{E}(\hat{\theta} - \theta)^2.$$

Hence

$$\epsilon^2 \, \mathrm{P}(\left|\hat{\theta} - \theta)\right| \geq \epsilon) \leq \mathrm{E}(\hat{\theta} - \theta)^2 = \mathrm{MSE}(\hat{\theta})$$

and therefore

$$\mathrm{P}(\left|\hat{\theta} - \theta\right|) < \epsilon) = 1 - \mathrm{P}(\left|\hat{\theta} - \theta\right| \geq \epsilon) \geq 1 - \frac{1}{\epsilon^2} \mathrm{MSE}(\hat{\theta})$$

Frequency interpretation

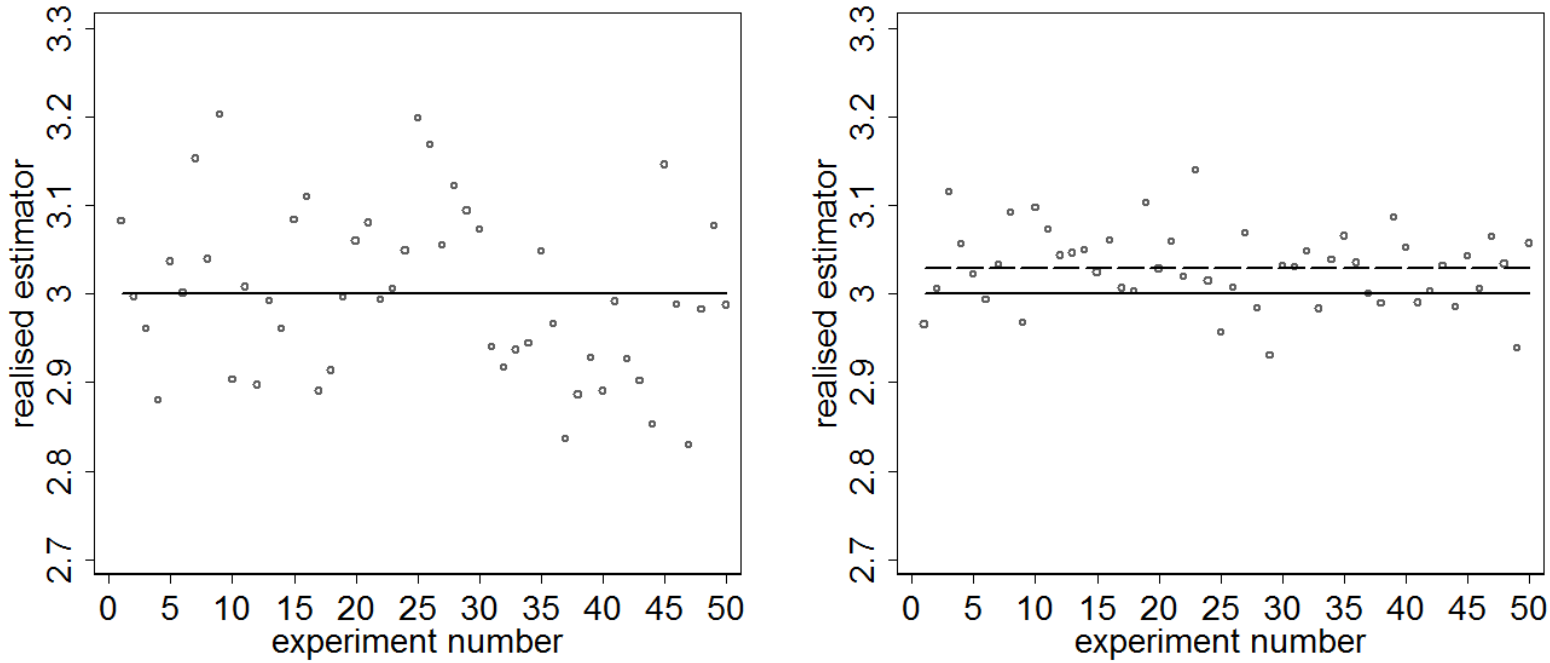


Figure: *Illustration of the MSE of estimators.*

Examples

Arbitrary distribution / normal distribution

$$\text{MSE}(\hat{\mu}) = \text{Var}(\hat{\mu}) = \frac{\sigma^2}{n}$$

Bernoulli distribution

$$\text{MSE}(\hat{\pi}) = \frac{\pi(1-\pi)}{n}$$

Poisson distribution

$$\text{MSE}(\hat{\lambda}_1) = \frac{\lambda}{n}$$

An unusual estimation problem

Suppose we are given a lake and we want to estimate the number N of fish in the lake. Estimation approach:

1. Draw a sample of size n_1, mark the fish and return the marked fish into the lake. Denote by

$$\pi = \frac{n_1}{N} \tag{1}$$

the share of marked fish

An unusual simulation problem (continued)

2. Draw a second sample of size n_2 and estimate the share of marked fish through

$$\hat{\pi} = \frac{X}{n_2},$$

where X is the number of marked fish in the sample. Insert $\hat{\pi}$ into (1) and solve for N to obtain

$$\hat{N} = \frac{n_1 n_2}{X}$$

(Although $\hat{\pi}$ is unbiased for π, $\hat{N}$ is not unbiased for N)

Case study: Doping - a rare phenomenon?

According to an article published in the “Süddeutsche Zeitung” on 23/12/2006:

Cool, cheerful, devotional, deceptive

A recent study has revealed that nearly every second German top athlete uses unauthorized doping substances.

“They used the well-known questioning technique RRT, which includes special instructions in addition to the normal questions. This method allows further probability calculations.”

“It was found out that 48.1% of all German top athletes are using doping.”

Initial situation

Sensitive yes-no questions

“Have you ever used unauthorized performance-enhancing drugs?”

“Have you ever made false statements in a tax declaration, for your own benefit?”

. . .

Objective

Estimate the number of doped athletes.

Estimate the number of taxpayers who have made false statements.

. . .

Sample

- Draw a random sample of $i = 1, \ldots, n$ persons.
- Define for each person the random variables

$$A_i = \begin{cases} 1 & i^{\text{th}} \text{ person answers with YES} \\ 0 & i^{\text{th}} \text{ person answers with NO} \end{cases}$$

and

$$R_i = \begin{cases} 1 & i^{\text{th}} \text{ person dopes} \\ 0 & i^{\text{th}} \text{ person does not dope} \end{cases}$$

- A_i and R_i follow a Bernoulli-distribution

$$\mathrm{P}(A_i = 1) = \pi_A = \mathrm{P}(\text{YES-answer}) \qquad \mathrm{P}(A_i = 0) = 1 - \pi_A$$

and

$$\mathrm{P}(R_i = 1) = \pi_R = \mathrm{P}(\text{Athlete dopes}) \qquad \mathrm{P}(R_i = 0) = 1 - \pi_R.$$

Honest answers

- Assuming that we have honest answers, the share π_A of YES-answers and the share π_R of doped athletes are equal, i.e. $\pi_A = \pi_R$.
- The relative frequency

$$\hat{\pi}_R = \bar{A} = \frac{1}{n}(A_1 + A_2 + \ldots + A_n)$$

  of YES-answers is an unbiased estimator for the share π_R of doped athletes.
- Expected value and variance of $\hat{\pi}_R = \bar{A}$ are given by

$$E(\bar{A}) = \pi_R \qquad Var(\bar{A}) = \frac{1}{n}\pi_R(1 - \pi_R).$$

Partially honest answers

- For partially false answers:

$$\begin{aligned} \mathrm{P}(\text{YES-answer} \,|\, \text{Athlete dopes}) &= \mathrm{P}(A_i = 1 | R_i = 1) = q < 1 \\ \mathrm{P}(\text{YES-answer} \,|\, \text{Athlete does not dope}) &= \mathrm{P}(A_i = 1 | R_i = 0) = 0. \end{aligned}$$

  hence

$$\mathrm{P}(\text{YES-answer}) = \mathrm{P}(A_i = 1) = \mathrm{P}(A_i = 1 | R_i = 1)\mathrm{P}(R_i = 1) = q \cdot \pi_R.$$

- Therefore the relative frequency $\bar{A}$ of YES-answers is a biased estimator for the share π_R of doped athletes with

$$E(\bar{A}) = q \cdot \pi_R < \pi_R.$$

  The share of doped athletes is underestimated!

Questioning techniques

- *Traditional interviews*
- *Questionnaire*
- *Computer-based questionnaire*
- *Randomized Response*

  Idea: Guarantee anonymous answers with the help of an additional experiment.

A simple RR-Model

- Every interviewed person tosses or rotates a fair coin, *unobserved* by the interviewer. (Note: a rotated 2 € coin is not fair!)
- If the coin shows “Head” the interviewed person should answer honestly.
- If the coin shows “Tail” the interviewed person must always answer with YES, independent of whether he/she has doped or not.
- The additional experiment guarantees the anonymity, since a YES-answer does not conclude that the person was doped.
- Assumption: The interviewees stick to the rules and answer correctly.

- Even if the answers are only partially honest, i.e. $\pi_A = q\pi_R < \pi_R$, we can make an unbiased estimate of the share π_A of YES-answers through the relative frequency $\bar{A}$, i.e. $\hat{\pi}_A = \bar{A}$.
- For the probability of YES-answers π_A the following is true

$$\pi_A = P(\text{YES-answer}) = \frac{1}{2}\pi_R + \frac{1}{2} \cdot 1 = \frac{1}{2}\pi_R + \frac{1}{2}$$

and therefore

$$\pi_R = 2\pi_A - 1$$

is true for the probability π_R of doping.

- Hence:

$$\hat{\pi}_R = 2\bar{A} - 1$$

is an unbiased estimator for the share of doped athletes.

- The variance is given by

$$\begin{aligned} Var(\hat{\pi}_R) &= \frac{1}{n}\pi_R(1-\pi_R) + \frac{1}{n}(1-\pi_R) \\ &= \text{Variance without randomization} + \frac{1}{n}(1-\pi_R) \end{aligned}$$

- In comparison to the situation of honest answers, the variance of the estimator increases!

The quality of RR in real life - Social insurance fraud

- A study of Heijden et al. (2000) concerning social insurance fraud compares RR, computer-based questionnaires (CASI) and traditional interviews.
- The characteristic is that only persons who evidently committed social insurance fraud were interviewed. The share of social insurance frauds in the sample is thus 100%! Neither the interviewees nor the interviewers were aware of this fact.
- Result: RR results in an estimated share of 43-49%, CASI leads to 19% and traditional interviews lead to 25%.
- RR significantly improves the results, nevertheless there is still a large share of dishonest answers.

Asymptotic properties of estimators

Random variables as mappings

Consider a random experiment with possible outcome in the **space** Ω and a corresponding **probability measure** P.

A **random variable** X **is a mapping** that assigns every $\omega \in \Omega$ a real value x, i.e. $X(\omega)=x$,
More specifically

$$X : \Omega \to \mathbb{R}$$
$$\omega \mapsto X(\omega) = x$$

Example

Rolling a dice twice

$$\Omega = \{(1,1),(1,2),\ldots,(6,6)\}$$

$$X((\omega_1,\omega_2)) = \omega_1 + \omega_2$$

e.g.

$$X((3,4)) = 3 + 4 = 7$$

$$\mathrm{P}(X = 3) = \mathrm{P}(\{(1,2),(2,1)\}) = \frac{2}{36},$$

More precise definition of a random variable

Let Ω be the set of possible outcomes of a random experiment.
An **event** is a subset of Ω.

In general it is not possible to assign probabilities to every subset of Ω in a consistant way. Instead we assign probabilities to a system of subsets F, called σ-field. A **σ-field F is a set of subsets of Ω** such that

i) $\Omega \in F$

ii) $A \in F \Rightarrow \bar{A} \in F$

iii) $A_1, A_2, \ldots$ in $F \Rightarrow \bigcup_{i=1}^{\infty} A_i \in F$

We call the **couple (Ω,F) a measurable space**.

More precise definition of a random variable (cont'd.)

A probability measure P is a mapping

$$P : F \rightarrow \left[0, 1\right]$$

such that

i) $P(\emptyset)=0$

ii) $P(\Omega)=1$

iii) $P(\bigcup_{i=1}^{\infty} A_i)= \sum_{i=1}^{\infty} P(A_i)$ *for sets A; with* $A_i \cap A_j = \emptyset$ *for all i, j*.

The **space (Ω, F, P) is called probability space**.

More precise definition of a random variable (cont'd.)

Let (Ω, F, P) be a probability space and (Ω', F') a measurable space.
A **random variable** X is a F-F'-measurable function

$$X : \Omega \to \Omega'.$$

X **is called measurable** if

$$X^{-1}(A') \in F$$

for all $A' \in F'$.
Measurability is important because for $A' \in F'$:

$$P(X \in A') = P(X^{-1}(A')).$$

Convergence concepts

We consider in the following a sequence of random variables

$$X_1, X_2, X_3, \ldots$$

and investigate the limit behavior for $n \to \infty$.

Example

Arbitrary distribution

Consider the estimator

$$\hat{\mu} = \bar{X}$$

in dependence of n:

$$\begin{aligned}
\bar{x}_1 &= x_1 \\
\bar{x}_2 &= \frac{1}{2}(x_1 + x_2) \\
\bar{x}_3 &= \frac{1}{3}(x_1 + x_2 + x_3) \\
&\vdots \\
\bar{x}_n &= \frac{1}{n}\sum_{i=1}^{n} x_i
\end{aligned}$$

Almost sure convergence

A sequence X_n of random variables **converges almost surely to a random variable** X if,

$$P(\lim_{n \to \infty} X_n = X) = 1.$$

The definition states that $X_n(\omega)$ converges to $X(\omega)$ for all $\omega \in \Omega$, except perhaps for $\omega \in N$ where $N \subset \Omega$ and $P(N) = 0$.

Notation: $X_n \xrightarrow{\text{a.s.}} X$

Example

Ω = [0,1] and P the uniform distribution. Define

$$X_n(\omega) = \omega + \omega^n$$

and

$$X(\omega) = \omega.$$

We show that the sequence X_n converges to X almost surely.

Example (continued)

- For all $\omega \in [0, 1)$ we have $\omega^n \to 0$ for $n \to \infty$ and therefore

$$X_n(\omega) = \omega + \omega^n \to \omega = X(\omega)$$

.

- For ω=1 we have $X_n(1) = 2$ for all n, such that $X_n(1)$ does not converge to $X(1) = 1$.
- Since $P([0, 1)) = 1$, X_n converges almost surely to X.

Convergence in probability

X_n **converges to** X **in probability**, if for $\epsilon > 0$

$$\lim_{n \to \infty} P(\left|X_n - X\right| \geq \epsilon) = 0$$

or

$$\lim_{n \to \infty} P(\left|X_n - X\right| < \epsilon) = 1.$$

Notation: $X_n \xrightarrow{\text{P}} X$.

Example

Weak law of large numbers

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with finite expected value μ and finite variance σ^2.
Then

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \xrightarrow{\text{P}} \mu \qquad \text{for} \quad n \to \infty$$

($\bar{X}_n \xrightarrow{\text{a.s.}} \mu$ can be established as well. Then we speak of the strong law of large numbers).
The law says, that for large n

$$\left| \bar{x}_n - \mu \right| < \epsilon$$

with high probability.

Example (continued)

Proof: Weak law of large numbers

- Using the Tschebyschov we obtain

$$P(\left|\bar{X} - \mu\right| \geq \epsilon) \geq \frac{\text{Var}\bar{X}}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0 \qquad n \to \infty$$

- Tschebyschov inequality: Let X be a random variable with expected value μ and variance σ^2. For $\epsilon > 0$ the inequality

$$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

or equivalently

$$P(|X - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{\epsilon^2}$$

holds.

Example

Convergence in probability, not almost sure

Let $\Omega = [0,1]$ and P is the uniform distribution. Define $X_1, X_2, \ldots$ as follows:

$$
\begin{aligned}
X_1(\omega) &= \omega + \mathrm{I}_{[0,1]}(\omega) \\
X_2(\omega) &= \omega + \mathrm{I}_{[0,\frac{1}{2}]}(\omega) \\
X_3(\omega) &= \omega + \mathrm{I}_{[\frac{1}{2},1]}(\omega) \\
X_4(\omega) &= \omega + \mathrm{I}_{[0,\frac{1}{3}]}(\omega) \\
X_5(\omega) &= \omega + \mathrm{I}_{[\frac{1}{3},\frac{2}{3}]}(\omega) \\
X_6(\omega) &= \omega + \mathrm{I}_{[\frac{2}{3},1]}(\omega) \\
&\vdots
\end{aligned}
$$

Define

$$X(\omega) = \omega.$$

Example (continued)

Convergence in probability, not almost sure

- X_n converges to X in probability as

$$P(\left|X_n - X\right| \geq \epsilon)$$

  is the probability of an interval of ω values whose length is going to 0.

- X_n does not converge to X almost surely as there is no $\omega \in \Omega$ for which

$$X_n(\omega) \to \omega = X(\omega).$$

Example (continued)

Convergence in probability, not almost sure

- For every ω, $X_n(\omega)$ alternates between ω and $\omega + 1$.
- For example $\omega = \frac{3}{8}$ yields

$$
\begin{aligned}
X_1(\frac{3}{8}) &= 1\frac{3}{8} \\
X_2(\frac{3}{8}) &= 1\frac{3}{8} \\
X_3(\frac{3}{8}) &= \frac{3}{8} \\
X_4(\frac{3}{8}) &= \frac{3}{8} \\
&\vdots
\end{aligned}
$$

No pointwise convergence occurs for this sequence.

Convergence in the r-th mean

X_n converges to X in the r-th mean, if

$$\mathrm{E}(\left|X_n^r\right|) < \infty \qquad \text{for all } n$$

and

$$\lim_{n\to\infty} \mathrm{E}(\left|X_n - X\right|^r) = 0.$$

For r=2 we say that **X_n converges in mean square to X**.

Notation: $X_n \xrightarrow{\mathrm{r}} X$.

Example

Arbitrary distribution

$$\mathrm{E}(\hat{\mu}) = \mathrm{E}(\bar{X}) = \mu$$

$$\mathrm{Var}(\hat{\mu}) = \frac{\sigma^2}{n}$$

$$\mathrm{MSE}(\hat{\mu}) = \mathrm{E}(\hat{\mu} - \mu)^2 = \frac{\sigma^2}{n} \to 0 \qquad \textit{with } n \to \infty,$$

i.e. $\hat{\mu}$ converges in mean square to μ.

Convergence in distribution

X_n converges to X in distribution, if

$$\lim_{n \to \infty} P(X_n \leq x) = P(X \leq x)$$

or

$$\lim_{n \to \infty} F_{X_n}(x) = F_x(x)$$

at all points x where $F_x(x)$ is continuous.

Remark

Note that if

$$f_n \to f \qquad \text{with} \qquad n \to \infty,$$

where f_n, f are the probability functions, then the distributions defined through f_n converge to the distribution defined through f.

Reverse is not correct, in general: **Convergence in distribution does not imply that the densities converge.**

Example

Convergence of the Binomial distribution to the Poisson distribution

Let $X_n \sim B(n,\pi)$ with probability function

$$f(x) = P(X = x) = \binom{n}{x} \pi^x (1-\pi)^{n-x}.$$

For $n \to \infty$ and $n\pi = \lambda$ $f(x)$ converges to the probability function of the Poisson distribution, i.e.

$$\lim_{n \to \infty} f(x) = \frac{\lambda^x exp(-\lambda)}{x!}.$$

Example (continued)

Convergence of the Binomial distribution to the Poisson distribution

$$\begin{aligned}
\lim_{n\to\infty} f(x) &= \lim_{n\to\infty} \binom{n}{x} \pi^x (1-\pi)^{n-x} \\
&= \lim_{n\to\infty} \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \\
&= \lim_{n\to\infty} \frac{\lambda^x}{x!} \underbrace{\frac{n(n-1)\cdots(n-x+1)}{n\cdots n}}_{\to 1} \underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\to \exp(-\lambda)} \underbrace{\left(1-\frac{\lambda}{n}\right)^{-x}}_{\to 1}
\end{aligned}$$

We have used

$$exp(y) = \lim_{n\to\infty} \left(1+\frac{y}{n}\right)^n = \sum_{n=o}^{\infty} \frac{y^n}{n!}$$

Application of the limit theorem

If

- $X \sim B(n, \pi)$,
- n large,
- π 'small', i.e. $\lambda = n\pi$ 'moderate' (rule of thumb $n > 30$, $\pi \leq 0.05$),

then X can be approximated by the Poisson distribution with parameters $\lambda = n\pi$.

Example

Rice grains

Distribute n rice grains randomly to N squares. Let

$$X = \text{'Number of rice grains in a square'}.$$

We have

$$X \sim B\left(n, \pi = \frac{1}{N}\right)$$

Because of $n > 30$, $\pi < 0.05$ we have

$$X \overset{a}{\sim} Po\left(\lambda = n\pi = \frac{n}{N}\right).$$

Example

The sentence *'at all x for which $F_x(x)$ is continuous'* matters!
Let

$$X_n \sim N\left(0, \frac{1}{n}\right)$$

and X a degenerated distribution at 0, i.e. $P(X{=}0) = 1$.

