Advanced Quantitative Methods: Statistical estimators

Jos Elkink

University College Dublin

January 26, 2011

Outline

Estimation

"Problems of estimation are those in which it is required to estimate the value of one or more of the population parameters from a random sample of the population."

(Fisher 1922, 310)

Estimators

- Ordinary Least Squares (OLS)
- Generalized Least Squares (GLS)
- Maximum Likelihood (ML)
- Simulated Maximum Likelihood (SML)
- General Method of Moments (GMM)
- Bayesian (usually Markov Chain Monte Carlo) (MCMC)
- etc.

Estimator criteria

Given that there are often many possible estimators for a particular population value, we want to be able to **evaluate** which one is most appropriate.

Estimator criteria

Given that there are often many possible estimators for a particular population value, we want to be able to **evaluate** which one is most appropriate.

We can make a distinction between two types of criteria:

- **Finite sample** properties - how well does the estimator do given a limited sample size?
- **Asymptotic** properties - how well does the estimator do as the sample size gets infinitely large?

Estimator criteria

Given that there are often many possible estimators for a particular population value, we want to be able to **evaluate** which one is most appropriate.

We can make a distinction between two types of criteria:

- **Finite sample** properties - how well does the estimator do given a limited sample size?
- **Asymptotic** properties - how well does the estimator do as the sample size gets infinitely large?

(We will assume single parameter estimations (β), rather than multivariate ones ($\boldsymbol{\beta}$) for the remainder of these slides.)

Outline

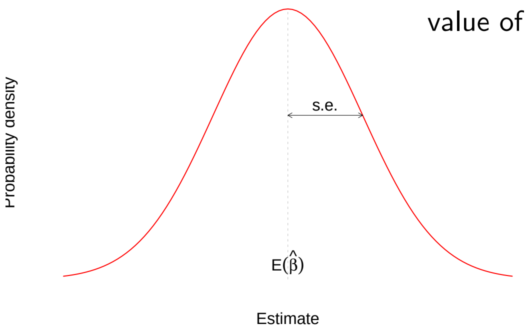
Sampling distribution

Imagine, that instead of having one sample, we take many samples.

If we do the same estimation in each of those randomly selected samples, we would get different results each time.

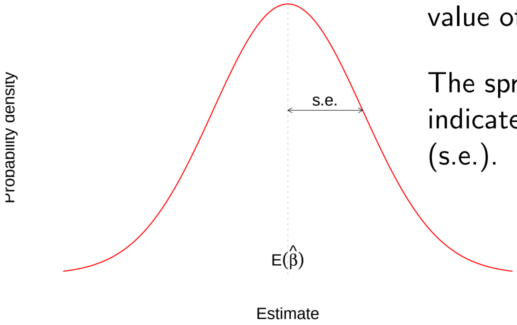
The distribution of these different estimates is the **sampling distribution** of the estimate.

Sampling distribution



The sampling distribution is a probability density function, with as mean the expected value of $\hat{\beta}$.

Sampling distribution



The sampling distribution is a probability density function, with as mean the expected value of $\hat{\beta}$.

The spread of this distribution is indicated by the **standard error** (s.e.).

$$se_{\hat{\beta}} = \sqrt{var(\hat{\beta})}$$

Unbiasedness

The **bias** of an estimator is the difference between the expected value of the sample distribution and the true value of the parameter to be estimated:

$$bias_{\hat{\beta}} = E(\hat{\beta}) - \beta$$

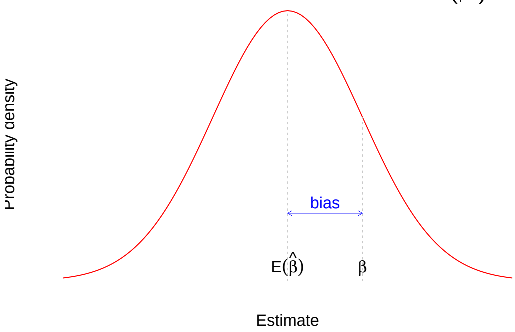
Unbiasedness

The **bias** of an estimator is the difference between the expected value of the sample distribution and the true value of the parameter to be estimated:

$$bias_{\hat{\beta}} = E(\hat{\beta}) - \beta$$

So an **unbiased estimator** is an estimator where $E(\hat{\beta}) = \beta$.

Bias: sampling distribution



For a biased estimator, $E(\hat{\beta}) \neq \beta$, and the bias is $E(\hat{\beta}) - \beta$.

Example: variance estimation

It can be shown (see notes) that, if s^2 is the **sample variance** and σ^2 the **population variance**, that:

$$E(s^2) = \frac{n-1}{n}\sigma^2,$$

in other words, s^2 is a **biased estimator** of σ^2.

Example: variance estimation

It can be shown (see notes) that, if s^2 is the **sample variance** and σ^2 the **population variance**, that:

$$E(s^2) = \frac{n-1}{n}\sigma^2,$$

in other words, s^2 is a **biased estimator** of σ^2.

$$E(\frac{n}{n-1}s^2) = \frac{n}{n-1}E(s^2) = \frac{n}{n-1} \cdot \frac{n-1}{n}\sigma^2 = \sigma^2,$$

so when we estimate the population variance, we calculate $\frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2$ instead of $\frac{1}{n}\sum_i^n (x_i - \bar{x})^2$.

Unbiasedness: vector of coefficients

If $\boldsymbol{\beta}$ is a vector of coefficients, rather than a single parameter β, it simply still holds that an unbiased estimator is one where $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ and a biased estimator where it is not.

Efficiency

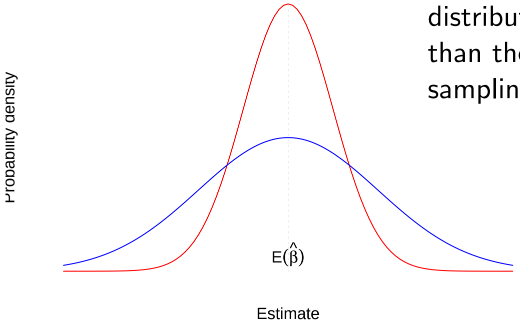
The estimator whose sampling distribution has the lowest variance is the more efficient estimator.

Efficiency

The estimator whose sampling distribution has the lowest variance is the more efficient estimator.

The most efficient unbiased estimator is called the **best unbiased** estimator.

Efficiency: sampling distribution



$E(\hat{\beta})$ is the same for both estimators, but the estimator with the red sampling distribution is more efficient than the one with the blue sampling distribution.

$$se_{\hat{\beta}_{blue}} > se_{\hat{\beta}_{red}}$$

BLUE

In the context of linear models, we often talk of the **best linear unbiased estimator** (BLUE), which is the estimator which is linear, unbiased, and has the lowest sampling variance of all possible unbiased linear estimators.

BLUE

In the context of linear models, we often talk of the **best linear unbiased estimator** (BLUE), which is the estimator which is linear, unbiased, and has the lowest sampling variance of all possible unbiased linear estimators.

When the assumptions underlying OLS hold, OLS is BLUE.

Efficiency: vector of coefficients

Often, $\boldsymbol{\beta}$ is a vector instead of just one parameter β. Instead of just the variance of β, we have a **variance-covariance matrix**:

$$var(\boldsymbol{\beta}) = \begin{bmatrix} var(\beta_1) & cov(\beta_1, \beta_2) & ... & cov(\beta_1, \beta_k) \\ cov(\beta_2, \beta_1) & var(\beta_2) & ... & cov(\beta_2, \beta_k) \\ ... & ... & ... & ... \\ cov(\beta_k, \beta_1) & cov(\beta_k, \beta_2) & ... & var(\beta_k) \end{bmatrix}$$

Efficiency: vector of coefficients

In this case we can consider various criteria for efficiency:

- smallest trace
- smallest determinant
- smallest variance of any linear combination of its elements
- some weighted sum of the variances and covariances

Efficiency: vector of coefficients

In this case we can consider various criteria for efficiency:

- smallest trace
- smallest determinant
- smallest variance of any linear combination of its elements
- some weighted sum of the variances and covariances

$$E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{W}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]$$

If **W** is selected such that this equation cannot lead to a negative result, minimizing this expectation leads to the most efficient estimator on all the above grounds.