Central Limit Theorem

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with finite mean μ and variance σ^2.
Then

$$\frac{1}{\sqrt{n}\sigma}\left(\sum_{i=1}^{n} X_i - n\mu\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i - \mu}{\sigma} = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \quad \xrightarrow{\text{D}} \quad N(0, 1).$$

Standardizing random variables

Assume X is a random variable mit expected value μ and variance σ^2. Then

$$Y = \frac{X - \mu}{\sigma} = \frac{X}{\sigma} - \frac{\mu}{\sigma}$$

is the standardized version of X with

$$\mathrm{E}(Y) = \mathrm{E}\left(\frac{X}{\sigma} - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma}\mathrm{E}(X) - \frac{\mu}{\sigma} = \frac{1}{\sigma}\mu - \frac{\mu}{\sigma} = 0$$

and

$$\mathrm{Var}(Y) = \mathrm{Var}\left(\frac{X}{\sigma} - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma^2}\mathrm{Var}(X) = \frac{1}{\sigma^2}\sigma^2 = 1$$

Standardizing random variables

Note that

$$\mathsf{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathsf{E}(X_i) = \sum_{i=1}^{n} \mu = n\mu$$

and

$$\mathsf{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathsf{Var}(X_i) = \sum_{i=1}^{n} \sigma^2 = n\sigma^2$$

such that

$$\frac{1}{\sqrt{n}\sigma}\left(\sum_{i=1}^{n} X_i - n\mu\right)$$

is the standardized version of the random variable

$$\sum_{i=1}^{n} X_i.$$

Utilities for the proof

For a proof of the central limit theorem we need the following facts from analysis:

Exponential limit:

If a sequence a_n converges to a, i.e. $a_n \longrightarrow a$, then

$$\lim_{n \to \infty} \left(1 + \frac{a_n}{n}\right)^n = exp(a) \tag{1}$$

Utilities for the proof

Taylor series expansion:

Let I be a real valued interval and $f : \mathrm{I} \longrightarrow \mathbb{R}$ a function that is $r + 1$ times continuously differentiable. Then we obtain in a neighborhood around $a \in \mathrm{I}$

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(r)}(a)}{r!}(x-a)^r + R(x),$$

where $R(x)$ is a function with

$$lim_{x \longrightarrow a} \frac{R(x)}{(x-a)^r} = 0. \tag{2}$$

Proof of the central limit theorem

- Proof in case that the moment generating function $M_{X_i}(t)$ exists in a neighborhood around 0. In the general case the proof is similar and based on the so called characteristic function (which is always defined).
- The proof uses property a) (uniqueness) of the moment generating function of random variables (slide 16 of the probability theory slides). We show that the moment generating function of

$$Z_n := \frac{\bar{X}_n - \mu}{\sigma}\sqrt{n}$$

converges for $n \to \infty$ to the moment generating function of the $N(0, 1)$ distribution.

1. Step: Standardizing

- Define the standardized random variable $Y_i = \frac{X_i - \mu}{\sigma}$ and let $M_Y(t)$ be the moment generating function of Y_i.
- We now define

$$Z_n := \frac{\bar{X}_n - \mu}{\sigma}\sqrt{n} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i$$

  and obtain the moment generating function of Z_n:

$$M_{Z_n}(t) = \left(M_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

- Here we used properties c) (with $a = 1/\sqrt{n}$, $b = 0$) and e) of the moment generating function (slide 16, probability theory slides).
- Note that $M_Y{}^{(r)}(0) = E(Y^r)$ (property b) page 16 of the slides). This implies $M_Y(0) = 1$, $M_Y{}'(0) = E(Y) = 0$, $M_Y{}''(0) = E(Y^2) = Var(Y) = 1$.

2. Step: Taylor series expansion

- A second order Taylor series expansion of $M_Y\left(\frac{t}{\sqrt{n}}\right)$ around 0 yields:

$$
\begin{aligned}
M_Y\left(\frac{t}{\sqrt{n}}\right) &= M_Y(0) + \frac{M_Y'(0)}{1!}\frac{t}{\sqrt{n}} + \frac{M_Y''(0)}{2!}\left(\frac{t}{\sqrt{n}}\right)^2 + R_Y\left(\frac{t}{\sqrt{n}}\right) \\
&= 1 + \frac{0}{1!}\frac{t}{\sqrt{n}} + \frac{1}{2!}\left(\frac{t}{\sqrt{n}}\right)^2 + R_Y\left(\frac{t}{\sqrt{n}}\right) \\
&= 1 + \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + R_Y\left(\frac{t}{\sqrt{n}}\right)
\end{aligned}
$$

2. Step: Taylor series expansion

- For fixed $t \neq 0$ we have $\frac{t}{\sqrt{n}} \to 0$ and because of (2)

$$\lim_{n \to \infty} \frac{R_Y\left(\frac{t}{\sqrt{n}}\right)}{\left(\frac{t}{\sqrt{n}}\right)^2} = 0$$

- Since $t \neq 0$ is fixed we also obtain

$$\lim_{n \to \infty} \frac{R_Y\left(\frac{t}{\sqrt{n}}\right)}{\left(\frac{1}{\sqrt{n}}\right)^2} = \lim_{n \to \infty} n\, R_Y\left(\frac{t}{\sqrt{n}}\right) = 0$$

- This statement is also valid for $t = 0$ because $R_Y\left(\frac{0}{\sqrt{n}}\right) = 0$.

3. Step: Limit

- For fixed t we finally obtain

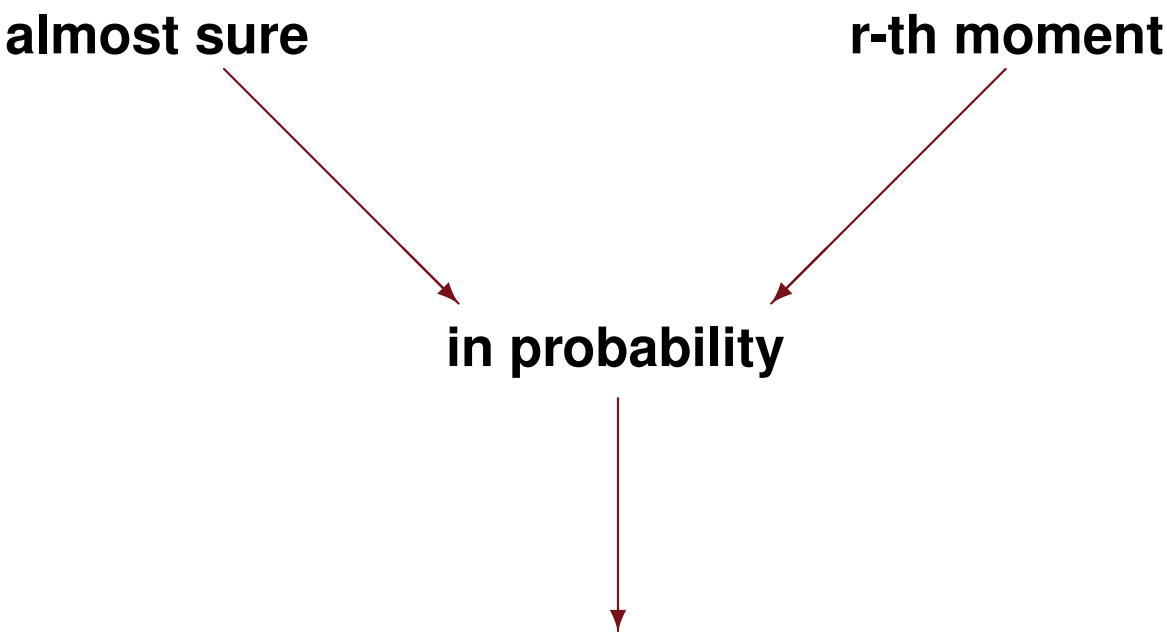
$$\begin{aligned}
\lim_{n\to\infty}\left(M_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n &= \lim_{n\to\infty}\left[1+\frac{\left(t/\sqrt{n}\right)^2}{2!}+R_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n \\
&= \lim_{n\to\infty}\left[1+\frac{1}{n}\left(\frac{t^2}{2}+n\,R_Y\left(\frac{t}{\sqrt{n}}\right)\right)\right]^n \\
&= \exp\left(\frac{t^2}{2}\right)
\end{aligned}$$

- Thereby we used the exponential limit (1) from above with

$$a_n = \frac{t^2}{2}+n\,R_Y\left(\frac{t}{\sqrt{n}}\right)\to\frac{t^2}{2}.$$

- Since $\exp\left(\frac{t^2}{2}\right)$ is the moment generating function of the $N(0,1)$ distribution the theorem is proven.

Relationship among modes of convergence



$$X_n \xrightarrow{D} c \quad \Rightarrow \quad X_n \xrightarrow{p} c$$

Construction of confidence intervals I

Let $X_1, \cdots, X_n$ be an i.i.d sample with $X_i \sim N(\mu, \sigma^2)$ and σ^2 known.

Then

$$\left[\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$$

is a $1 - \alpha$ **confidence interval** for μ.

Proof

We have

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

respectively

$$\frac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim N(0, 1). \tag{3}$$

Hence for $0<\alpha<1$ we obtain

$$P\big(-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \leq z_{1-\frac{\alpha}{2}}\big) = 1 - \alpha,$$

where $z_{1-\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ Quantile of the $N(0,1)$ distribution.

Proof (continued)

Rearranging terms yields

$$P(\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha,$$

such that

$$\left[\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$$

is a 1-α confidence interval for μ.
Because of the CLT(3) and with it the CI is valid for arbitrary distributions provided that n is large enough (rule of thumb $n \geq 30$).

Special Case: Approximation of the Binomial Distribution

Let $X \sim B(n, \pi)$. X can be represented as $X = X_1 + \cdots + X_n$, where

$$X_i \in \{0, 1\}, \quad P(X_i = 1) = \pi, \quad P(X_i = 0) = 1 - \pi$$

Applying the CLT we obtain

$$\frac{X - nE(X_i)}{\sqrt{n \operatorname{Var}(X_i)}} = \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}} \overset{\text{D}}{\sim} N(0, 1)$$

Thus approximately we have

$$X \overset{\text{a}}{\sim} N\big(n\pi, n\pi(1-\pi)\big)$$

$$P(X \leq x) \approx \phi\left(\frac{x - n\pi}{\sqrt{n\pi(1-\pi)}}\right).$$

Unknown variance

So far we have only considered the case where the variance σ^2 is known.

To construct **confidence intervals for unknown σ^2** we would like to replace

$$\frac{\bar{X} - \mu}{\sigma}\sqrt{n} \quad \overset{\text{D}}{\sim} \quad N(0, 1)$$

by

$$\frac{\bar{X} - \mu}{S}\sqrt{n} \quad \overset{\text{D}}{\sim} \quad N(0, 1).$$

Continuous mapping theorem

If $g : \mathbb{R} \to \mathbb{R}$ is a continuous function, then

$$\textbf{(i)} \quad X_n \xrightarrow{a.s.} X \quad \Rightarrow \quad g(X_n) \xrightarrow{a.s.} g(X)$$

$$\textbf{(ii)} \quad X_n \xrightarrow{p} X \quad \Rightarrow \quad g(X_n) \xrightarrow{p} g(X)$$

$$\textbf{(iii)} \quad X_n \xrightarrow{d} X \quad \Rightarrow \quad g(X_n) \xrightarrow{d} g(X)$$

Sums and products

(a) $X_n \xrightarrow{p} X,\ Y_n \xrightarrow{p} Y \Rightarrow X_n + Y_n \xrightarrow{p} X + Y$

(b) $X_n \xrightarrow{r} X,\ Y_n \xrightarrow{r} Y \Rightarrow X_n + Y_n \xrightarrow{r} X + Y$

(c) $X_n \xrightarrow{d} X,\ Y_n \xrightarrow{d} a \Rightarrow X_n + Y_n \xrightarrow{d} X + a$ (Slutzky)

(d) $X_n \xrightarrow{p} X,\ Y_n \xrightarrow{p} Y \Rightarrow X_n \cdot Y_n \xrightarrow{p} X \cdot Y$

(e) $X_n \xrightarrow{d} X,\ Y_n \xrightarrow{d} a \Rightarrow X_n \cdot Y_n \xrightarrow{d} a \cdot X$ (Slutzky)

Example

Sampling variance

$$
\begin{aligned}
S^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \\
&= \frac{n}{n-1} \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \\
&= \frac{n}{n-1} \left[\frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}^2 \right]
\end{aligned}
$$

We show

$$
S^2 \xrightarrow{\text{P}} \sigma^2.
$$

Proof

- We have
$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{\text{P}} E(X^2)$$
because of the weak law of large numbers.
- As $\bar{X} \xrightarrow{\text{P}} \mu$ the continuous mapping theorem implies $\bar{X}^2 \xrightarrow{\text{P}} \mu^2$.
- Then
$$\frac{1}{n}\sum X_i^2 - \bar{X}^2 \xrightarrow{\text{P}} E(X^2) - \mu^2 = \sigma^2$$
applying slide 96 a).
- Finally as $n/(n-1) \to 1$ we obtain
$$S^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right] \xrightarrow{\text{P}} \sigma^2.$$

Construction of confidence intervals II

Let $X_1, X_2, \ldots$ be an i.i.d. sample with finite mean μ and variance σ^2.

We show that

$$\frac{\bar{X} - \mu}{S}\sqrt{n} \quad \overset{\text{D}}{\sim} \quad N(0, 1).$$

Proof

- Since $S^2 \xrightarrow{\text{P}} \sigma^2$ we have $S \xrightarrow{\text{P}} \sigma$ according to the continuous mapping theorem.
- This implies (using slide 96 d)

$$\frac{1}{S} \xrightarrow{\text{P}} \frac{1}{\sigma} \quad \text{and} \quad \frac{\sigma}{S} \xrightarrow{\text{P}} 1.$$

- According to Slutsky we obtain

$$\frac{\bar{X} - \mu}{S}\sqrt{n} = \underbrace{\frac{\sigma}{S}}_{\xrightarrow{\text{P}} 1} \underbrace{\frac{\bar{X} - \mu}{\sigma}\sqrt{n}}_{\xrightarrow{\text{D}} N(0,1)} \xrightarrow{\text{D}} N(0,1)$$

- This result implies the following CI for μ and large n:

$$\left[\bar{X} - z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$$

Asymptotically unbiased estimators

An **estimator $\hat{\theta}$ for θ is asymptotically unbiased**, if for all θ

$$\lim_{n \to \infty} E(\hat{\theta}) = \theta$$

Obviously every unbiased estimator $\hat{\theta}$ is also asymptotically unbiased.

Example

Arbitrary distribution

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

is unbiased for σ^2.

$$\hat{\sigma}^2 = \tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

is however biased. Since

$$E(\tilde{S}^2) = E\left(\frac{n-1}{n}S^2\right) = \frac{n-1}{n}\sigma^2$$

$\tilde{S}^2$ is at least asymptotically unbiased.

Consistency of an estimator

An **estimator $\hat{\theta}$ for θ is called (weakly) consistent** , if $\hat{\theta}$ converges on probability to θ, i.e.

$$\hat{\theta} \xrightarrow{\mathsf{P}} \theta \qquad\qquad n \to \infty$$

(*strong consistency means convergence almost surely*).

An **estimator $\hat{\theta}$ is consistent in square mean**, if $\hat{\theta}$ converges to θ in square mean, i.e.

$$\hat{\theta} \xrightarrow{r=2} \theta.$$

Example

Consistency of $\bar{X}$ for μ

$X_1, X_2, \ldots$ i.i.d. with finite expected value μ and variance σ^2.

As $\bar{X} \xrightarrow{\text{P}} \mu$ (*law of large numbers*), $\hat{\mu}=\bar{X}$ is consistent for μ.

It follows that

$$\hat{\pi} = \bar{X}$$

and

$$\hat{\lambda} = \bar{X}$$

are consistent for π in the Binomial case and λ in the Poisson case.

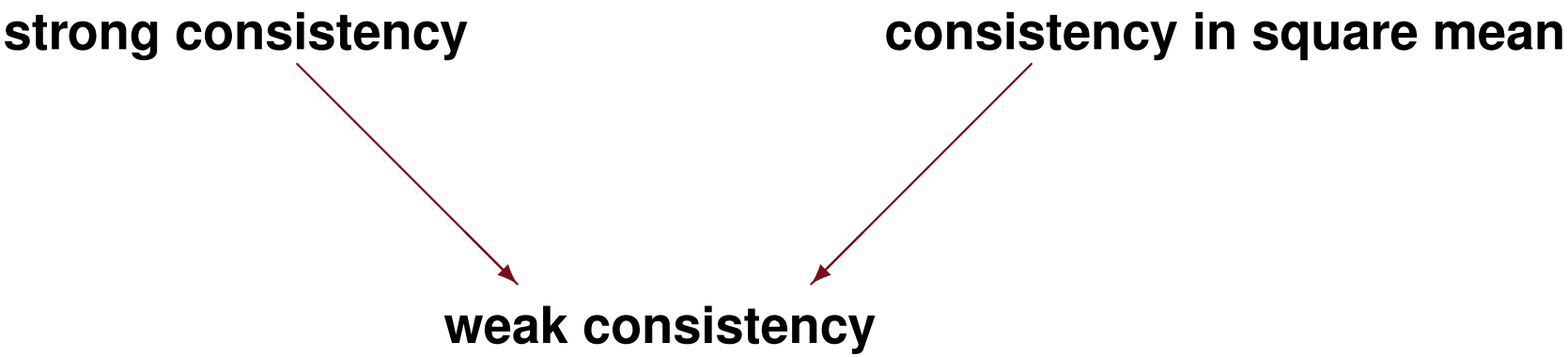
Example

Consistency of the variance estimator

We have $S^2 \xrightarrow{\text{P}} \sigma^2$. It follows that S^2 is consistent for σ^2.

Because of the continuous mapping theorem $S \xrightarrow{\text{P}} \sigma$, so that S is consistent for σ.

Relationships between the various consistency concepts



If the *MSE* converges to 0, then $\hat{\theta} \xrightarrow{r=2} \theta$, because $MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2$. This implies consistency in square mean and in turn weak consistency. Note that **every consistent estimator is asymptotically unbiased.**

Note also, that **the consistency of $\hat{\theta}$ for θ implies the consistency of $g(\hat{\theta})$ for $g(\theta)$**, if g is continuous (*because of the continuous mapping theorem*).

Standard errors

Consider an i.i.d. random sample $X_1, \ldots, X_n$.
Let $\hat{\theta} = T(X_1, \ldots, X_n)$ be an estimator for θ. Let V be a consistant estimator for $Var(T)$.

Then $\sqrt{V}$ is consistent for $\sqrt{Var(T)}$.

The quantity

$$se(\hat{\theta}) = \sqrt{V}$$

is called **standard error of $\hat{\theta}$**.

Example

Arbitrary distribution

$X_1, \ldots, X_n$ i.i.d. random sample with finite expected value μ and variance σ^2. Then

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

and

$$V = \frac{S^2}{n}$$

is consistent for $Var(\bar{X})$. Therefore

$$se(\hat{\mu}) = \frac{S}{\sqrt{n}}.$$

Statistical Inference

Chapter 3

Likelihood based inference

Likelihood based inference

Likelihood and log-likelihood

Likelihood and log-likelihood

We throw a coin $n = 10$ times and observe $y = 8$ times tail.

Which value for

$$\pi = \textit{'Probability of tail'}$$

is most likely?

Let X='Number of tails', then

$$X \sim B(10, \pi).$$

Likelihood and log-likelihood (continued)

The following table provides $P_\pi(X{=}8)$ in dependence of π:

π	$P_\pi(X{=}8)$
0.1	0
0.2	0.00007373
⋮	⋮
0.7	0.23347444
0.8	0.3019898
0.9	0.1937
1	0

Likelihood and log-likelihood (continued)

The **most plausible value for the parameter** π is $\pi = 0.8$!

Likelihood theory uses the estimator $\hat{\pi}$ **that maximizes the function**

$$L(\pi) = P_\pi(X = 8)$$

with respect to π.

$L(\pi)$ is denoted as likelihood. In likelihood theory, all inferential conclusions are based on the likelihood.

Definition: Likelihood

Let x **be the realized value** of a random variable X. Let $f(x, \theta)$ **be the corresponding probability function** (if X is discrete) respectively density (if X is continuous).

Then

$$L(\theta) = f(x, \theta)$$

is called likelihood (function).

Definition: Maximum likelihood estimator

The **maximum likelihood estimator** $\hat{\theta}_{\text{ML}}$ of a parameter θ is given by maximizing the likelihood:

$$\hat{\theta}_{\text{ML}} = \underset{\theta}{\text{argmax}}\, L(\theta).$$

Usually **it is more convenient to maximize the log-likelihood**

$$\ell(\theta) = \log L(\theta).$$

Because of the monotonicity of the logarithm we have

$$\hat{\theta}_{\text{ML}} = \underset{\theta}{\text{argmax}}\, L(\theta) = \underset{\theta}{\text{argmax}}\, \ell(\theta).$$

Example

Binomial distribution

Let $X \sim B(n, \pi)$ and x be the realized value.
Then

$$L(\pi) = \binom{n}{k} \pi^x (1-\pi)^{n-x} \propto \pi^x (1-\pi)^{n-x}$$

$$l(\pi) = x\log(\pi) + (n-x)\log(1-\pi)$$

$$l'(\pi) = \frac{x}{\pi} + \frac{(n-x)}{1-\pi}(-1) = \frac{x}{\pi} - \frac{n-x}{1-\pi}$$

Example (continued)

Binomial distribution

Setting to zero yields

$$\frac{x}{\pi} - \frac{n-x}{1-\pi} = 0$$

and therefore

$$\hat{\pi}_{\text{ML}} = \frac{x}{n}.$$

Likelihood and ML-estimator for a random sample

Consider an **i.i.d. sample $X_1,\ldots,X_n$ with probability function or density of X_i given by** $f(x_i,\theta)$. If

$$x = (x_1, \ldots, x_n)'$$

is the realized sample of

$$X = (X_1, \ldots, X_n)',$$

then because of the independence we have

$$L(\theta) = f(x_1, \theta) \cdot f(x_2, \theta) \cdot \ \cdots \ \cdot f(x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

and

$$\ell(\theta) = \log f(x_1, \theta) + \cdots + \log f(x_n, \theta) = \sum_{i=1}^{n} \log f(x_i, \theta).$$

Example

Poisson distribution

Let $X_1, \ldots, X_n$ be a random sample with $X_i \sim Po(\lambda)$.
The realizations are given by $x_1, x_2, \ldots, x_n$.
The probability function for X_i is given by

$$f(x_i) = P(X_i = x_i) = \frac{\lambda^{x_i}}{x_i!}\exp(-\lambda).$$

Example (continued)

Poisson distribution

$$L_i(\lambda) = \frac{\lambda^{x_i}}{x_i!}\exp(-\lambda) \propto \lambda^{x_i}\exp(-\lambda)$$

$$l_i(\lambda) = \log\left(\lambda^{x_i}\exp(-\lambda)\right) = x_i\log(\lambda) - \lambda$$

$$l_i'(\lambda) = \frac{x_i}{\lambda} - 1$$

$$l'(\lambda) = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n$$

Setting to zero and solving for λ yields

$$\hat{\lambda}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

Score function and Fisher information

Definition: Score function

The first derivative of the log-likelihood

$$s(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \ell'(\theta)$$

is called **score function**.