(Kennedy 2008, 34)

Mean Square Error

Sometimes an estimator can be slightly biased, but be so much more efficient, that is becomes preferable.

Mean Square Error

Sometimes an estimator can be slightly biased, but be so much more efficient, that is becomes preferable.

So an estimator which is BLUE is not necessarily the most efficient estimator - it is simply the most efficient *unbiased* estimator.

Mean Square Error

Sometimes an estimator can be slightly biased, but be so much more efficient, that is becomes preferable.

So an estimator which is BLUE is not necessarily the most efficient estimator - it is simply the most efficient *unbiased* estimator.

The **Mean Square Error** (MSE) is the sum of the variance and the square of the bias of an estimator:

$$\begin{aligned} MSE_{\hat{\beta}} &= E[(\hat{\beta} - \beta)^2] \\ &= var(\hat{\beta}) + E(\hat{\beta} - \beta)^2 \end{aligned}$$

MSE: vector of coefficients

When $\boldsymbol{\beta}$ is a vector of coefficients instead of a single parameter β, we could look at the **MSE matrix**:

$$\mathbf{MSE} = E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'$$

MSE: vector of coefficients

When $\boldsymbol{\beta}$ is a vector of coefficients instead of a single parameter β, we could look at the **MSE matrix**:

$$\mathbf{MSE} = E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'$$

Or we could just look at the trace of this matrix.

Outline

Asymptotics

The **asymptotic properties** of an estimator concern the estimator's sampling distribution in extremely (or infinitely) large samples.

Asymptotics

The **asymptotic properties** of an estimator concern the estimator's sampling distribution in extremely (or infinitely) large samples.

We need these properties when we cannot provide the same proofs of unbiasedness, efficiency, etc. for small samples. To what extend the properties hold for small samples is then left for further exploration (e.g. through simulation).

Limits

"The real-valued sequence $\{x_n\}$ has the real number x^* for its **limit**, or converges to x^*, if for any positive ε, no matter how small, it is possible to find a positive integer N such that for all integers n greater than N, $|x_n - x^*| < \varepsilon$".

We can write:

$$\lim_{n \to \infty} x_n = x^*.$$

(Davidson & MacKinnon 1993: 102)

Limits: examples

$$\lim_{n \to \infty} \frac{1}{n} = 0$$

Limits: examples

$$\lim_{n \to \infty} \frac{1}{n} = 0$$

$$\lim_{n \to \infty} \frac{2n+1}{n-1} = 2$$

Limits: examples

$$\lim_{n\to\infty} \frac{1}{n} = 0$$

$$\lim_{n\to\infty} \frac{2n+1}{n-1} = 2$$

$$\lim_{n\to\infty} \frac{n-1}{n}\sigma^2 = \sigma^2$$

Probability limits

$$P(\|\mathbf{x}_n - \mathbf{x}^*\| > \varepsilon) < \delta,$$

whereby ε and δ are arbitrarily small, positive real numbers. We can write:

$$\operatorname*{plim}_{n\to\infty} \mathbf{x}_n = \mathbf{x}^*.$$

(Davidson & MacKinnon 1993: 103)

Convergence in distribution

$$\lim_{n \to \infty} P(\mathbf{x}_n \leq \mathbf{b}) = P(\mathbf{x}^* \leq \mathbf{b})$$

for any arbitrary $\mathbf{b}$. In other words, the probability distribution of $\{\mathbf{x}_n\}$ approaches that of $\mathbf{x}^*$ as n increases. This can be written as:

$$\mathbf{x}_n \xrightarrow{D} \mathbf{x}^*.$$

(Davidson & MacKinnon 1993: 107)

Asymptotic unbiasedness

Just was we can look at the bias in finite samples, we can also talk of the **asymptotic bias** in infinitely large samples:

$$asy.bias = \lim_{n \to \infty} E(\hat{\beta}) - \beta$$

(But see Greene 2003, 917)

Asymptotic unbiasedness: example

$$E(s^2) = \frac{n-1}{n}\sigma^2,$$

so s^2 is a **biased** estimator of σ^2.