Definition: Fisher information

The negative second derivative of the log-likelihood

$$\mathrm{I}(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \theta^2} = -\frac{\partial \mathrm{S}(\theta)}{\partial \theta} = -\ell''(\theta)$$

is called **Fisher information**.

Evaluating the Fisher information at the maximum $\hat{\theta}_{\mathrm{ML}}$ yields the **observed Fisher information**.

First interpretation of the (observed) Fisher information

At the maximum $\hat{\theta}_{\text{ML}}$ the second derivative of the log-likelihood is negative and therefore $\mathrm{I}(\hat{\theta}_{\text{ML}})$ positive.

The larger $\mathrm{I}(\hat{\theta}_{\text{ML}})$, the larger is the curvature of the log-likelihood, i.e. the more *'information'* to locate the maximum.

Example

Binomial distribution

Let $X \sim B(n, \pi)$.

$$l(\pi) = x \log(\pi) + (n - x) \log(1 - \pi)$$

$$S(\pi) = l'(\pi) = \frac{x}{\pi} - \frac{n - x}{1 - \pi}$$

$$\mathrm{I}(\pi) = -l''(\pi) = -\left(-\frac{x}{\pi^2} + \frac{n - x}{(1 - \pi)^2}(-1)\right) = \frac{x}{\pi^2} + \frac{n - x}{(1 - \pi)^2}$$

$$\mathrm{I}(\hat{\pi}) = \frac{x}{(\frac{x}{n})^2} + \frac{n - x}{\left(1 - (\frac{x}{n})\right)^2} = \cdots = \frac{n}{\hat{\pi}(1 - \hat{\pi})}$$

Example

Poisson distribution

$X_1, \ldots, X_n$ i.i.d. Poisson distributed with $X_i \sim Po(\lambda)$ and realizations $x_1, \ldots, x_n$.

$$S_i(\lambda) = l_i'(\lambda) = \frac{x_i}{\lambda} - 1$$

$$S(\lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n\lambda$$

$$\mathrm{I}_i(\lambda) = -l_i''(\lambda) = \frac{x_i}{\lambda^2}$$

$$\mathrm{I}(\lambda) = \frac{1}{\lambda^2} \sum_{i=1}^{n} x_i$$

$$\mathrm{I}(\hat{\lambda}) = \frac{n^2}{\left(\sum_{i=1}^{n} x_i\right)^2} \sum_{i=1}^{n} x_i = \frac{n^2}{\sum_{i=1}^{n} x_i} = \frac{n}{\bar{x}}$$

Definition: Expected Fisher information

The expected value of the Fisher information $\mathrm{I}(\theta)$, considered as a function of the sample variables $X=(X_1,\ldots,X_n)$', is called **expected Fisher information**:

$$J(\theta) = \mathrm{E}\big(\mathrm{I}(\theta)\big).$$

Note that $J(\theta)$ is additive for an i.i.d. sample with density $f(x,\theta)$.
Let $J_1(\theta)$ be the expected Fisher information of x_i. Then $J(\theta) = nJ_1(\theta)$.

Example

Binomial distribution

Let $X \sim B(n, \pi)$. Then

$$I(\pi) = \frac{X}{\pi^2} + \frac{n - X}{(1-\pi)^2}$$

$$\begin{aligned} J(\pi) &= E\left(\frac{X}{\pi^2} + \frac{n - X}{(1-\pi)^2}\right) \\ &= \frac{n\pi}{\pi^2} + \frac{n - n\pi}{(1-\pi)^2} = \frac{n}{\pi} + \frac{n}{1-\pi} \\ &= \frac{n(1-\pi) + n\pi}{\pi(1-\pi)} = \frac{n}{\pi(1-\pi)} \\ &= \frac{1}{\text{Var}(\hat{\pi})}, \end{aligned}$$

Expected value and variance of the score function

Expected value and variance

Under regularity conditions we show

$$E(S(\theta)) = 0 \text{ and } Var(S(\theta)) = J(\theta).$$

Regularity conditions

Exchangeability of Differentiation and Integration (or Differentiation and Summation).

For example fulfilled for fixed integral limits, f and $\frac{\partial f}{\partial t}$ continuous, then

$$\frac{\partial}{\partial t} \int_a^b f(x, t) \, \mathrm{d}x = \int_a^b \frac{\partial f(x, t)}{\partial t} \, \mathrm{d}x$$

Proof (for continuous distributions)

- Because
$$\mathsf{S}(\theta) = \sum_{i=1}^{n} \mathsf{S}_i(\theta),$$
it is sufficient to show $\mathsf{E}(\mathsf{S}_i(\theta)) = 0$ and we can restrict ourselves to $n = 1$, then $L(\theta) = f(x, \theta)$.
- We have the following equivalent representations of the score function:
$$\mathsf{S}(\theta) = \frac{\partial l(\theta)}{\partial \theta} = \frac{\partial \log L(\theta)}{\partial \theta} = \frac{1}{L(\theta)} \frac{\partial L(\theta)}{\partial \theta}$$

Proof (continued)

$$
\begin{aligned}
\mathrm{E}(\mathrm{S}(\theta)) &= \int \mathrm{S}(\theta) f(x, \theta) \, \mathrm{d}x \\
&= \int \frac{1}{\mathrm{L}(\theta)} \frac{\partial \mathrm{L}(\theta)}{\partial \theta} f(x, \theta) \, \mathrm{d}x \\
&= \int \frac{\partial \mathrm{L}(\theta)}{\partial \theta} \, \mathrm{d}x \\
&= \frac{\partial}{\partial \theta} \left(\int \mathrm{L}(\theta) \, \mathrm{d}x \right) \\
&= \frac{\partial}{\partial \theta} (1) \\
&= 0
\end{aligned}
$$

Proof (continued)

Since $\mathrm{E}(S(\theta)) = 0$ we have $\mathrm{Var}(S(\theta)) = \mathrm{E}(S(\theta)^2)$. Then

$$
\begin{aligned}
J(\theta) &= \mathrm{E}\left(-\frac{\partial^2 \log \mathrm{L}(\theta)}{\partial \theta^2}\right) = \mathrm{E}\left(-\frac{\partial}{\partial \theta}\left(\underbrace{\frac{\partial \mathrm{L}(\theta)}{\partial \theta}}_{\hat{=} f}\underbrace{\frac{1}{\mathrm{L}(\theta)}}_{\hat{=} g}\right)\right) \\
&= \mathrm{E}\left(-\frac{\frac{\partial^2 \mathrm{L}(\theta)}{\partial \theta^2}\mathrm{L}(\theta) - \left(\frac{\partial \mathrm{L}(\theta)}{\partial \theta}\right)^2}{\mathrm{L}(\theta)^2}\right) = -\mathrm{E}\left(\frac{\frac{\partial^2 \mathrm{L}(\theta)}{\partial \theta^2}}{\mathrm{L}(\theta)}\right) + \mathrm{E}\left(\frac{\left(\frac{\partial \mathrm{L}(\theta)}{\partial \theta}\right)^2}{\mathrm{L}(\theta)^2}\right) \\
&= -\int \frac{\frac{\partial^2 \mathrm{L}(\theta)}{\partial \theta^2}}{\mathrm{L}(\theta)} f(x, \theta)\, \mathrm{d}x + \int \frac{\left(\frac{\partial \mathrm{L}(\theta)}{\partial \theta}\right)^2}{\mathrm{L}(\theta)^2} f(x, \theta)\, \mathrm{d}x \\
&= -\frac{\partial^2}{\partial \theta^2}\underbrace{\int \mathrm{L}(\theta)\, \mathrm{d}x}_{=1} + \int \left(\frac{\partial}{\partial \theta} \log \mathrm{L}(\theta)\right)^2 f(x, \theta)\, \mathrm{d}x \\
&= \mathrm{E}\big(\mathrm{S}(\theta)^2\big) = \mathrm{Var}\big(\mathrm{S}(\theta)\big)
\end{aligned}
$$

Proof (continued)

We thereby used the following rules regarding derivatives

$$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$$

and

$$\frac{\partial}{\partial\theta}\log L(\theta) = \frac{\frac{\partial L(\theta)}{\partial\theta}}{L(\theta)}$$

Cramer-Rao bound

Cramer-Rao bound

Consider an unbiased estimator $\hat{\theta}$ for θ, i.e.

$$E(\hat{\theta}) = \theta.$$

Let $J(\theta)$ be the expected Fisher information. Then we have

$$Var(\hat{\theta}) \geq \frac{1}{J(\theta)} \qquad \textbf{(Cramer-Rao bound)}$$

under regularity conditions, i.e. exchangeability of integration and differentiation.

Application of the Cramer-Rao bound

If we can show in applications, that

$$Var(\hat{\theta}) = \frac{1}{J(\theta)},$$

i.e. the bound is reached, then $\hat{\theta}$ is the estimator with lowest variance among all unbiased estimators.

Then $\hat{\theta}$ is called **efficient**.

Proof

Let

$$f(\vec{x}, \theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

Then the squared correlation between $S(\theta)$ and $\hat{\theta}$ is

$$\rho^2\left(\mathrm{S}(\theta), \hat{\theta}\right) = \frac{\mathrm{Cov}\left(\mathrm{S}(\theta), \hat{\theta}\right)^2}{\mathrm{Var}\left(\mathrm{S}(\theta)\right)\mathrm{Var}\left(\hat{\theta}\right)} \leq 1$$

Proof (continued)

Because of $\text{Var}(S(\theta)) = J(\theta)$ we have:

$$\frac{\text{Cov}\left(\text{S}(\theta), \hat{\theta}\right)^2}{J(\theta)\text{Var}(\hat{\theta})} \leq 1.$$

Rearranging yields

$$\text{Var}(\hat{\theta}) \geq \frac{\text{Cov}\left(\text{S}(\theta), \hat{\theta}\right)^2}{J(\theta)}$$

It remains to show that $Cov(S(\theta), \hat{\theta}) = 1$.

Proof (continued)

$$
\begin{aligned}
\mathrm{Cov}\big(\mathrm{S}(\theta), \hat{\theta}\big) &= \mathrm{E}\big(\mathrm{S}(\theta)\hat{\theta}\big) - \underbrace{\mathrm{E}\big(\mathrm{S}(\theta)\big)}_{=0}\,\mathrm{E}\big(\hat{\theta}\big) \\
&= \int \mathrm{S}(\theta)\hat{\theta} f(\vec{x}, \theta)\,\mathrm{d}\vec{x} = \int \frac{\partial f(\vec{x}, \theta)}{\partial \theta} \frac{1}{f(\vec{x}, \theta)} \hat{\theta} f(\vec{x}, \theta)\,\mathrm{d}\vec{x} \\
&= \int \frac{\partial f(\vec{x}, \theta)}{\partial \theta} \hat{\theta}\,\mathrm{d}\vec{x} \\
&= \frac{\partial}{\partial \theta} \int f(\vec{x}, \theta)\hat{\theta}\,\mathrm{d}\vec{x} \\
&= \frac{\partial}{\partial \theta} \Big(\mathrm{E}(\hat{\theta}) \Big) \\
&= \frac{\partial}{\partial \theta} (\theta) \\
&= 1
\end{aligned}
$$

Example

Binomial distribution

$$X \sim B(n, \pi).$$

Then

$$\hat{\pi} = \frac{X}{n},$$
$$J(\pi) = \frac{n}{\pi(1-\pi)}.$$

Finally

$$\text{Var}(\hat{\pi}) = \frac{n\pi(1-\pi)}{n^2} = \frac{\pi(1-\pi)}{n} = \frac{1}{J(\pi)},$$

i.e. $\hat{\pi}$ reaches the Cramer-Rao lower bound and is therefore efficient.

Distribution of the score statistic

Distribution of the score statistic

Let $X_1, \ldots, X_n$ be a random sample with density $f(x, \theta)$. Then

$$\frac{\mathrm{S}(\theta)}{\sqrt{J(\theta)}} \xrightarrow{\mathrm{D}} N\big(0, 1\big) \tag{1}$$

i.e. approximately we have

$$\mathrm{S}(\theta) \overset{\mathrm{a}}{\sim} N\big(0, J(\theta)\big).$$

The proof is a direct application of the CLT

$$\frac{S(\theta)}{\sqrt{J(\theta)}} = \frac{\sum S_i(\theta) - 0}{\sqrt{J(\theta)}} \xrightarrow{\mathrm{D}} N\big(0, 1\big)$$

Remark

In (1) the expected Fisher information can be replaced by the (observed) Fisher information, i.e.

$$\frac{\mathrm{S}(\theta)}{\sqrt{\mathrm{I}(\theta)}}, \qquad \frac{\mathrm{S}(\theta)}{\sqrt{J(\hat{\theta}_{\mathrm{ML}})}}, \qquad \frac{\mathrm{S}(\theta)}{\sqrt{\mathrm{I}(\hat{\theta}_{\mathrm{ML}})}}.$$

Example

Poisson distribution

$$T = \frac{S(\lambda)}{\sqrt{J(\lambda)}} = \frac{\frac{\sum x_i}{\lambda} - n}{\sqrt{\frac{n}{\lambda}}} \overset{a}{\sim} N(0, 1).$$

Application: Score Test

Aim is to test the hypothesis

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0$$

Use the **test statistic**

$$\frac{S(\theta_0)}{\sqrt{J(\theta_0)}} \overset{a}{\sim} N(0,1) \qquad \text{(under } H_0\text{)}.$$

Application: Score Test (continued)

If $\left|S(\theta_0)\right|$ large, then the difference between θ_0 and the ML-estimator $\hat{\theta}_{ML}$ is large (note that $S(\hat{\theta}_{ML})=0$), i.e. we reject H_0.

More precisely we reject H_0 if

$$\frac{\left|S(\theta_0)\right|}{\sqrt{J(\theta_0)}} > z_{1-\frac{\alpha}{2}}.$$

Note that the computation of the ML-estimator is not required to use the score test.

Example

Poisson distribution

$$H_0 : \lambda = \lambda_0 \qquad \text{versus} \qquad H_1 : \lambda \neq \lambda_0$$

Use

$$T = \frac{\frac{\sum x_i}{\lambda_0} - n}{\sqrt{\frac{n}{\lambda_0}}} \overset{a}{\sim} N(0, 1)$$

as the test statistic.

Recap Hypotheses testing

- Aim is to test the hypothesis

$$\mathrm{H}_0 : \theta = \theta_0 \qquad \text{versus} \qquad \mathrm{H}_1 : \theta \neq \theta_0.$$

- Two sources of error:
  - H_0 is true, but we decide for H_1 (type I error)
  - H_1 is true, but we decide to keep H_0 (type II error)
- Test is constructed asymmetrically such that type 1 error is under our control, e.g. $\alpha = 0.05$. Type II error is not controlled directly, but can be reduced usually through sample size.
- Test decision is based on a test statistics whose distribution under H_0 must be known. Often distribution is not known under H_1, then type II error is not available/computable.

Recap Hypotheses testing

- Caution I:
  - A nonsignificant result, i.e. H_0 is kept, does not necessarily imply that H_0 is true!
  - Suppose we toss a coin 4 times and obtain 3 times head. The aim is to test $H_0 : \pi = 0.5$ versus $H_1 : \pi \neq 0.5$ where π is the probability of head.
  - Obviously any reliable test will not reject H_0 on basis of the given data.
  - However, already our intuition says, that the nonsignificant results does not imply that H_0 is true.
- Caution II:
  Think of the power of the test. Is there a real chance for a significant result with the test at hand?

Distribution of the ML-estimator

Distribution of the ML-estimator

Let $X_1, \ldots, X_n$ be an i.i.d random sample with density (or probability function) $f(x, \theta)$ and $\hat{\theta}_{\text{ML}}$ the ML-estimator for θ. Let θ_0 be the true parameter.

Then under regularity conditions the **ML-estimator is consistent, i.e.**

$$\hat{\theta}_{ML} \xrightarrow{\text{P}} \theta_0$$

and

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{\text{d}} N\left(0, \frac{1}{J_1(\theta_0)}\right)$$

or

$$(\hat{\theta}_{ML} - \theta_0)\sqrt{nJ_1(\theta_0)} = (\hat{\theta}_{ML} - \theta_0)\sqrt{J(\theta_0)} \xrightarrow{\text{d}} N(0, 1),$$

i.e. we obtain approximately

$$\hat{\theta}_{ML} \overset{\text{a}}{\sim} N(\theta_0, J(\theta_0)^{-1})$$

Remarks I

The asymptotic results suggest the following statistical properties of the maximum likelihood estimator:

- $\hat{\theta}_{ML}$ is weakly consistent, i.e. $\hat{\theta}_{ML} \xrightarrow{P} \theta_0$.
- Consistency implies that $\hat{\theta}_{ML}$ is asymptotically unbiased. Note however, that $\hat{\theta}_{ML}$ is generally not unbiased in finite samples!
- For large n the approximate variance of $\hat{\theta}_{ML}$ is

$$\text{Var}(\hat{\theta}_{ML}) \approx \frac{1}{J(\theta_0)},$$

  i.e. the ML estimator reaches the Cramer- Rao bound as n tends to infinity.
- This in turn implies that $\hat{\theta}_{ML}$ is asymptotically efficient, i.e. among all asymptotically unbiased estimators it is the estimator with the lowest variance.

Remarks II

- In $\hat{\theta}_{ML} \overset{a}{\sim} N\big(\theta_0, J(\theta_0)^{-1}\big)$ we can replace $J(\theta_0)$ by $I(\theta_0)$.
- Because $\hat{\theta}_{ML}$ is consistent for θ_0 we also have

$$\hat{\theta}_{ML} \overset{a}{\sim} N\big(\theta_0, J(\hat{\theta}_{ML})^{-1}\big) \quad \text{and} \quad \hat{\theta}_{ML} \overset{a}{\sim} N\big(\theta_0, I(\hat{\theta}_{ML})^{-1}\big).$$

  For the construction of tests and CI's this is important because $J(\hat{\theta}_{ML})$ does not depend on the unknown parameter.
- The second remark yields approximate standard errors

$$se\big(\hat{\theta}_{ML}\big) \approx \sqrt{\frac{1}{J(\hat{\theta}_{ML})}}$$

or

$$se\big(\hat{\theta}_{ML}\big) \approx \sqrt{\frac{1}{\mathrm{I}(\hat{\theta}_{ML})}}$$

Regularity conditions for consistency

R1: The parameter is identifiable, i.e. if $\theta \neq \theta'$ then $f(x,\theta) \neq f(x,\theta')$.

R2: The densities $f(x,\theta)$ have common support and $f(x,\theta)$ is differentiable in θ.

R3: The parameter space Θ contains an open set of which the true parameter is an interior point.

Additional regularity conditions for asymptotic normality

R4: The density $f(x,\theta)$ is three times differentiable with respect to θ, the third derivative is continuous in θ and $\int f(x,\theta)dx$ can be differentiated three times under the integral sign.

R5: For any θ_0, there exists a positive number c and a function $M(x)$ (both of which may depend on θ_0) such that

$$\left| \frac{\partial^3 \log f(x,\theta)}{\partial \theta^3} \right| \leq M(x)$$

for all x and $\theta_0 - c < \theta < \theta_0 + c$ with $E(M(x)) < \infty$.