Asymptotic unbiasedness: example

$$E(s^2) = \frac{n-1}{n}\sigma^2,$$

so s^2 is a **biased** estimator of σ^2.

$$\lim_{n\to\infty} E(S^2) = \lim_{n\to\infty} \frac{n-1}{n}\sigma^2 = \sigma^2,$$

so s^2 is an **asymptotically unbiased** estimator of σ^2.

(See for the asymptotic variance of this estimator: Greene 2003: 917-918)

Consistency

"A statistics satisfies the criterion of consistency, if, when it is calculated from the whole population, it is equal to the required parameter."

(Fisher 1922, 309)

Consistency

An estimator is **consistent** iff:

$$\operatorname*{plim}_{n\to\infty} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$$

Consistency

An estimator is **consistent** iff:

$$\underset{n\to\infty}{\text{plim}}\ \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$$

In other words, the **asymptotic variance** of $\hat{\boldsymbol{\beta}}$ becomes very small and $\hat{\boldsymbol{\beta}}$ is **asymptotically unbiased**, so that the probability distribution of $\hat{\boldsymbol{\beta}}$ "collapses" to a very tight distribution around $\boldsymbol{\beta}$.

Consistency: sample mean

The **Central Limit Theorem** states:

$$E(\bar{x}) = \mu_x \qquad var(\bar{x}) = \frac{\sigma^2}{n}$$

Consistency: sample mean

The **Central Limit Theorem** states:

$$E(\bar{x}) = \mu_x \qquad var(\bar{x}) = \frac{\sigma^2}{n}$$

therefore:

$$\operatorname*{plim}_{n\to\infty} E(\bar{x}) = \mu_x \qquad \operatorname*{plim}_{n\to\infty} var(\bar{x}) = 0$$

Consistency: sample mean

The **Central Limit Theorem** states:

$$E(\bar{x}) = \mu_x \qquad var(\bar{x}) = \frac{\sigma^2}{n}$$

therefore:

$$\operatorname*{plim}_{n\to\infty} E(\bar{x}) = \mu_x \qquad \operatorname*{plim}_{n\to\infty} var(\bar{x}) = 0$$

The mean of a random sample ($\bar{x}$) is a **consistent estimator** of the mean of the population (μ_x).

Efficiency

"The criterion of efficiency is satisfied by those statistics which, when derived from large samples, tend to a normal distribution with the least possible standard deviation."

(Fisher 1922, 310)

Asymptotic variance

The **asymptotic variance** of an estimator is:

$$\lim_{n \to \infty} var(\hat{\beta})$$

Asymptotic efficiency

"An estimator is **asymptotically efficient** if it is consistent, asymptotically normally distributed, and has an asymptotic covariance matrix that is not larger than the asymptotic covariance matrix of any other consistent, asymptotically normally distributed estimator."

(Greene 2003, 71)

Outline

Monte Carlo studies

Used to study **small-sample properties** of estimators when only asymptotics are known.

Based on computer **simulations** and due to computational costs only recently becoming common.

Monte Carlo: steps

1. Model the data-generating process (DGP)

Monte Carlo: steps

1. Model the data-generating process (DGP)
2. Generate artificial data sets

Monte Carlo: steps

1. Model the data-generating process (DGP)
2. Generate artificial data sets
3. Create estimates of the underlying parameters using the estimator you are testing

Monte Carlo: steps

1. Model the data-generating process (DGP)
2. Generate artificial data sets
3. Create estimates of the underlying parameters using the estimator you are testing
4. Assess the estimator's efficiency, bias and MSE relative to the (known) data

Monte Carlo: steps

1. Model the data-generating process (DGP)
2. Generate artificial data sets
3. Create estimates of the underlying parameters using the estimator you are testing
4. Assess the estimator's efficiency, bias and MSE relative to the (known) data
5. … or, for tests, check proportion Type I and Type II errors

Monte Carlo: example

In R, the linear model is estimated with the following command:

```
summary(m <- lm(y ~ x))
```

or explicitly without constant:

```
summary(m <- lm(y ~ 0 + x))
```

Monte Carlo: example

In R, the linear model is estimated with the following command:

```
summary(m <- lm(y ~ x))
```

or explicitly without constant:

```
summary(m <- lm(y ~ 0 + x))
```

We can test this estimator by creating fake datasets, following the steps outlined.

Monte Carlo example: data generating process

Data generating process (DGP):

$$
\begin{aligned}
x_i &\sim N(5, 2) \\
\varepsilon_i &\sim N(0, 1) \\
y_i &= 3x_i + \varepsilon_i
\end{aligned}
$$

Monte Carlo example: artificial datasets

Artificial datasets (*R* datasets of *N* cases each):

```
N <- 50
R <- 1000
for (i in 1:R) {
 x <- rnorm(N, 5, 2)
 e <- rnorm(N, 0, 1)
 y <- 3 * x + e
}
```

Monte Carlo example: estimations

Insert the estimation itself:

```
N <- 50
R <- 1000
estimates <- rep(NA, R)
for (i in 1:R) {
 x <- rnorm(N, 5, 2)
 e <- rnorm(N, 0, 1)
 y <- 3 * x + e
 estimates[i] <- coef(lm(y ~ 0 + x))
}
```

Monte Carlo example: assessment

- Efficiency

```
plot(density(estimates))
sd(estimates)
```

- Bias

```
mean(estimates - 3)
```

- Mean squared error (MSE)

```
var(estimates) + mean(estimates - 3)^2
```

Monte Carlo example: assessment

- Efficiency

```
plot(density(estimates))
sd(estimates)
```

- Bias

```
mean(estimates - 3)
```

- Mean squared error (MSE)

```
var(estimates) + mean(estimates - 3)^2
```

Probably not very useful without comparison to other methods.

Monte Carlo: exercise 1

Run Monte Carlo experiments with the following data generation process:

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \\
&= 1 + \frac{1}{2} x_i + \varepsilon_i \\
\varepsilon_i &\sim N(0,1) \\
x_i &\sim N(5,2)
\end{aligned}
$$

- Plot histograms of β_0 and β_1 ...
- ... for sample sizes $n = \{10, 100, 1000\}$
- Repeat with ε_i having a t-distribution with 1 d.f.

Monte Carlo: exercise 2

Run Monte Carlo experiments with the following data generation process:

$$
\begin{aligned}
x_i &= a + x_{i-1} + u_i \\
y_i &= b + y_{i-1} + v_i \\
x_0 = y_0 &= 0 \\
u_i &\sim N(0,1) \\
v_i &\sim N(0,1) \\
a = b &= \frac{1}{2}
\end{aligned}
$$

- Plot histograms of β_0 and β_1 when regressing y on x
- Repeat for $a = b = 0$

Outline

Least squares

We can look at the difference between the predicted and the observed values of the dependent variable:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

and we want this prediction error to be as small as possible.

Least squares

We can look at the difference between the predicted and the observed values of the dependent variable:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

and we want this prediction error to be as small as possible.

Many options:

- minimizing absolute errors, $min(|\mathbf{e}|)$
- minimizing squared errors, $min(\mathbf{e}'\mathbf{e})$
- using weights, $min(\mathbf{e}'\mathbf{W}\mathbf{e})$
- etc.

Least squares

We can look at the difference between the predicted and the observed values of the dependent variable:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

and we want this prediction error to be as small as possible.