Proof of asymptotic normality

- A second order Taylor series expansion of the score function at the ML-estimator $\hat{\theta}_{ML}$ about the true value θ_0 yields

$$s\left(\hat{\theta}_{ML}\right) = s(\theta_0) + \left(\hat{\theta}_{ML} - \theta_0\right) l''(\theta_0) + \frac{1}{2}\left(\hat{\theta}_{ML} - \theta_0\right)^2 l'''(\theta^*) \quad (3)$$

  with θ^* between θ_0 and $\hat{\theta}_{ML}$.
- Using $s(\hat{\theta}_{ML}) = 0$ and rearranging (3) yields

$$(\hat{\theta}_{ML} - \theta_0)\sqrt{n} = \frac{\frac{s(\theta_0)}{\sqrt{n}}}{-\frac{1}{n} l''(\theta_0) - \frac{1}{2n}(\hat{\theta}_{ML} - \theta_0) l'''(\theta^*)}. \quad (4)$$

Now we can derive the following three properties:

1. Because of the asymptotic properties of the score function we have

$$\frac{s(\theta_0)}{\sqrt{n}} \xrightarrow{d} N\left(0, J_1(\theta_0)\right).$$

2. For $-l''(\theta_0)/n$ the central limit theorem yields

$$-\frac{l''(\theta_0)}{n} = \frac{I(\theta_0)}{n} = \frac{1}{n}\sum_{i=1}^{n} I_i(\theta_0) \xrightarrow{d} J_1(\theta_0). \tag{5}$$

3. The second summand in the denominator tends in probability to 0 as $n \to \infty$, i.e.

$$\frac{1}{2n}(\hat{\theta}_{ML} - \theta_0) l'''(\theta^*) \xrightarrow{p} 0. \tag{6}$$

This is true because $\hat{\theta}_{ML} - \theta_0 \xrightarrow{p} 0$ (consistency of the ML-estimator) and $1/n\, l'''(\theta^*)$ is bounded in probability. The latter is due to regularity condition R5. More specifically,

$$\left|\frac{1}{n} l'''(\theta^*)\right| = \left|\frac{1}{n}\sum_{i=1}^{n} l_i'''(\theta^*)\right| \leq \frac{1}{n}\sum_{i=1}^{n} M(X_i) \xrightarrow{d} E(M(X)) < \infty.$$

- Properties 2 and 3 imply that the denominator in (4) converges to $J_1(\theta_0)$ in probability, while property 1 shows that the numerator converges in distribution to $N(0, J_1(\theta_0))$.
- Slutzky's theorem shows that the ratio converges in distribution to $N(0, 1/J_1(\theta_0))$, which completes the proof.

Approximate Tests and confidence intervals

Wald-statistic

- We investigate the hypotheses

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0$$

- For testing we use

$$T_1 = \left(\hat{\theta}_{ML} - \theta_0\right) \sqrt{I\left(\hat{\theta}_{ML}\right)}$$

or

$$T_2 = \left(\hat{\theta}_{ML} - \theta_0\right) \sqrt{J\left(\hat{\theta}_{ML}\right)}.$$

Wald-statistic

- Under H_0 we have

$$T_1 \overset{a}{\sim} N(0,1)$$
$$T_2 \overset{a}{\sim} N(0,1)$$

- Rejection:

  We reject H_0 if

$$\left|T_1\right| > z_{1-\frac{\alpha}{2}}$$

  respectively

$$\left|T_2\right| > z_{1-\frac{\alpha}{2}}.$$

Example

Poisson distribution

$$\mathsf{H}_0 : \lambda = \lambda_0 \qquad \text{versus} \qquad \mathsf{H}_1 : \lambda \neq \lambda_0$$

The various variants of the Fisher information are given by

$$\mathsf{I}(\lambda) = \frac{n\bar{X}}{\lambda^2} \qquad J(\lambda) = \frac{n}{\lambda}$$

$$\mathsf{I}(\hat{\lambda}_{\mathsf{ML}}) = \frac{n\bar{X}}{\bar{X}^2} = \frac{n}{\bar{X}} \qquad J(\hat{\lambda}_{\mathsf{ML}}) = \frac{n}{\bar{X}}.$$

Then

$$\mathsf{T}_1 = \mathsf{T}_2 = (\bar{X} - \lambda_0)\sqrt{\frac{n}{\bar{X}}} \quad \overset{\text{a}}{\sim} \quad N(0, 1).$$

Example

Binomial distribution

$$H_0 : \pi = \pi_0 \qquad \text{versus} \qquad H_1 : \pi \neq \pi_0$$

$$J(\pi) = \frac{n}{\pi(1-\pi)}, \quad J(\hat{\pi}) = \frac{n}{\bar{X}(1-\bar{X})}, \quad \bar{X} = \frac{X}{n}$$

$$T_2 = (\hat{\pi} - \pi_0)\sqrt{\frac{n}{\hat{\pi}(1-\hat{\pi})}}$$

Wald confidence intervals

Because of the duality between tests and confidence intervals we obtain the following approximate confidence intervals:

$$\left[\hat{\theta}_{\text{ML}} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{I(\hat{\theta}_{\text{ML}})}}\right] = \left[\hat{\theta}_{\text{ML}} \pm z_{1-\frac{\alpha}{2}} \text{se}(\hat{\theta}_{\text{ML}})\right]$$

respectively

$$\left[\hat{\theta}_{\text{ML}} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{J(\hat{\theta}_{\text{ML}})}}\right]$$

Example

Poisson distribution

An approximate confidence interval is given by

$$\left[\bar{X} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{X}}{n}}\right].$$

Example

Binomial distribution

An approximate confidence interval is given by

$$\left[\hat{\pi} \pm z_{1-\frac{\alpha}{2}} \frac{\hat{\pi}(1-\hat{\pi})}{n}\right]$$

as

$$J(\hat{\pi}) = \frac{n}{\bar{X}(1-\bar{X})}.$$

Likelihood ratio test

Likelihood ratio test

- Our goal is again to investigate the hypotheses

$$\mathsf{H}_0 : \theta = \theta_0 \quad \text{versus} \quad \mathsf{H}_1 : \theta \neq \theta_0$$

- The idea of the likelihood ratio test is to compare $\ell(\hat{\theta}_{\mathsf{ML}})$ with $\ell(\theta_0)$. If the difference $\ell(\hat{\theta}_{\mathsf{ML}}) - \ell(\theta_0)$ is "large", then θ_0 is not plausible and H_0 is rejected.
- The exact test statistic is

$$\mathsf{W} = 2\left(\ell(\hat{\theta}_{\mathsf{ML}}) - \ell(\theta_0)\right) = 2\log\frac{\mathsf{L}(\hat{\theta}_{\mathsf{ML}})}{\mathsf{L}(\theta_0)}.$$

- Under H_0 we have

$$W \overset{a}{\sim} \chi^2_1$$

and we reject H_0 if $W > \chi^2_1(1-\alpha)$.

Example

Poisson distribution

$$H_0 : \lambda = \lambda_0 \qquad \text{versus} \qquad H_1 : \lambda \neq \lambda_0$$

$$l(\lambda) = \sum_{i=1}^{n} X_i \log(\lambda) - n\lambda = n\bar{X} \log(\lambda) - n\lambda$$

$$W = 2\big(l(\hat{\lambda}) - l(\lambda_0)\big) = 2\big(n\bar{X} \log \bar{X} - n\bar{X} - n\bar{X} \log(\lambda_0) + n\lambda_0\big)$$

Derivation of the asymptotic distribution of W

- To derive the asymptotic distribution of W under H_0 we compute the first order Taylor series expansion of the log-likelihood about θ_0 at $\hat{\theta}_{ML}$

$$l(\hat{\theta}_{ML}) = (\theta_0) + \left(\hat{\theta}_{ML} - \theta_0\right) l'(\theta_0) + \frac{1}{2}\left(\hat{\theta}_{ML} - \theta_0\right)^2 l''\left(\theta^{**}\right) \quad (7)$$

  with θ^{**} between $\hat{\theta}_{ML}$ and θ_0.

- For the derivation of the asymptotic distribution of the ML-estimator we also computed the second order Taylor series expansion of $l'(\hat{\theta}_{ML}) = s(\hat{\theta}_{ML})$ given by (3). Because of $l'(\hat{\theta}_{ML}) = 0$ we can rearrange (3) to obtain

$$-l'(\theta_0) = \left(\hat{\theta}_{ML} - \theta_0\right) l''(\theta_0) + \frac{1}{2}\left(\hat{\theta}_{ML} - \theta_0\right)^2 l'''(\theta^*).$$

- Inserting this into (7) yields

$$\begin{aligned} l\left(\hat{\theta}_{ML}\right) - l\left(\theta_0\right) &= -\left(\hat{\theta}_{ML} - \theta_0\right)\left[\left(\hat{\theta}_{ML} - \theta_0\right) l''(\theta_0) + \tfrac{1}{2}\left(\hat{\theta}_{ML} - \theta_0\right)^2 l'''(\theta^*)\right] \\ &\quad + \tfrac{1}{2}\left(\hat{\theta}_{ML} - \theta_0\right) l''(\theta^{**}) \\ &= -n\left(\hat{\theta}_{ML} - \theta_0\right)^2 \left[\tfrac{1}{n} l''(\theta_0) + \tfrac{1}{2n}(\hat{\theta}_{ML} - \theta_0) l'''(\theta^*) - \tfrac{1}{2n} l''(\theta^{**})\right] \end{aligned}$$

- We already know from (6) that the second summand in brackets tends to 0 in probability.
- The other two summands in brackets both tend to $-J_1(\theta_0)$ (see the derivation of (5)). Hence $2(l(\hat{\theta}_{ML}) - l(\theta_0))$ has the same limit distribution as $\left(\hat{\theta}_{ML} - \theta_0\right)^2 n J_1(\theta_0)$. The asymptotic normality of the ML-estimator provides

$$\left(\hat{\theta}_{ML} - \theta_0\right) \sqrt{n J_1(\theta_0)} \xrightarrow{d} N(0, 1).$$

- Invoking the continuous mapping theorem we finally arrive at

$$\left(\hat{\theta}_{ML} - \theta_0\right)^2 n J_1(\theta_0) \xrightarrow{d} \chi^2_1.$$

(Note that if X i.i.d. $N(0, 1)$ then $X^2 \sim \chi^2_1$.)

Universität Innsbruck

Statistical Inference

Chapter 4

Likelihood inference for vector valued parameters

Likelihood inference for vector valued parameters

Situation

Situation

- Let $X_1,...,X_n$ be a random sample with probability function or density $f_i(x_i,\theta)$.
- The X_i's are still independent but no longer identically distributed.
- Our goal is to estimate θ.

Example

(Normal distribution)

The random sample $X_1, ..., X_n$ is assumed to be i.i.d. with $X_i \sim N(\mu, \sigma^2)$, i.e.

$$\boldsymbol{\theta} = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}.$$

Example

Linear Model

- $Y_1,...,Y_n$ independent with $Y_i \sim N(\mu_i,\sigma^2)$ and

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}.$$

- Define $\theta = \left(\beta_0, \beta_1, \ldots, \beta_k, \sigma^2\right)'$.
- We assume $X_{i1},...,X_{ik}$ non stochastic.

Example

Binary regression models

- We assume an independent random sample $Y_1, ..., Y_n$ with $Y_i \sim B(1, \pi_i)$ being Bernoulli distributed, i.e. $P(Y_i = 1) = \pi_i$ and $P(Y_i = 0) = 1 - \pi_i$.
- Logit model:

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}$$

$$= \frac{\exp(\boldsymbol{x}_i' \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i' \boldsymbol{\beta})}$$

$$= \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

Example

Binary regression models

- Probit model

$$\pi_i = \Phi(\beta_0 + \cdots + \beta_k x_{ik}) = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}) = \Phi(\eta_i),$$

  where $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})'$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ and $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$.
- Here $\boldsymbol{\theta} = (\beta_0, \ldots, \beta_k)'$.
- For the logit model we further obtain

$$g(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik},$$

  where g is called link function (here the logit-link).

Example

Binary regression models

- The ratio $\log \frac{\pi_i}{1-\pi_i}$ is called log-odds, which can be regarded as a linear combination of the covariates.
- For the odds ratio $\frac{\pi_i}{1-\pi_i}$ we have a multiplicative model, i.e.

$$\frac{\pi_i}{1-\pi_i} = \exp(\beta_0)\exp(\beta_1 x_{i1}) \cdot \cdots \exp(\beta_k x_{ik}).$$

Likelihood, score function and Fisher information

Likelihood

- The likelihood of the sample $X_1,...,X_n$ is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(X_i, \boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta})$$

with

$$L_i(\boldsymbol{\theta}) = f_i(X_i, \boldsymbol{\theta}).$$

- The log-likelihood is given by

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\big(f_i(X_i, \boldsymbol{\theta})\big),$$

where

$$l_i(\boldsymbol{\theta}) = \log\big(f_i(X_i, \boldsymbol{\theta})\big).$$

Example

Normal distribution

$$
\begin{aligned}
L_i(\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\
&\propto \frac{1}{\sqrt{\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\
l_i(\mu, \sigma^2) &= \log(1) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(x_i - \mu)^2 \\
&= -\frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(x_i - \mu)^2 \\
l(\mu, \sigma^2) &= \sum_{i=1}^{n} l_i(\mu, \sigma^2)
\end{aligned}
$$

Example

Linear regression

$$
\begin{aligned}
L_i(\boldsymbol{\beta}, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2\right) \\
&\propto \frac{1}{\sqrt{\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2\right) \\
l_i(\boldsymbol{\beta}, \sigma^2) &= -\frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \\
l(\boldsymbol{\beta}, \sigma^2) &= -\frac{1}{2}n\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2
\end{aligned}
$$

Example

Binary regression

We obtain

$$L_i(\boldsymbol{\beta}) = \pi_i^{y_i} \left(1 - \pi_i\right)^{1-y_i}$$
$$l_i(\boldsymbol{\beta}) = y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)$$

for the likelihood and log-likelihood.

Example (continued)

Binary regression

Because of

$$\pi_i = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}, \quad 1 - \pi_i = \frac{1}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}$$

and

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

we further obtain

$$\begin{aligned} l_i(\boldsymbol{\beta}) &= y_i \boldsymbol{x}_i'\boldsymbol{\beta} + \log\left(\frac{1}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right) \\ &= y_i \boldsymbol{x}_i'\boldsymbol{\beta} - \log\big(1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\big) \end{aligned}$$

in case of the logit model.

Score Function

$$
\begin{aligned}
S_i(\boldsymbol{\theta}) &= \left(\frac{\partial l_i(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial l_i(\boldsymbol{\theta})}{\partial \theta_p} \right)' \\
S(\boldsymbol{\theta}) &= \sum_{i=1}^{n} S_i(\boldsymbol{\theta})
\end{aligned}
$$

The ML-estimator is the solution to the following system of equations

$$S(\boldsymbol{\theta}) = \mathbf{0}.$$

Fisher Information

$$
\begin{aligned}
\mathsf{I}_i(\boldsymbol{\theta}) &= -\begin{pmatrix} \frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial\theta_1\partial\theta_1} & \cdots & \frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial\theta_1\partial\theta_p} \\ \vdots & & \vdots \\ \frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial\theta_p\partial\theta_1} & \cdots & \frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial\theta_p\partial\theta_p} \end{pmatrix} \\
\mathsf{I}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \mathsf{I}_i(\boldsymbol{\theta})
\end{aligned}
$$

The observed Fisher-Information $\mathsf{I}(\hat{\boldsymbol{\theta}}_{ML})$ can be calculated by inserting the ML-estimator $\hat{\boldsymbol{\theta}}_{ML}$.
Expected Fisher-Information $J(\boldsymbol{\theta})$ is the expected value of $\mathsf{I}(\boldsymbol{\theta})$.

Example

Normal distribution

$$
\begin{aligned}
l_i(\mu, \sigma^2) &= -\frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(x_i - \mu)^2 \\
\frac{\partial l_i(\mu, \sigma^2)}{\partial \mu} &= -\frac{1}{2\sigma^2} 2(-1)(x_i - \mu) = \frac{1}{\sigma^2}(x_i - \mu) \\
\frac{\partial l_i(\mu, \sigma^2)}{\partial \sigma^2} &= -\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(x_i - \mu)^2
\end{aligned}
$$

$$
\mathrm{S}\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \\ -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 \end{pmatrix}
$$

Example (continued)

Normal distribution

The ML-estimator is given as the solution to

$$\text{I}: \quad \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right) = 0$$

$$\text{II}: \quad -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

Example (continued)

Normal distribution

From I we have $\hat{\mu} = \bar{x}$.

Inserting $\hat{\mu}$=$\bar{x}$ in II yields

$$-\frac{n}{\sigma^2} + \frac{1}{(\sigma^2)^2} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 = 0$$

We obtain

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2.$$

Remark

- The ML-estimator for σ^2 in the previous example is biased.
- Alternatively to the ML-estimator for variance parameters, the so called Restricted ML-estimator (REML) is often used.
- It maximizes the marginal likelihood

$$\begin{aligned} \text{RL}(\sigma^2) &= \int_{-\infty}^{\infty} L(\mu, \sigma^2) \, \text{d}\mu \\ &= \int_{-\infty}^{\infty} \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \left(x_i - \mu \right)^2 \right) \right) d\mu. \end{aligned}$$

- REML estimators are used regularly for variance parameters, e.g. in linear models or linear mixed models, in the context of likelihood based inference.

Example

Linear regression

$$
\begin{aligned}
l_i(\boldsymbol{\beta}, \sigma^2) &= -\frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \\
\mathrm{l}(\boldsymbol{\beta}, \sigma^2) &= -\frac{1}{2}n\log(\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta) \\
&= -\frac{1}{2}n\log(\sigma^2) - \frac{1}{2\sigma^2}\left(Y'Y - 2Y'X\beta + \beta'(X'X)\beta\right) \\
\frac{\partial \mathrm{l}(\boldsymbol{\beta}, \sigma^2)}{\partial \beta} &= -\frac{1}{2\sigma^2}\left(-2X'Y + 2X'X\beta\right) \\
&= -\frac{1}{\sigma^2}\left(X'X\beta - X'Y\right) \\
\frac{\partial \mathrm{l}(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}(Y - X\beta)'(Y - X\beta)
\end{aligned}
$$

Example (continued)

Linear regression

Hence the score function is given by

$$\mathrm{S}\begin{pmatrix}\beta\\ \sigma^2\end{pmatrix} = \begin{pmatrix} -\frac{1}{\sigma^2}\left(X'X\beta - X'Y\right) \\ -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(Y - X\beta)'(Y - X\beta)\end{pmatrix}$$

To obtain the ML-estimator we solve

$$\mathrm{S}\begin{pmatrix}\beta\\ \sigma^2\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ 0\end{pmatrix}$$

Example (continued)

Linear regression

We immediately obtain

$$\hat{\beta}_{\text{ML}} = (X'X)^{-1}X'Y$$
$$\hat{\sigma}^2_{\text{ML}} = \frac{1}{n}(Y - X\hat{\beta})'(Y - X\hat{\beta}).$$

Again $\hat{\sigma}^2_{\text{ML}}$ is biased. The REML estimator

$$\hat{\sigma}^2_{\text{REML}} = \frac{1}{n-k-1}(Y - X\hat{\beta})'(Y - X\hat{\beta})$$

is unbiased and maximizes the marginal likelihood

$$\text{RL}(\sigma^2) = \int \text{L}\begin{pmatrix}\beta \\ \sigma^2\end{pmatrix} \text{d}\beta.$$

Example

Binary regression: Logit model

$$l_i(\beta) = y_i \boldsymbol{x}_i' \beta - \log\big(1 + \exp(\boldsymbol{x}_i' \beta)\big) = y_i \eta_i - \log\big(1 + \exp(\eta_i)\big)$$

The individual score function is given by

$$\begin{aligned} s_i(\beta) = \frac{\partial \mathsf{l}_i(\beta)}{\partial \beta} &= \frac{\partial \mathsf{l}_i(\beta)}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta} \\ &= \left(y_i - \frac{1}{1 + \exp(\eta_i)} \exp(\eta_i) \right) \boldsymbol{x}_i \\ &= (y_i - \pi_i) \boldsymbol{x}_i. \end{aligned}$$

Example (continued)

Binary regression: Logit model

This implies

$$s(\beta) = \sum_{i=1}^{n} (y_i - \pi_i) \boldsymbol{x}_i = \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{\pi}),$$

where

$$\boldsymbol{x} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \ldots & \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}$$

is a design matrix and $\boldsymbol{y} = (y_1, \ldots, y_n)'$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)'$. The ML-estimator solves the nonlinear equation system

$$s(\beta) = \boldsymbol{X}'\big(\boldsymbol{y} - \boldsymbol{\pi}\big) = \boldsymbol{0}$$

numerically.

Example (continued)

Binary regression: Logit model

$$J(\beta) = \mathrm{E}\left(-\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta'}\right) = \mathrm{Cov}(\boldsymbol{S}(\beta)) = \mathrm{E}(\boldsymbol{S}(\beta)\boldsymbol{S}'(\beta)).$$

After some tedious calculations we finally obtain

$$J(\beta) = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \pi_i (1 - \pi_i).$$

Numerical computation of the ML–estimator

- In most practical situations

$$s(\boldsymbol{\theta}) = \mathbf{0}$$

  can not be solved analytically. In this case we need numerical algorithms to determine the ML estimator.
- Here: Newton method and Fisher scoring.

Iterative solution: θ one dimensional

1. Set $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)}$
2. Fit at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)}$ a tangent to $s(\boldsymbol{\theta})$. The tangent is given by

$$y = s\left(\boldsymbol{\theta}^{(0)}\right) + \frac{\partial s\left(\boldsymbol{\theta}^{(0)}\right)}{\partial \boldsymbol{\theta}} \cdot \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}\right)$$

3. Set $\boldsymbol{\theta} = \boldsymbol{\theta}^{(1)}$ as the root of the tangent, i.e.

$$\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} - \left(\frac{\partial s\left(\boldsymbol{\theta}^{(0)}\right)}{\partial \boldsymbol{\theta}}\right)^{-1} \cdot s\left(\boldsymbol{\theta}^{(0)}\right) = \boldsymbol{\theta}^{(0)} + \mathsf{I}^{-1}\left(\boldsymbol{\theta}^{(0)}\right) \cdot s\left(\boldsymbol{\theta}^{(0)}\right)$$

4. Proceed with step 2 thereby replacing $\boldsymbol{\theta}^{(0)}$ by $\boldsymbol{\theta}^{(1)}$. Iterate until the new solution $\boldsymbol{\theta}^{(1)}$ and the previous solution do not change.

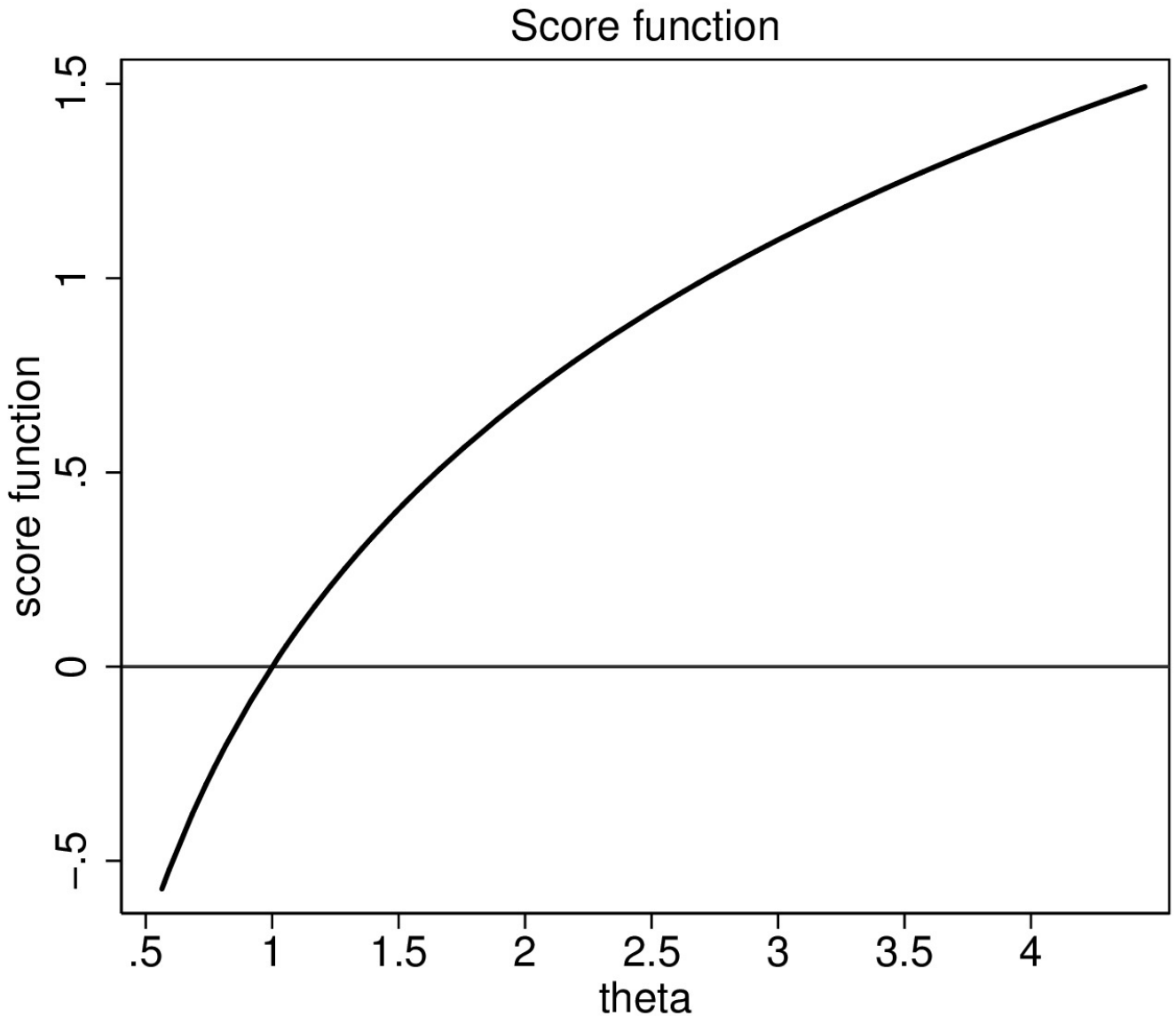


Figure: *Illustration Newton method.*

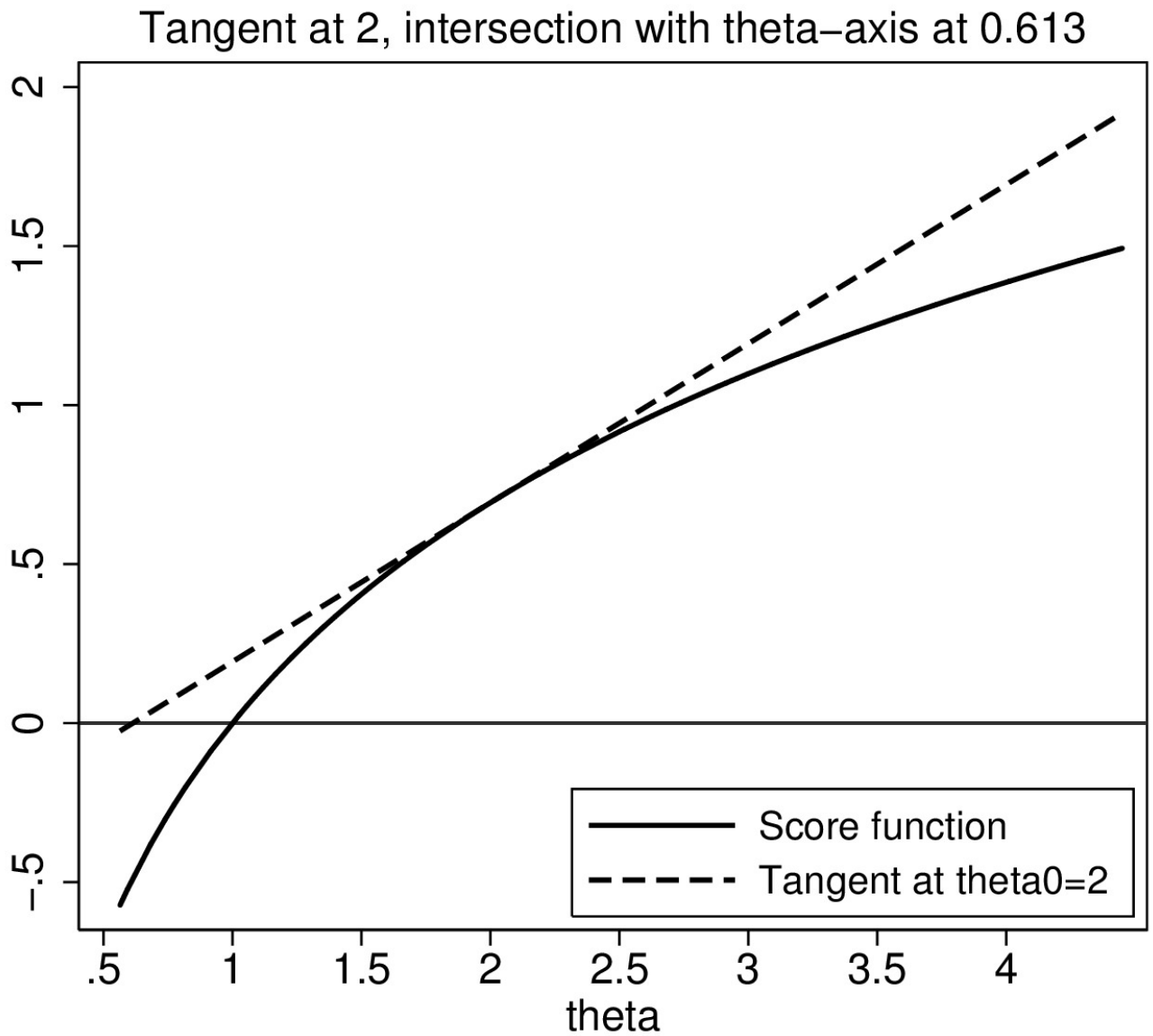


Figure: *Illustration Newton method.*

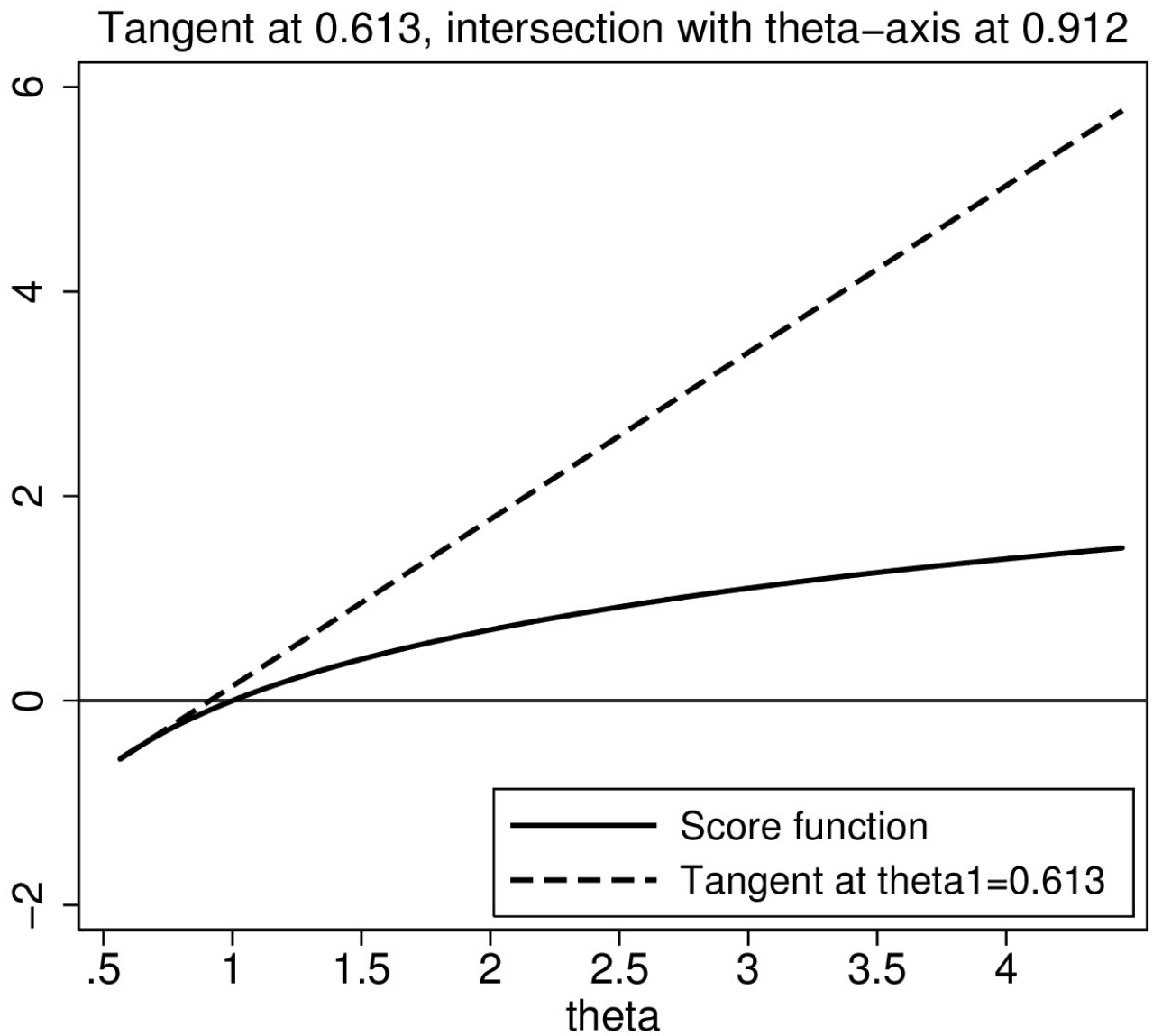


Figure: *Illustration Newton method.*

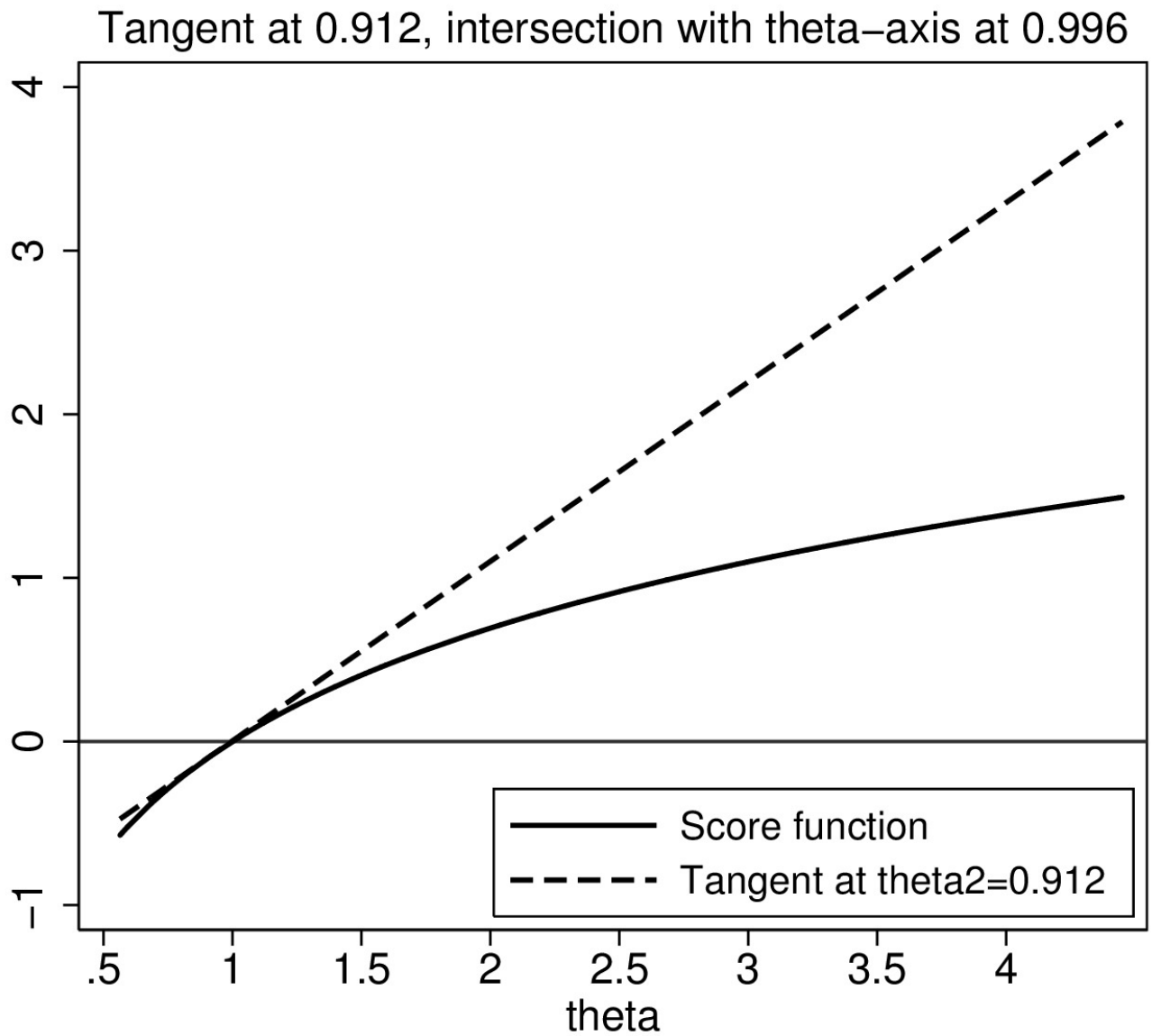


Figure: *Illustration Newton method.*

General multiparameter case $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$:

$$\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} - \left(\frac{\partial s\left(\boldsymbol{\theta}^{(0)}\right)}{\partial \boldsymbol{\theta}'} \right)^{-1} s\left(\boldsymbol{\theta}^{(0)}\right) = \boldsymbol{\theta}^{(0)} + \left(I\left(\boldsymbol{\theta}^{(0)}\right) \right)^{-1} s\left(\boldsymbol{\theta}^{(0)}\right)$$

with

$$s(\boldsymbol{\theta}) = \left(\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_p} \right)'$$

$$I(\boldsymbol{\theta}) = -\frac{\partial s(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = - \begin{pmatrix} \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_1} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_p} \\ \vdots & & \vdots \\ \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_p \partial \theta_1} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_p \partial \theta_p} \end{pmatrix}$$

Fisher scoring

Replace the Fisher information $\mathrm{I}(\boldsymbol{\theta})$ by the expected Fisher Information $J(\boldsymbol{\theta}) = E(I(\boldsymbol{\theta}))$, i.e.

$$\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} + \left(J\left(\boldsymbol{\theta}^{(0)}\right)\right)^{-1} s\left(\boldsymbol{\theta}^{(0)}\right).$$

Advantage: In many cases numerically more favorable.

Properties of the ML-estimator

a) The ML-estimator is asymptotically unbiased and consistent.

b) Distributional properties

$$\hat{\boldsymbol{\theta}}_{ML} \overset{a}{\sim} N_p\left(\boldsymbol{\theta}, J(\boldsymbol{\theta})^{-1}\right)$$

$J(\boldsymbol{\theta})$ can be substituted by $J(\hat{\boldsymbol{\theta}}_{ML})$, $\mathrm{I}(\boldsymbol{\theta})$ or $\mathrm{I}(\hat{\boldsymbol{\theta}}_{ML})$.

c) For $n \to \infty$ the ML-estimator converges to the Cramér - Rao bound:

$$Cov(\hat{\boldsymbol{\theta}}_{ML}) = J(\boldsymbol{\theta})^{-1}.$$

Hence, the ML-estimator is efficient for $n \to \infty$.

Statistical tests

Linear hypotheses

$$H_0 : \mathbf{C}\boldsymbol{\theta} = \boldsymbol{d} \qquad \text{versus} \qquad H_1 : \mathbf{C}\boldsymbol{\theta} \neq \boldsymbol{d},$$

where $\boldsymbol{C}$ is a $r \times p$ with full column rank.

Some special cases

- j-th parameter zero

$$H_0 : \theta_j = 0 \qquad \text{versus} \qquad H_1 : \theta_j \neq 0$$

$$\mathbf{C} = (0 \dots \underbrace{1}_{j} \dots 0) \qquad \boldsymbol{d} = 0$$

- Parameter θ_l equal to θ_j

$$H_0 : \theta_l = \theta_j \qquad \text{versus} \qquad H_1 : \theta_l \neq \theta_j$$

$$\mathbf{C} = (0 \dots \underbrace{1}_{l} \dots \underbrace{-1}_{j} \dots 0) \qquad \boldsymbol{d} = 0.$$

Test principles

- **Likelihood–ratio–test**
  $\tilde{\theta}$ the ML-estimator for θ under $\mathbf{C}\theta = \boldsymbol{d}$.

$$W := 2(l(\hat{\theta}_{ML}) - l(\tilde{\theta})) \overset{a}{\sim} \chi^2_r$$

- **Wald–test**

$$T := (\mathbf{C}\hat{\theta}_{ML} - \boldsymbol{d})'(\mathbf{C}\,\widehat{Cov(\hat{\theta}_{ML})}\,\mathbf{C}')^{-1}(\mathbf{C}\hat{\theta}_{ML} - \boldsymbol{d}) \overset{a}{\sim} \chi^2_r$$

- **Score–test**

$$U := s(\tilde{\theta})'\widehat{Cov(\hat{\theta}_{ML})}^{-1} s(\tilde{\theta}) \overset{a}{\sim} \chi^2_r$$

Wald-Test (special case)

- Test the hypotheses

  $H_0 : \theta_j = \theta_{j0}$ against $H_1 : \theta_j \neq \theta_{j0}$

- Possible test statistic

$$T = \frac{\hat{\theta}_j - \theta_{j0}}{se\left(\hat{\theta}_j\right)} \overset{a}{\sim} N(0,1)$$

  Reject H_0 if $|T| > z_{1-\frac{\alpha}{2}}$.

- Alternative test statistic

$$T^2 = \frac{(\hat{\theta}_j - \theta_{j0})^2}{se\left(\hat{\theta}_j\right)^2} \overset{a}{\sim} \chi^2(1)$$

  Reject H_0 if $T^2 > \chi^2_1(1-\alpha)$.

Statistical Inference

Chapter 5

Model choice and variable selecion

Model choice and variable selecion

Introduction

Introduction

- In many applications, a large (potentially enormous) number of candidate predictor variables is available, and we face the challenge and decision as to which of these variables to include in the regression model.
- The following are two naive (but often practiced) approaches to the model selection problem:
  - Strategy 1: Estimate the most complex model which includes all available covariates.
  - Strategy 2: First estimate a model with all variables. Then, remove all insignificant variables from the model.

Introduction

- We investigate the simulated data illustrated in Figure 8 a).
- The true model used for simulation is $y_i = -1 + 0.3x_i + 0.4x_i^2 - 0.8x_i^3 + \varepsilon_i$ with $\varepsilon_i \sim \text{N}(0, 0.07^2)$.
- The scatter plot suggests polynomial modeling of the relationship between y and x resulting in the regression model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_l x_i^l + \varepsilon_i.$$

- Figure 8 c) - e) show the estimated relationship for $l = 1$ (regression line), $l = 2$, and $l = 5$.

Introduction

- Figure 8 f) additionally displays the mean squared error

$$\text{MSE}(l) = \frac{1}{50} \sum_{i=1}^{50} (y_i - \hat{y}_i(l))^2$$

  of the fitted models depending on the order of the polynomial (continuous line). Clearly, $\text{MSE}(l)$ decreases monotonically with increased l.
- Figure 8 b) shows additionally simulated observations for every design point x_i, $i = 1, \ldots, 50$. We refer to this data set as the validation sample, whereas we refer to the first data set (used for estimation) as the training set.