Many options:

- minimizing absolute errors, $min(|\mathbf{e}|)$
- minimizing squared errors, $min(\mathbf{e}'\mathbf{e})$
- using weights, $min(\mathbf{e}'\mathbf{W}\mathbf{e})$
- etc.

OLS minimizes $\mathbf{e}'\mathbf{e}$.

R^2

R^2 represents the explained variance in $\mathbf{y}$, as a proportion of the total variance in $\mathbf{y}$.

R^2

Only appropriate when:

- using OLS for estimation
- explained variation is linear only
- there is an intercept in the model

R^2

R^2 represents the explained variance in $\mathbf{y}$, as a proportion of the total variance in $\mathbf{y}$.

- OLS maximizes R^2 implicitly

R^2

R^2 represents the explained variance in **y**, as a proportion of the total variance in **y**.

- OLS maximizes R^2 implicitly
- High R^2 can be caused by:
    - Using a lot of regressors
    - Including lagged dependent variable

R^2

R^2 represents the explained variance in **y**, as a proportion of the total variance in **y**.

- OLS maximizes R^2 implicitly
- High R^2 can be caused by:
  - Using a lot of regressors
  - Including lagged dependent variable
- No good equivalent for non-linear of models

R^2

R^2 represents the explained variance in $\mathbf{y}$, as a proportion of the total variance in $\mathbf{y}$.

- OLS maximizes R^2 implicitly
- High R^2 can be caused by:
    - Using a lot of regressors
    - Including lagged dependent variable
- No good equivalent for non-linear of models
- When interested in causal effect, high R^2 is not all too interesting

R^2

"These measures of goodness of fit have a fatal attraction. Although it is generally conceded among insiders that they do not mean a thing, high values are still a source of pride and satisfaction to their authors, however hard they may try to conceal these feelings."

(Cramer 1987, 253, as cited in Kennedy 2008, 27)

R^2 and least squares

Too much emphasis on these criteria can lead to **overfitting**, where you get excellent results for the sample at hand, but not if you would use the same estimates on any other data.

R^2 and least squares

Too much emphasis on these criteria can lead to **overfitting**, where you get excellent results for the sample at hand, but not if you would use the same estimates on any other data.

One way to reduce the problem of overfitting is to **split the sample** in two, use one half for estimation, and then use the estimated values to predict the dependent variable in the other half of the sample, and check errors and R^2.

Likelihood

"The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed."

(Fisher 1922, 310)

Maximum Likelihood

The likelihood is proportional to the probability of observing the data you have (give or take some arbitrarily small deviation), given some parameter estimate:

$$L(\beta|\mathbf{y},\mathbf{X}) = \alpha P(\mathbf{y},\mathbf{X}|\beta)$$

Maximum Likelihood

The likelihood is proportional to the probability of observing the data you have (give or take some arbitrarily small deviation), given some parameter estimate:

$$L(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \alpha P(\mathbf{y}, \mathbf{X}|\boldsymbol{\beta})$$

The **likelihood function** is thus proportional to the **probability density function** of the given sample, as a function of the parameter values.

Maximum Likelihood

The likelihood is proportional to the probability of observing the data you have (give or take some arbitrarily small deviation), given some parameter estimate:

$$L(\beta|\mathbf{y}, \mathbf{X}) = \alpha P(\mathbf{y}, \mathbf{X}|\beta)$$

The **likelihood function** is thus proportional to the **probability density function** of the given sample, as a function of the parameter values.

The estimator that maximizes this function also maximizes this probability density function and is the **Maximum Likelihood Estimator** (MLE or ML).

(Fisher 1922)

Sufficiency

"A statistic satisfies the criterion of sufficiency when no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter to be estimated."

(Fisher 1922, 310)

Computational costs

Used to be a major problem ...

... not so much anymore.

Still worth considering for very large datasets.

Computational costs

Used to be a major problem ...

... not so much anymore.

Still worth considering for very large datasets.

E.g. $|(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}|$ in a spatial regression model.