Introduction

- Figure 8 f) shows the mean squared error of $\hat{y}_i^*$ for the data y_i^* (dashed line) in the validation set.
- Apparently, the fit to the new data is initially getting better with an increase of the polynomial order. However, from the polynomial order $l = 3$ onward, the fit is getting worse.
- The more complex the model, the better is the fit to the data that were used for estimation. However with new data resulting from the same data generating process, models that are too complex can cause a poorer fit.

Introduction

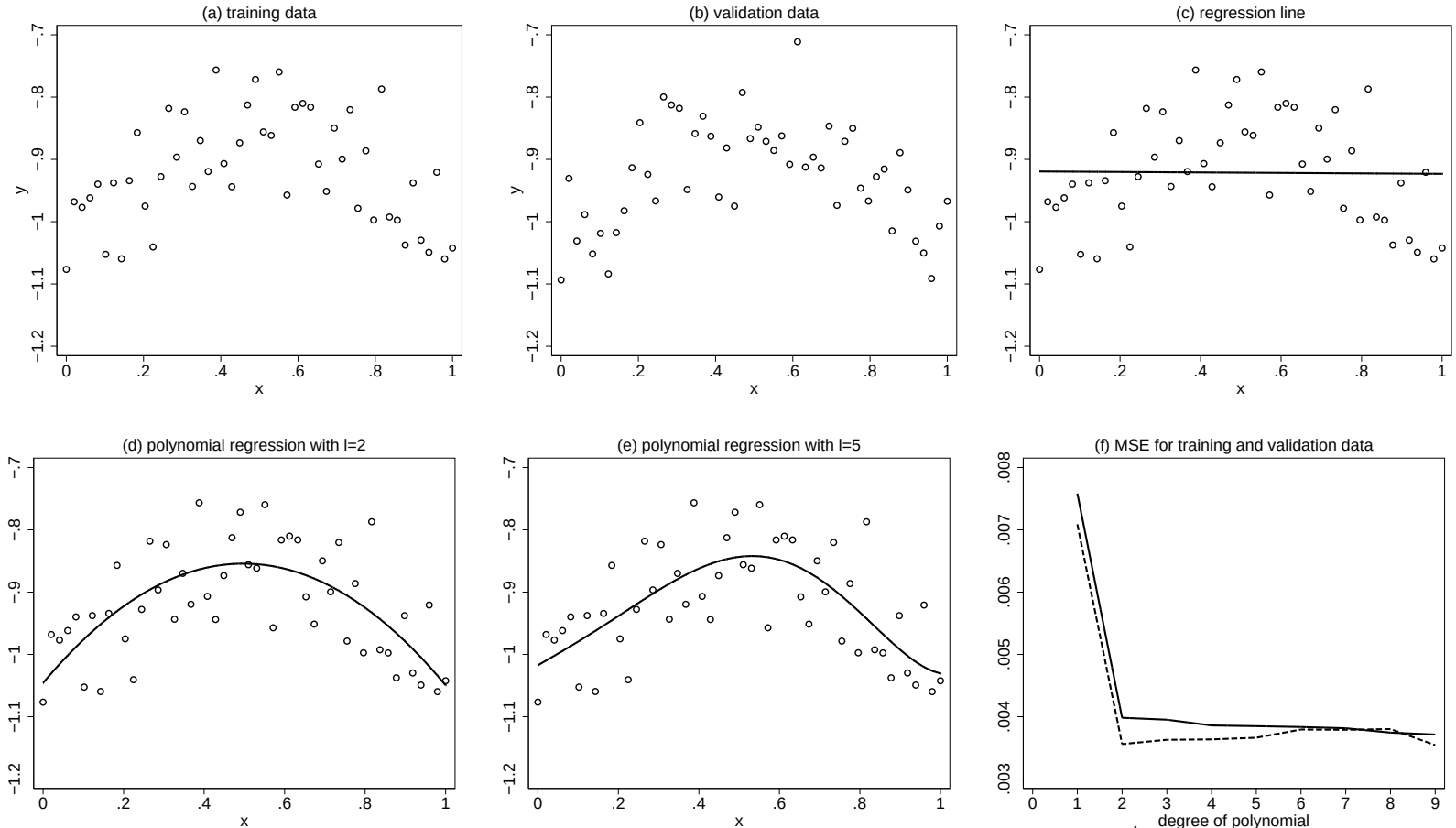


Figure: *Simulated training data y_i (panel a) and validation data y_i^* (panel b). Panels c)-e) show estimated polynomials based on the training set. Panel f) displays MSE(l) in relation to the polynomial degree (solid line). The dashed line shows MSE(l), if the estimated polynomials are used to predict the validation data y_i^*.*

Introduction

- Consider the $n = 150$ observations $(y_i, x_{i1}, x_{i2}, x_{i3})$, $i = 1, \cdots, 150$, in the scatter plot matrix in Figure 9.
- The variables x_1 and x_3 are independent and uniformly distributed on [0,1]. The variable x_2 is defined as $x_2 = x_1 + u$, where u is also uniformly distributed on [0,1].
- The response variable y is simulated according to the model

$$y \mid x_1, x_2, x_3 \sim \mathrm{N}(-1 + 0.3x_1 + 0.2x_3, 0.2^2).$$

  The conditional mean of y is thus dependent on x_1 and x_3, but not on x_2.

Introduction

- We first estimate a regression model with all available covariates x_1, x_2, and x_3, see Table 1.
- Clearly, x_1 and x_2 are nonsignificant. If we followed strategy 2, we would not only eliminate the nonrelevant covariate x_2, but also the relevant variable x_1.
- If we estimate a correctly specified model with true predictor variables x_1 and x_3, we obtain the results shown in Table 2.
- When having a correct model specification, not only is x_3 significant, but so is the previously insignificant variable x_1.

Introduction

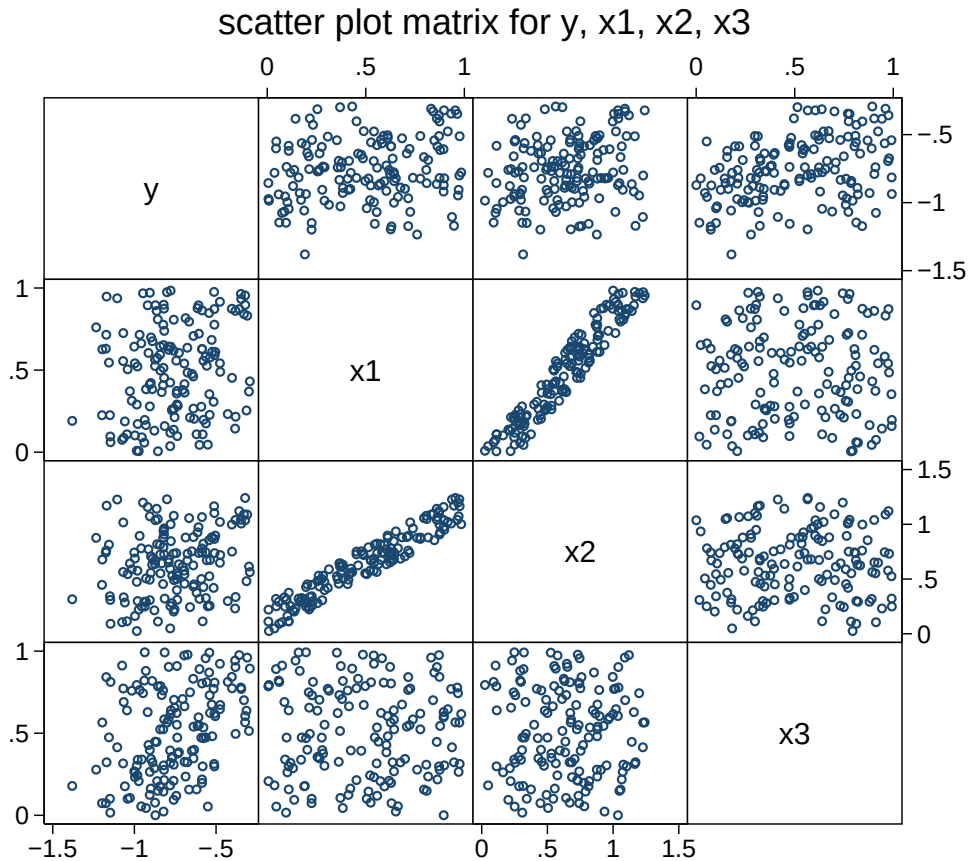


Figure: *Scatter plot matrix for the variables y, x_1, x_2 and x_3.*

Introduction

Variable	Coefficient	Standard-error	t-value	p-value	95% Confidence-interval	
intercept	-0.970	0.047	-20.46	$<$0.001	-1.064	-0.877
x_1	0.146	0.187	0.78	0.436	-0.224	0.516
x_2	0.027	0.177	0.15	0.880	-0.323	0.377
x_3	0.227	0.052	4.32	$<$0.001	0.123	0.331

Table: *Results for the model based on covariates x_1, x_2 and x_3.*

Introduction

Variable	Coefficient	Standard-error	t-value	p-value	95% Confidence-interval	
intercept	-0.967	0.039	-24.91	$<$0.001	-1.042	-0.889
x_1	0.173	0.055	3.17	0.002	0.065	0.281
x_3	0.226	0.052	4.33	$<$0.001	0.123	0.330

Table: *Results for the correctly specified model based on covariates x_1 and x_3.*

Theoretical insights

Theoretical insights

We focus on the following questions:

1. *Irrelevant Variables:* What can be said about the bias and the variance of the least squares estimator, in the case that we include irrelevant variables in the model?
2. *Missing Variables:* What can be said about the bias and the variance of the least squares estimator, if we omit relevant variables in the model?
3. *Prediction Quality:* What effect does the model specification, more specifically the selected variables in the model, have on prediction?

Bias and Variance of the Least Squares Estimator

- Consider a partition of the available explanatory variables $\boldsymbol{x} = (x_0, x_1, \ldots, x_k)'$ with $x_0 \equiv 1$ into the subsets $\boldsymbol{x}_1 = (x_0, x_1, \ldots, x_{k_1})'$ and $\boldsymbol{x}_2 = (x_{k_1+1}, \ldots, x_k)'$.
- We look at the two models

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

and

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}.$$

- The first model uses all available variables. The second model uses only the subset $\boldsymbol{x}_1$.

Bias and Variance of the Least Squares Estimator

- For the submodel we obtain the least squares estimators

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \quad \text{and} \quad \tilde{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}$$

  respectively.
- For the estimator $\tilde{\boldsymbol{\beta}}_1$ of the submodel, we obtain

$$\mathrm{E}(\tilde{\boldsymbol{\beta}}_1) = \boldsymbol{\beta}_1 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{\beta}_2$$

  and

$$\mathrm{Cov}(\tilde{\boldsymbol{\beta}}_1) = \sigma^2(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}.$$

- For the estimator $\hat{\boldsymbol{\beta}}$ in the full model we obtain

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \qquad \mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2\,(\mathbf{X}'\mathbf{X})^{-1}.$$

Bias and Variance of the Least Squares Estimator

We now investigate the following two situations:

- *Missing variables:* Even though the complete model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is correct, we mistakenly estimate the reduced model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}$. In this case we neglect the relevant variables $\boldsymbol{x}_2$.
- *Irrelevant Variables:* Even though the reduced model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}$ is correct, we mistakenly estimate the full model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. In this case, we included irrelevant variables in the model. The variables in $\boldsymbol{x}_2$ are redundant.

Bias and Variance of the Least Squares Estimator

In the first case of missing variables the following applies:

- $\boldsymbol{\beta}_1$ is biased. An exception is the case when $\boldsymbol{X}_1'\boldsymbol{X}_2 = \mathbf{0}$, i.e. every variable in $\boldsymbol{X}_1$ is uncorrelated to every variable in $\boldsymbol{X}_2$.
- It can be shown that the difference $\text{Cov}(\hat{\boldsymbol{\beta}}_1) - \text{Cov}(\tilde{\boldsymbol{\beta}}_1)$ of covariance matrices is positive semi-definite. This implies that the components of the estimator $\tilde{\boldsymbol{\beta}}_1$ based on the submodel $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}$ show a smaller variance than the corresponding components of the estimator $\hat{\boldsymbol{\beta}}_1$ based on the correct model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$. Thus we have $\text{Var}(\hat{\beta}_j) \geq \text{Var}(\tilde{\beta}_j)$.
- It can be shown that situations exist, in which the components in $\tilde{\boldsymbol{\beta}}_1$ based on the misspecified submodel actually show a smaller MSE than the components in $\hat{\boldsymbol{\beta}}_1$, which are based on the full model, i.e. $\text{MSE}(\hat{\beta}_j) \geq \text{MSE}(\tilde{\beta}_j)$.

Bias and Variance of the Least Squares Estimator

In the second case of irrelevant variables, we have:

- Even though irrelevant variables were considered, $\hat{\boldsymbol{\beta}}$ is unbiased. Of course, the estimator $\tilde{\boldsymbol{\beta}}_1$ based on the true model is also unbiased.
- It can be shown that the estimators for the components in $\boldsymbol{\beta}_1$ based on $\hat{\boldsymbol{\beta}}$ have larger variance than based on $\tilde{\boldsymbol{\beta}}_1$. Thus we have $\text{Var}(\hat{\beta}_j) \geq \text{Var}(\tilde{\beta}_j)$. If the estimated model contains irrelevant variables, then the precision of the estimators decreases.

We can reach the following conclusion: Preferably, the specified model should not contain irrelevant covariates. Moreover, we should aim for a sparse model so that bias and variance, and thus MSE, are small.

Consequences of the Model Specification on Prediction

- Next we take a look at prediction quality in linear models. Thereby, we do not necessarily assume that the model is correctly specified.
- We assume independent observations y_i, $i = 1, \ldots, n$, with expectation $\mathrm{E}(y_i) = \mu_i$ and variance $\mathrm{Var}(y_i) = \sigma^2$.
- The variables $x_0 = 1, x_1, \ldots, x_k$ are available as potential regressors.
- We assume that a subset $M \subset \{0, 1, 2, \ldots, k\}$ of the available variables will be used for estimation.
- The specified model is defined by the subset of included covariates with corresponding design matrix $\boldsymbol{X}_M$.

Consequences of the Model Specification on Prediction

- For the least squares estimator we obtain

$$\hat{\boldsymbol{\beta}}_M = (\boldsymbol{X}_M'\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M'\boldsymbol{y}.$$

- An estimator $\hat{\boldsymbol{y}}_M$ for the vector $\boldsymbol{\mu}$ of means $\mu_i = \mathrm{E}(y_i)$ is given by

$$\hat{\boldsymbol{y}}_M = \boldsymbol{X}_M\hat{\boldsymbol{\beta}}_M.$$

- We can view the estimator $\hat{y}_{iM}$ also as a prediction for future observations $y_{n+i} = \mu_i + \varepsilon_{n+i}$, $i = 1, \ldots, n$, with given covariates $x_{i1}, \ldots, x_{ik}$.
- In the following, we derive a formula for the sum of the expected squared prediction errors $\sum \mathrm{E}(y_{n+i} - \hat{y}_{iM})^2$.

Consequences of the Model Specification on Prediction

To do so, we need the following, easily verifiable, properties of $\hat{\boldsymbol{y}}_M$:

- Expectation:
$$\mathrm{E}(\hat{\boldsymbol{y}}_M) = \boldsymbol{X}_M(\boldsymbol{X}_M'\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M'\mathrm{E}(\boldsymbol{y}).$$
- Covariance matrix:
$$\mathrm{Cov}(\hat{\boldsymbol{y}}_M) = \sigma^2\boldsymbol{X}_M(\boldsymbol{X}_M'\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M'.$$
- Sum of the variances:
$$\sum_{i=1}^{n}\mathrm{Var}(\hat{y}_{iM}) = \sigma^2\mathrm{tr}(\boldsymbol{X}_M(\boldsymbol{X}_M'\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M') = |M|\,\sigma^2,$$

  where $|M|$ represents the cardinal number of M, i.e. the number of the covariates included in the model. The sum of the variances increases as more covariates are included in the model.

Consequences of the Model Specification on Prediction

- Sum of the mean squared errors (SMSE):

$$\begin{aligned}
\text{SMSE} &= \sum_{i=1}^{n} \mathrm{E}(\hat{y}_{iM} - \mu_i)^2 \\
&= \sum_{i=1}^{n} \mathrm{E}\left((\hat{y}_{iM} - \mu_{iM}) + (\mu_{iM} - \mu_i)\right)^2 \\
&= \sum_{i=1}^{n} \mathrm{Var}(\hat{y}_{iM}) + 2\sum_{i=1}^{n} \mathrm{E}\left((\hat{y}_{iM} - \mu_{iM})(\mu_{iM} - \mu_i)\right) + \sum_{i=1}^{n} (\mu_{iM} - \mu_i)^2 \\
&= |M|\,\sigma^2 + \sum_{i=1}^{n} (\mu_{iM} - \mu_i)^2.
\end{aligned}$$

Here we used $\mu_{iM} = \mathrm{E}(\hat{y}_{iM})$ as an abbreviation for the expectation of the estimator $\hat{y}_{iM}$.

Consequences of the Model Specification on Prediction

These properties provide us with the expected squared prediction error:

$$
\begin{aligned}
\text{SPSE} &= \sum_{i=1}^{n} \mathrm{E}(y_{n+i} - \hat{y}_{iM})^2 \\
&= \sum_{i=1}^{n} \mathrm{E}((y_{n+i} - \mu_i) - (\hat{y}_{iM} - \mu_i))^2 \\
&= \sum_{i=1}^{n} (\mathrm{E}(y_{n+i} - \mu_i)^2 - 2\mathrm{E}((y_{n+i} - \mu_i)(\hat{y}_{iM} - \mu_i)) + \mathrm{E}(\hat{y}_{iM} - \mu_i)^2) \\
&= \sum_{i=1}^{n} \mathrm{E}(y_{n+i} - \mu_i)^2 + \sum_{i=1}^{n} \mathrm{E}(\hat{y}_{iM} - \mu_i)^2 \\
&= n\sigma^2 + \text{SMSE} \\
&= n\sigma^2 + |M|\,\sigma^2 + \sum_{i=1}^{n} (\mu_{iM} - \mu_i)^2
\end{aligned}
$$

Note that in line 3 of the above derivation, the expectation for the cross product term can be written as the product of expectations due to the independence of $\hat{y}_{iM}$ and y_{n+i}. This way the entire term becomes zero.

Consequences of the Model Specification on Prediction

Thus, the expected squared prediction error can be decomposed into three additive terms:

- *Irreducible Prediction Error:* The first term $n\sigma^2$ depends on the error variance. Hence, it cannot be reduced, even by sophisticated inference techniques. This term is therefore referred to as the irreducible prediction error.
- *Variance:* The second term consists of the sum of variances $\text{Var}(\hat{y}_{iM})$ of the estimators $\hat{y}_{iM}$. This term can be manipulated through model choice. It becomes smaller as fewer variables are included in the model.
- *Squared Bias:* The last term $\sum(\mu_{iM} - \mu_i)^2$ can be seen as a bias term. It consists of the squared bias of the estimator $\hat{y}_{iM}$ for the expectation μ_i. This term can also be manipulated through model choice and becomes smaller as more variables are included in the model.

Consequences of the Model Specification on Prediction

- The decomposition of the expected prediction error into an irreducible error, a variance term, and a squared bias term is not limited to linear models, but rather a *fundamental property of prediction* in all statistical models.
- The formula for SPSE shows a classical bias-variance trade-off.
- The more complex the model, the smaller the squared bias and the greater the variance. On the contrary, simpler models show a greater squared bias and in return for that a smaller variance.
- This bias-variance trade-off is not only characteristic for linear models, but for all statistical models.

Model choice

Model choice: Minimizing SPSE

Estimate SPSE using new and independent data

- If in fact additional observations y_{n+i} are available, we are able to estimate $\text{SPSE} = \sum \text{E}(y_{n+i} - \hat{y}_{iM})^2$ simply by $\widehat{\text{SPSE}} = \sum_{i=1}^{n}(y_{n+i} - \hat{y}_{iM})^2$.
- In practice, it is usually not possible to use this approach, as it rarely happens that additional observations are collected.
- An alternative procedure is the following:
  - Randomly split the data into two parts, i.e. a test and a validation sample.
  - Use the test data set to estimate the specified model.
  - Use the validation set to assess the goodness-of-fit, i.e. for the estimation of SPSE.

Model choice: Minimizing SPSE

Estimate SPSE using existing data

- A naive estimator for SPSE would be the use of the squared sum of residuals $\sum(y_i - \hat{y}_{iM})^2$.
- Note that this sum *underestimates* on average the expected prediction error, as it can be shown that

$$\mathrm{E}\left(\sum_{i=1}^{n}(y_i - \hat{y}_{iM})^2\right) = \mathrm{SPSE} - 2|M|\sigma^2.$$

- Thus a better estimate for SPSE is given by

$$\widehat{\mathrm{SPSE}} = \sum_{i=1}^{n}(y_i - \hat{y}_{iM})^2 + 2|M|\hat{\sigma}^2.$$

Model choice: Minimizing SPSE

- Accordingly, we choose a model that minimizes $\widehat{\text{SPSE}}$. In doing so we have to keep in mind that we always use the same estimator for $\hat{\sigma}^2$. Preferably, this estimator should be based on the full model with all available variables, in order to keep the bias in $\hat{\sigma}^2$ small.
- The criterion $\widehat{\text{SPSE}}$ has the typical structure of many model choice criteria. It consists of two terms: The first term, the sum of squared residuals, measures the goodness-of-fit and becomes smaller the more complex the model becomes. The second term $2|M|\hat{\sigma}^2$ measures model complexity and becomes smaller as models become simpler.

Model choice: Corrected coefficient of determination

- When comparing different models, the use of the coefficient of determination is limited, since the coefficient of determination will always increase (never decrease) with the addition of a new covariate into the model.
- The *corrected coefficient of determination* adjusts for this problem, by including a correction term for the number of parameters.
- The corrected coefficient of determination is defined by

$$\bar{R}^2 = 1 - \frac{n-1}{n-p}(1 - R^2).$$

We advice against its usage, since the "penalty" for newly included covariates appears to be too small. It can be shown that $\bar{R}^2$ already increases when a variable with a t-value greater than 1 is included in the model, implying we would include variables with a p-value of about 0.3.

Model choice: Mallow's C_p

- Mallow's C_p ("Complexity parameter") is defined by

$$C_p = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{iM})^2}{\hat{\sigma}^2} - n + 2|M|.$$

- C_p can be understood as an estimate of SMSE/σ^2.
- Thus minimizing C_p produces the same optimal model as minimizing $\widehat{\text{SPSE}}$.

Model choice: Akaike information criterion

- AIC is defined by

$$\text{AIC} = -2 \cdot l(\hat{\boldsymbol{\beta}}_M, \hat{\sigma}^2) + 2\,(|M| + 1),$$

  where $l(\hat{\boldsymbol{\beta}}_M, \hat{\sigma}^2)$ is the maximum value of the log-likelihood.
- Smaller values of the AIC correspond to a better model fit.
- In a linear model with Gaussian errors, we obtain

$$\text{AIC} = n \cdot \log(\hat{\sigma}^2) + 2\,(|M| + 1).$$

Model choice: Akaike information criterion

- Figure 10 plots AIC for the simulated data from the Example on page 237 as a function of the polynomial degree.
- AIC obtains a minimum for $l = 2$ resulting in a reasonable model, even though we do not obtain the polynomial order of the true model with $l = 3$.

Model choice: Akaike information criterion

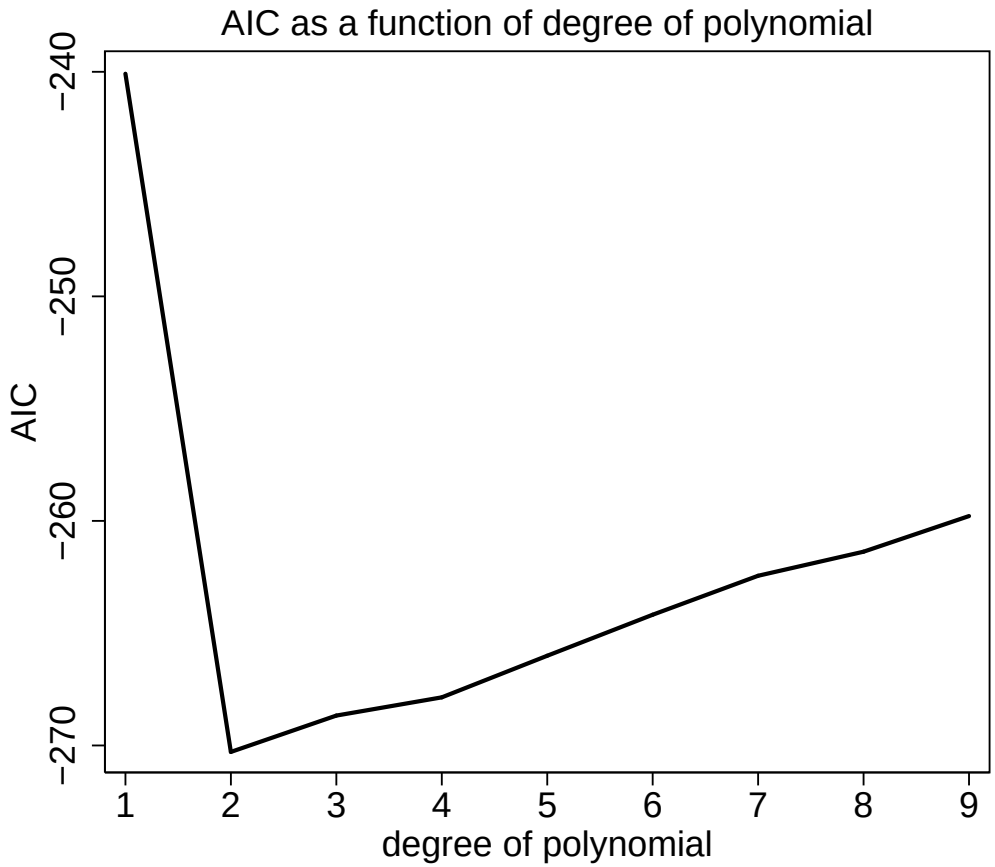


Figure: *AIC as a function of the polynomial degree for the simulated data of the example on page 237.*

Model choice: Cross Validation

Cross validation is based on the following general principle:

- Partition the data set into r subsets $1, \ldots, r$, of similar size.
- Start with the first data set as validation set and use the combined remaining $r - 1$ data sets for parameter estimation. Based on the estimates, obtain predictions for the validation set and determine the sum of the squared prediction errors.
- Cycle through the partitions using the second, third, up to the rth data set as validation sample, and all other data sets for estimation. Determine the sum of squared prediction errors.
- Use the model with the smallest sum of squared prediction errors, where the final parameter estimates reflect all data. The partition into test and validation samples serves only to estimate the expected squared prediction error.

Model choice: “leave-one-out” Cross Validation

- “leave-one-out” cross validation uses all observations with the exception of *one* for the estimation of model parameters.
- We use this “leave-one-out” estimator to predict the deleted observation and to determine the squared prediction error.
- If we denote the “leave-one-out” estimator with $\hat{y}_{iM}^{-i}$, we obtain the cross validation score

$$\text{CV} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{iM}^{-i})^2.$$

- It can be shown

$$\text{CV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_{iM}}{1 - h_{iiM}}\right)^2,$$

  where h_{iiM} are the diagonal elements of the hat matrix $\mathbf{H_M} = \boldsymbol{X}_\mathbf{M}(\boldsymbol{X}_\mathbf{M}'\boldsymbol{X}_\mathbf{M})^{-1}\boldsymbol{X}_\mathbf{M}'$.

Practical Use of Model Choice Criteria

Practical Use of Model Choice Criteria

We can use the various model choice criteria to select the most promising models from candidate models. We recommend the following approach:

- On the basis of scientific knowledge, perhaps gained from previous research, we obtain a preselection of potential models. The models can differ in the number of variables but also in model type (e.g. linear vs. nonlinear). The total number of potential models should be as small as possible.
- All potential models can now be assessed with the aid of one of the various model choice criteria. The summary of the results should not be restricted to the “best” model. As a rule, there are a number of competitive models having approximately equal model fit, differing only in small aspects from each other. These differences cause some uncertainty regarding the conclusions.

Practical Use of Model Choice Criteria

If the number of potential models is large, we can use the following partially heuristic methods:

- *Complete Model Selection (All-Subset-Selection)*: In case that the number of covariates is smaller than about 40, we can determine the best model (in the sense of a model choice criterion) with the "leaps and bounds" algorithm. An implementation can be found e.g. in the R package `leaps`.
- *Forward Selection:* Based on a starting model, forward-selection includes one additional variable in every iteration of the algorithm. The variable which offers the greatest reduction of a preselected model choice criteria (C_p, AIC, CV) is chosen. The algorithm terminates if no further reduction is possible.

Practical Use of Model Choice Criteria

- *Backward Elimination:* Backward elimination starts with the full model containing all potential covariates. Subsequently, in every iteration, the covariate which provides the greatest reduction of the model choice criteria (C_p, AIC, CV) is eliminated from the model. The algorithm terminates if no further reduction is possible.
- *Stepwise Selection:* Stepwise selection is a combination of forward selection and backward elimination. In every iteration of the algorithm, the inclusion and the deletion of a variable are both possible.

Practical Use of Model Choice Criteria

- The listed procedures should not be confounded with an algorithm proposed by Efroymson in the 1960s, even though the approach is similar.
- In contrast to what has been proposed above, the Efroymson algorithm includes or excludes those variables in/from the model, which have the highest or lowest t-value.
- The procedure terminates when no variable that potentially needs to be included has a p-value of less than a previously fixed maximal p-value (e.g. 0.05) and when no variable that needs to be excluded has a p-value greater than a minimal p-value (e.g. 0.1).

Practical Use of Model Choice Criteria

This automatic procedure, which is implemented in all major statistical software packages, is often viewed as obscure among statisticians due to the following two reasons:

- Forward, backward, and stepwise selection usually provide different results. This also happens when using a global model choice criterion such as AIC. We can, however, compare the different selected models with the help of the global model choice criterion. When using the Efroymson approach, discrimination between the different models is impossible.
- The repetitive use of the t-test statistic, to assess whether or not a regression coefficient is different from zero, suggests exact tests. However, the t-test statistic does not follow a t-distribution under the null hypothesis, since during the selection process we do not test an arbitrary variable, but rather the variable with the *maximal* t-value.

Case study: prices of used cars

- We illustrate the approaches for model choice using data from the sales price of pre-owned VW Golf automobiles.
- Our goal is to model the relationship between the sales price in 1000 Euro (variable *price*) and the five explanatory variables “age of the car in months” (*age*), “kilometer reading in 1000 km” (*kilometer*), “number of months until the next appointment with the Technical Inspection Agency” (*TIA*), “ABS brake yes/no” (*extras1*), and “sunroof yes/no” (*extras2*).

Case study: prices of used cars

The plots of the Figure on page 281 suggest the following effects:

- We can assume a linear or monotonically decreasing nonlinear effect for the variables *age* and *kilometer*, which could be appropriately modeled using (orthogonal) polynomials of degree three or less.
- The variable *TIA* appears to either have no effect or a very weak linear effect on the average sales price.
- Cars with ABS (*extras1*) seem to be slightly more expensive than cars without ABS; the effect, however, remains arguable.
- We can attest to no difference in the average sales price for models with or without sunroof (*extras2*).
- All in all, there seems to be a relationship with age and the kilometer reading. The effects of the remaining variables appear doubtful.

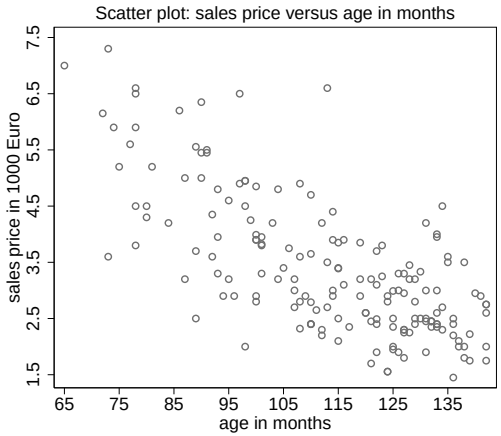
Case study: prices of used cars

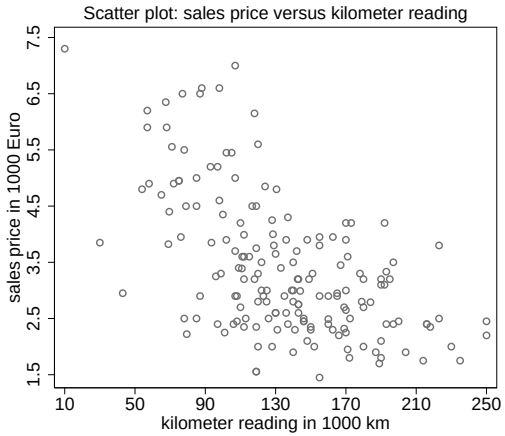
- We first examine eight regression models (see Table 3), which do not differ in the modeling of the variables *age* and *kilometer*. For the remaining three regressors, all possible model combinations will be tested.
- Using the AIC criterion, we obtain the first model in Table 3 as the preliminary best model. Figure 11 displays the AIC values for the eight models under consideration. In addition, the AIC for a ninth model based on automatic variable selection is provided; see below.

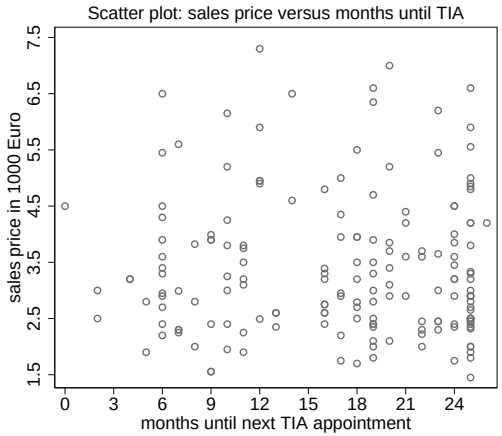
Case study: prices of used cars

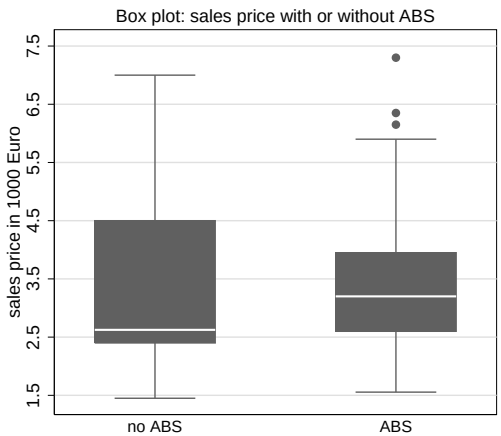
- Since only five explanatory variables are available, we can even determine the AIC best model with the help of the “Leaps and Bounds” algorithm. This model attains an AIC value of 389.35. It differs from the current “best model", in that it only makes use of polynomials of second degree (not third) for the variables *age* and *kilometer*.
- Figure 11 shows that the obtained AIC for this ninth model is considerably smaller than the best AIC value of all the models that we examined so far.

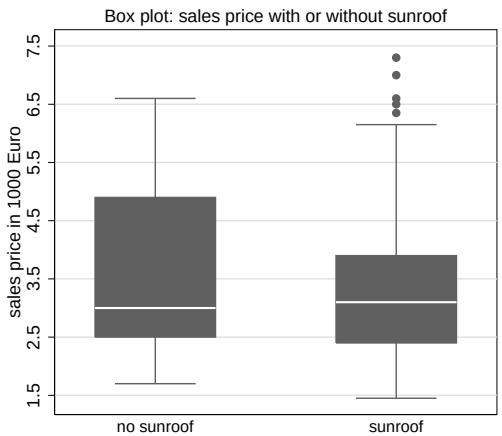
Case study: prices of used cars











Case study: prices of used cars

Model	*kilometer* degree 3	*age* degree 3	*extras1*	*extras2* linear	*TIA*	**AIC**
1	x	x				393.234 (1)
2	x	x	x			394.566 (2)
3	x	x		x		395.119 (4)
4	x	x			x	394.973 (3)
5	x	x	x	x		396.481 (6)
6	x	x	x		x	396.143 (5)
7	x	x		x	x	396.881 (7)
8	x	x	x	x	x	398.085 (8)

Table: *Prices of used cars: potential models. The values in brackets indicate the rank of the models according to AIC.*

Case study: prices of used cars

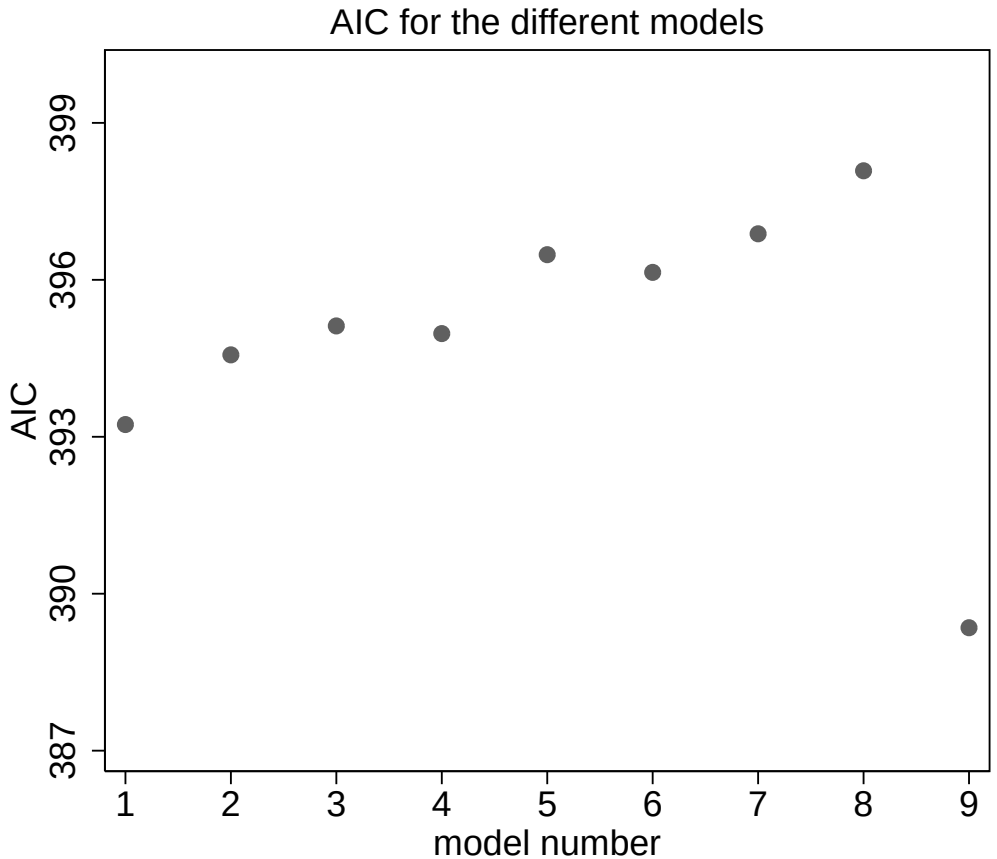


Figure: *Prices of used cars: AIC values for the potential models.*

Case study: prices of used cars

Variable	Coefficient	Standard-error	t-value	p-value	95% Confidence-interval	
intercept	3.397	0.056	60.220	$<$0.001	3.285	3.508
ageop1	-0.705	0.061	-11.470	$<$0.001	-0.826	-0.584
ageop2	0.187	0.057	3.270	0.001	0.074	0.300
kilometerop1	-0.439	0.061	-7.170	$<$0.001	-0.560	-0.318
kilometerop2	0.141	0.057	2.460	0.015	0.028	0.254

Table: *Prices of used cars: estimation results for the best model according to AIC.*

Case study: prices of used cars

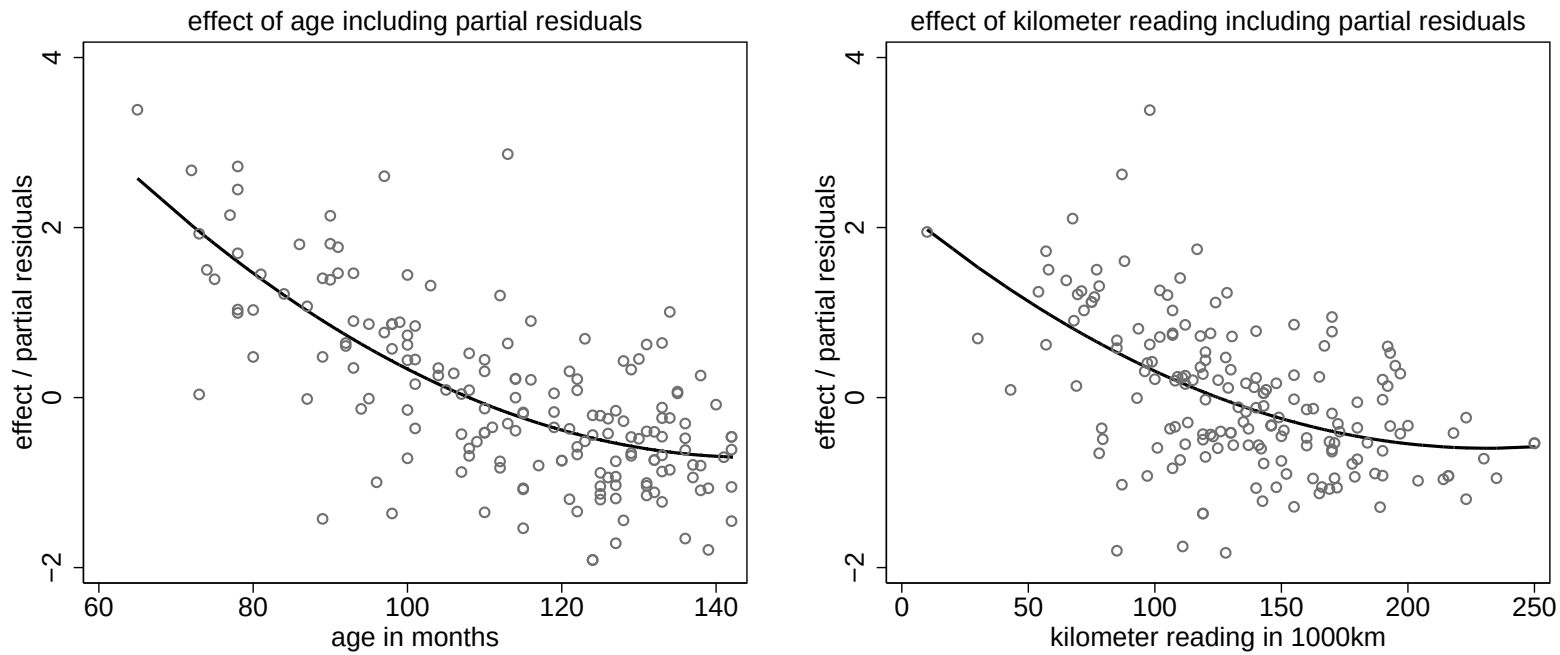


Figure: *Prices of used cars: model 9 based on all-subset selection, effects of age and kilometer reading including partial residuals.*

Statistical Inference

Chapter 6

Bayesian Inference

Bayesian Inference

Basic Concepts

Basic Concepts

- The fundamental difference to likelihood-based inference is that the unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ are not considered as fixed, deterministic quantities but as random variables with a *prior distribution*.
- *Prior distribution:* Any (subjective) information about the unknown parameter $\boldsymbol{\theta}$ is expressed by specifying a probability distribution $p(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$. The prior describes the *degree of uncertainty* about the unknown parameters prior to the statistical analysis.
- *Observation model:* The observation model specifies the conditional distribution of observable quantities, that is the random sample variables $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$, given the parameters. The p.d.f. or probability function of this conditional distribution is proportional to the likelihood $L(\boldsymbol{\theta})$ and will be denoted by $p(\boldsymbol{y} \mid \boldsymbol{\theta})$.

Basic Concepts

- Based on the prior and the observation model, Bayes' theorem determines the distribution of $\boldsymbol{\theta}$ after the data are known through the statistical experiment, that is the conditional distribution of $\boldsymbol{\theta}$ given the observations $\boldsymbol{y} = (y_1, \ldots, y_n)'$.
- We obtain

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} = c \cdot p(\boldsymbol{y} \mid \boldsymbol{\theta})\; p(\boldsymbol{\theta}),$$

  with the normalizing constant $c = [\int p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}]^{-1}$. This conditional distribution is called *posterior (distribution)*.

Example

Poisson Distribution

- Consider an i.i.d. sample $Y_1, \ldots, Y_n$ from a Poisson distribution, i.e. $Y_i \sim Po(\lambda)$.
- The joint probability for the observed sample $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is

$$p(\boldsymbol{y} \mid \lambda) = \frac{1}{y_1! \cdot \cdots \cdot y_n!} \lambda^{\sum_{i=1}^{n} y_i} \exp(-n\lambda).$$

- We specify a gamma distribution with parameters a and b for λ, i.e. $\lambda \sim Ga(a, b)$. It follows that λ has p.d.f.

$$p(\lambda) = k\, \lambda^{a-1} \exp(-b\lambda)$$

with $k = \frac{b^a}{\Gamma(a)}$.

Poisson Distribution

- The posterior is obtained as

$$\begin{aligned} p(\lambda \mid \boldsymbol{y}) &= \frac{p(\boldsymbol{y} \mid \lambda)\, p(\lambda)}{\int p(\boldsymbol{y} \mid \lambda)\, p(\lambda)\, d\lambda} \\ &= c\, \frac{1}{y_1! \cdots y_n!} \lambda^{\sum_{i=1}^n y_i} \exp(-n\lambda)\, k \lambda^{a-1} \exp(-b\lambda). \end{aligned}$$

- To determine the type of this distribution, we can ignore all factors that do not depend on λ. This gives

$$\begin{aligned} p(\lambda \mid \boldsymbol{y}) &\propto \lambda^{\sum_{i=1}^n y_i} \exp(-n\lambda)\, \lambda^{a-1} \exp(-b\lambda) \\ &= \lambda^{a+ \sum_{i=1}^n y_i - 1} \exp(-(b+n)\lambda). \end{aligned}$$

Poisson Distribution

- This has the form of a gamma distribution with parameters $a' = a + \sum_{i=1}^{n} y_i$ and $b' = b + n$, i.e.

$$\lambda \,|\, \boldsymbol{y} \sim Ga\left(a + \sum_{i=1}^{n} y_i, b + n\right),$$

  and the posterior has the same type of distribution as the prior.
- We call the prior as conjugate to the Poisson model because the posterior is of the same type as the prior.

Bayesian Logic Model - Diffuse Prior

- We consider a logit model with a single covariate x:

$$Y_i = B(1, \pi_i), \quad \pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}, \quad \eta_i = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n.$$

- Assuming, as usual, (conditionally) independent response variables, the observation model is given by

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}) \propto L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ is the vector of regression coefficients.

Bayesian Logic Model - Diffuse Prior

- Since estimated regression coefficients are often approximately normally distributed, it is reasonable to assume a two-dimensional normal prior, i.e.

$$p(\beta) \sim N_2(\boldsymbol{m}, \boldsymbol{M})$$

  with prior mean $\boldsymbol{m}$ and prior covariance matrix $\boldsymbol{M}$.
- If results from a previous statistical analysis are available, we could choose the previous point estimate as $\boldsymbol{m}$ and its estimated covariance matrix as $\boldsymbol{M}$.
- If the previous analysis has been carried out some time ago, we may also multiply $\boldsymbol{M}$ with a factor $a > 1$ to express increased uncertainty.

Bayesian Logic Model - Diffuse Prior

- Increasing the variances in $\boldsymbol{M}$, the normal prior becomes very flat and approximates a uniform distribution.
- In the limiting case the prior becomes proportional to a constant, i.e.

$$p(\boldsymbol{\beta}) \propto \text{const.}$$

  We also write $p(\boldsymbol{\beta}) \propto 1$.
- The integral of this flat prior over $\mathbb{R}^2$ is not finite, so that $p(\boldsymbol{\beta})$ is not a density in the usual sense. Such a prior is called improper or diffuse.

Bayesian Logic Model - Diffuse Prior

- Such diffuse priors are admissible as long as the posterior, resulting from Bayes' theorem, is a proper distribution. i.e. its integral over $\mathbb{R}^2$ is finite. In a Bayesian logit model this is the case if a finite MLE exists.
- With a flat, diffuse prior the posterior density is

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{\beta}) p(\boldsymbol{y} \mid \boldsymbol{\beta})}{\int p(\boldsymbol{\beta}) p(\boldsymbol{y} \mid \boldsymbol{\beta}) d\boldsymbol{\beta}} \propto p(\boldsymbol{y} \mid \boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i}.$$

- Although the posterior is proper, it has no known distributional type.

Bayesian Point Estimates

- The posterior mean is given by

$$\hat{\boldsymbol{\theta}} = EW(\boldsymbol{\theta} \mid \boldsymbol{y}) = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta} = c \cdot \int \boldsymbol{\theta}\, p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

- The posterior mode is the value $\hat{\boldsymbol{\theta}}$ that (globally) maximizes the posterior density, i.e.

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\ p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \underset{\boldsymbol{\theta}}{\text{argmax}}\, p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}).$$

  The second expression shows that no integration is necessary to compute the posterior mode, because the normalizing constant is not needed.

- The posterior median, that is the median of the posterior distribution, is sometimes preferred to the posterior mean because it is more robust against outliers.

Poisson Distribution

- The posterior mean is

$$EW(\lambda \mid \boldsymbol{y}) = \frac{a + \sum_{i=1}^{n} y_i}{b + n}.$$

- The smaller a (in relation to $\sum y_i$) and b (in relation to n), the closer the posterior mean is to the usual MLE $\hat{\lambda} = \bar{y}$.
- The larger the prior information, i.e. the larger a and b are, the more the posterior mean and the MLE differ from each other.

Bayesian Interval Estimates

- For the posterior mean, a natural measure is the posterior variance.
- For the posterior median, the interquartile distance seems to be appropriate to measure its variability.
- In case of the posterior mode, the curvature of the posterior at the mode, i.e. the observed Fisher information, is a natural choice.
- Another way of assessing uncertainty are Bayesian confidence intervals or *credible intervals* or, more generally, *credible regions*: A region $C \subset \boldsymbol{\Theta}$ of the parameter space is said to be a $(1-\alpha)$-credible region for $\boldsymbol{\theta}$ if

$$\mathrm{P}(\boldsymbol{\theta} \in C \mid \boldsymbol{y}) = 1 - \alpha.$$

  If $C \subseteq \mathbb{R}$ is an interval it is called credible interval.
- A credible region contains (at least) a probability mass $1-\alpha$ of the posterior.

Bayesian Inference

Markov Chain Monte Carlo Methods

Basic Idea

- MCMC methods allow to draw samples from posterior distributions (and, in principle, from any distribution) that are usually not available analytically and to estimate characteristics of the posterior such as the mean, the variance or quantiles, or the posterior density itself.
- The most important advantage compared to more traditional methods of drawing a sample from a distribution, for example importance or rejection sampling, is that samples can be drawn from high-dimensional densities, even for dimensions in the thousands.
- Another advantage is that the normalizing constant, often a high-dimensional integral that cannot be computed with traditional numerical methods, does not have to be known.

MCMC Methods - Basis Idea

- Let $\boldsymbol{\theta}$ be the unknown vector of parameters in a Bayesian model and $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ the posterior density (we assume here that $\boldsymbol{\theta}$ is continuous).
- Instead of directly drawing an i.i.d. sample from $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, a Markov chain is generated such that the iterations of the transition kernel converge to the posterior of interest.
- In this way random numbers are generated that can be considered as a (correlated) sample from the posterior after some time of convergence, the *burn-in phase*.

Metropolis-Hastings Algorithm

To draw random numbers from the density $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, the Metropolis-Hastings algorithm proceeds as follows:

1. Initialize $\boldsymbol{\theta}^{(0)}$ and the number T of iterations. Set $t = 1$.
2. Draw a random number $\boldsymbol{\theta}^*$ from the proposal density $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)})$ and accept it as the new state $\boldsymbol{\theta}^{(t)}$ with probability $\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)})$, otherwise set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.
3. Stop if $t = T$, otherwise set $t = t + 1$ and go to 2.

After a *burn-in phase* t_0, the random numbers $\boldsymbol{\theta}^{(t_0+1)}, \ldots, \boldsymbol{\theta}^{(T)}$ can be considered as a (correlated) sample from $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.

Bayesian Logit Model

- We consider the following simulated logit model with two covariates x_1 and x_2:

$$\begin{aligned} Y_i &= B(1, \pi_i) \qquad i = 1, \ldots, 500, \\ \pi_i &= \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}, \\ \eta_i &= -0.5 + 0.6\, x_{i1} - 0.3\, x_{i2}. \end{aligned}$$

- The covariates x_1 and x_2 are drawn independently from a standard normal distribution.
- We want to construct a Metropolis-Hastings algorithm to estimate the parameter $\boldsymbol{\beta} = (-0.5, 0.6, -0.3)'$ given this simulated data.

Bayesian Logit Model

- We specify independent diffuse priors $p(\beta_j) \propto$ const.
- The posterior is then proportional to the likelihood:

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) \propto \prod_{i=1}^{500} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}.$$

- As a proposal density for the Metropolis-Hastings algorithm we choose a 3-dimensional normal distribution, with the current state $\boldsymbol{\beta}^{(t-1)}$ as its mean.
- For its covariance matrix, we start with the diagonal matrix $\Sigma = \mathit{diag}(0.4^2, 0.4^2, 0.4^2)$.
- Figure 13 (first row) shows the first 2000 random numbers for β_0 and β_1 drawn with this proposal density.
- Since we have specified diffuse priors, Bayes estimates for the regression coefficients should not differ too much from the MLEs. Therefore, the MLEs are displayed as horizontal lines in the plots.

Bayesian Logit Model

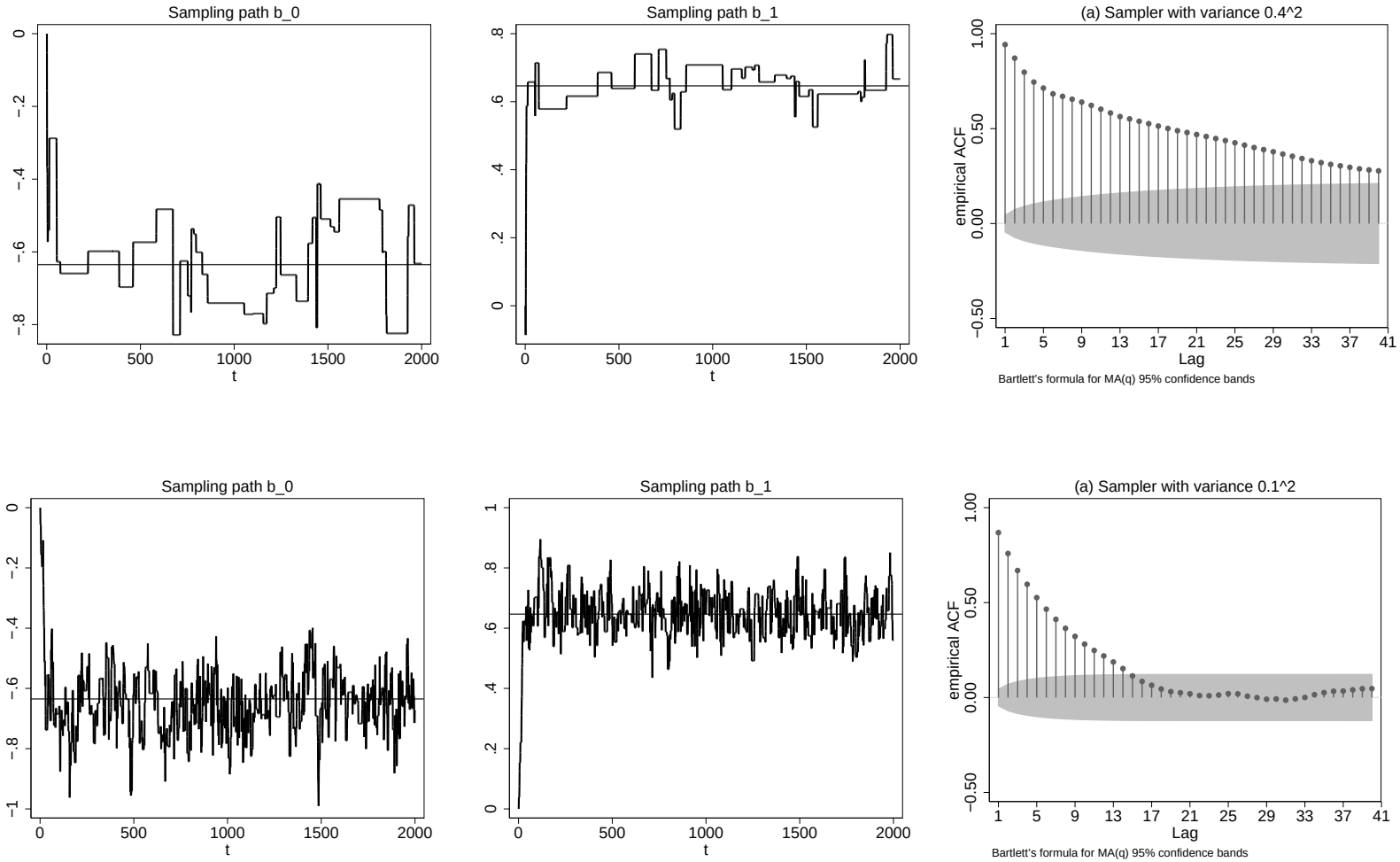


Figure: *Sampling paths for β_0 and β_1 for different MH algorithms. The third column shows the respective autocorrelation functions for β_1.*

Bayesian Logit Model

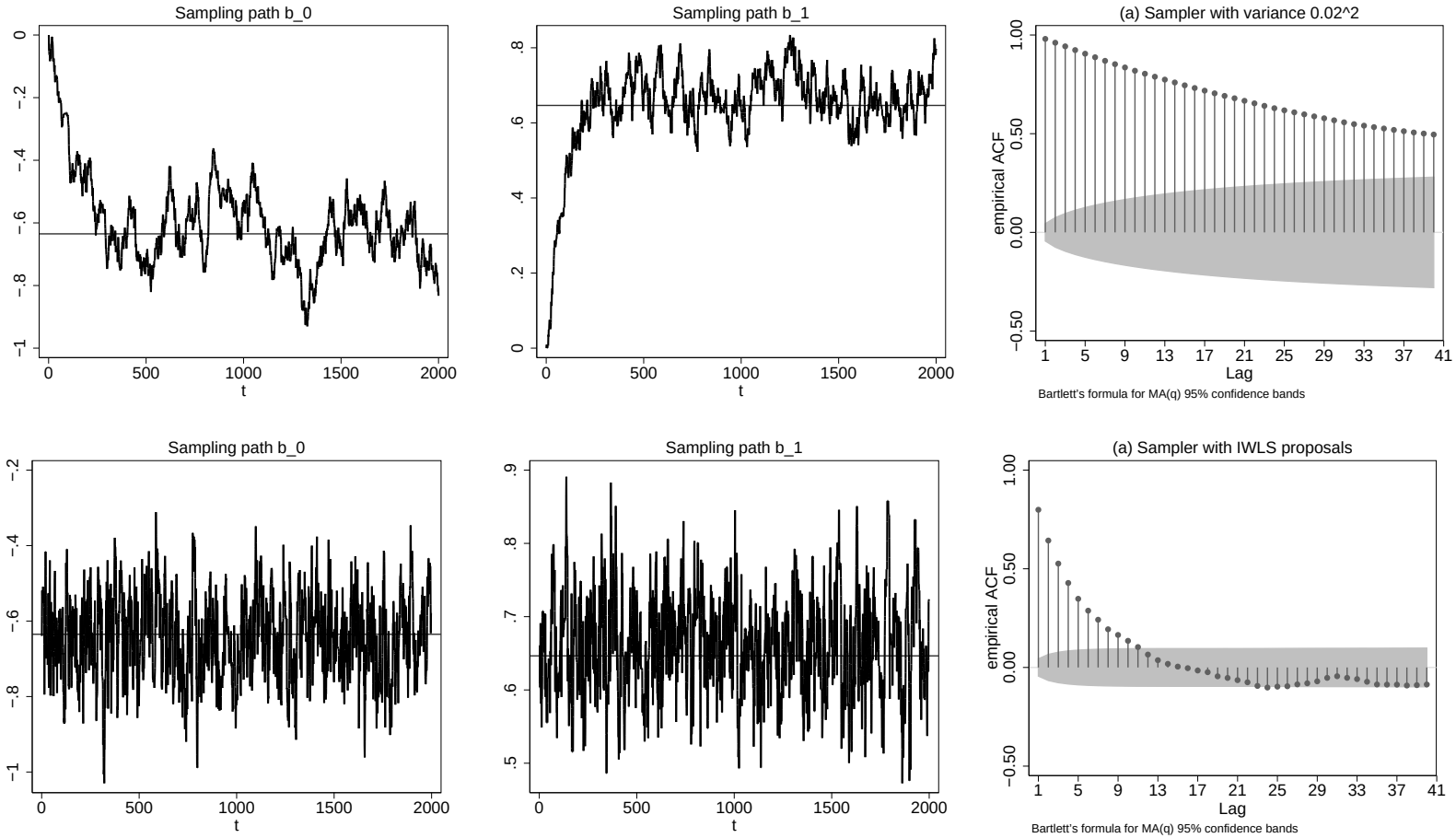


Figure: *Sampling paths for β_0 and β_1 for different MH algorithms. The third column shows the respective autocorrelation functions for β_1.*

Bayesian Logit Model

- Clearly, only a few of the proposed random numbers are accepted with this first algorithm, sometimes the state remains unchanged for more than 100 iterations.
- Thus, the acceptance probabilities are far too small.
- We obtain larger acceptance probabilities if the variances of the proposal density are decreased to $\Sigma = \mathit{diag}(0.1^2, 0.1^2, 0.1^2)$.
- The second row in Figure 13 shows the first 2000 random numbers for β_0 and β_1 resulting from this second MH algorithm.
- We recognize a short burn-in phase of about 50 iterations, followed by reasonable iterations with relatively large acceptance rates.

Bayesian Logit Model

- If we further decrease the variance to $\Sigma = diag(0.02^2, 0.02^2, 0.02^2)$, acceptance rates are further increased, but successive draws remain almost in the same state, see the first row in Figure 14.
- A useful and important tool for assessing the quality of MCMC algorithms is the autocorrelation function of the sample.
- Ideally, autocorrelations should rapidly converge to zero with increasing lags. The smaller the autocorrelation of successive parameters, the better the characteristics of the posterior can be estimated, based on the same length T of the sample.

Bayesian Logit Model

- For practical work, 'thinning' is carried out for the original sample, i.e. only every kth random number is kept in the sample, so that the remaining random numbers are almost uncorrelated. In this way, memory space can be saved without worsening estimation results.
- To generate an approximately uncorrelated sample of size 1000 with our second MH algorithm, we would have to generate about 20000 random numbers after a short burn-in phase and then keep only every 20th random number in the thinned sample.

Bayesian Logit Model

We can conclude the following:

- Small variances of the proposal density lead to high acceptance rates.
- In contrast, acceptance rates become small for large variances.
- For very large or very small variances autocorrelations of successive random numbers are high.
- The art of designing good MH algorithms is therefore the choice of appropriate proposal densities that combine high acceptance rates with low autocorrelations.
- Furthermore, automated methods are desirable that do not require subjective tuning of parameters of the proposal density.

Bayesian Logit Model

- An algorithm with these desirable properties is the MH algorithm based on IWLS proposals, see the last column of Fig. 14.
- Using this algorithm a Markov chain was generated and, after the burn-in phase, 20000 random numbers were drawn. Saving every 20th random number led to a thinned sample of size 1000. Based on this thinned sample all characteristics of interest of the posterior can be approximated.
- To approximate the posterior mean we compute the arithmetic means for the sample, resulting in $\hat{\boldsymbol{\beta}} = (-0.64, 0.65, -0.38)'$.
- Estimation of credible intervals can be based on the quantiles of the sampled random numbers. For example, we obtain 95% credible intervals by choosing the 2.5% quantiles as lower and 97.5% quantiles as upper bounds. This results in the intervals $[-0.87, -0.42]$, $[0.52 - 0.78]$ and $[-0.52, -0.26]$ for the sample generated in our example.