E 31501/4150
Properties of OLS estimators by Monte Carlo simulation
Ragnar Nymoen

Department of Economics, University of Oslo

19 January 2012
Repeated sampling

- Section 2.4.3 of the HGL book is called *Repeated sampling*
- The point is that by drawing repeated samples of size 40 from a large population of households, they repeat the food expenditure regression several times (they do 10 repetitions).
- The average of the estimates for $\beta_2$ (the 10 $b_2$ values) is 78.8
- The interpretation is that *if* the regression model is correctly specified, then 78.8 is a better estimate than a single one.
Monte Carlo analysis I

- Repeated sampling is costly in practice, but we can do something similar at a low cost:
  - We can use the computer to generate a large number of data sets (call it $M$),
  - estimate the regression model on all $M$ data sets and
  - average all of the $M$ estimates of $\beta_2$.

- Since the data sets are independent (the computer takes care of that), the average

$$\hat{b}_2 = \frac{\sum_{j=1}^{M} \hat{\beta}_{2j}}{M}.$$ 

will come close to the true $E(b_2)$ when $M$ is large. This is with reference to the Law of Large Numbers.
Monte Carlo analysis II

- The procedure in an example of Monte Carlo analysis.
- In the lecture we have shown that if the classical assumptions hold, the OLS estimator $b_2$ is unbiased:
  \[ E(\hat{\beta}_{2j}) = \beta_2 \]
- We can use Monte Carlo analysis to check this result.
- To do that, we instruct the computer to generate data in accordance with the regression model with classical assumptions.
- We say that the model is the data generating process, and we expect to find that:
  \[ b_2 \xrightarrow{M \to \infty} \beta_2 \]
Monte Carlo analysis III

- We can also use Monte Carlo simulation to check the theory of $\text{var}(\hat{\beta}_2)$ that we have developed. Does $\text{var}(\hat{\beta}_2)$ fall with how large the sample is for example.

- In this course we illustrate Monte Carlo methodology in two ways.
  - Manual Monte Carlo simulations, where $M$ is very small. But it still illustrates the idea
    - This will be done in two of the exercises to the seminars ($M = 10$ in those exercises)
  - Automatic Monte Carlo when $M$ can be as large as we want—The following
Monte Carlo 1: Deterministic X

\[
y_i = 10 + 2x_i + \varepsilon_i
\]
\[
\beta_1 \quad \beta_2
\]
\[
\varepsilon_i \sim N(0, 0.1^2)
\]
homoskedasticity and
no disturbance correlation
\[
x_i : \text{Deterministic}
\]
\[
n = 100
\]

We first investigate if (\(\tilde{b}_2 - 2\)) and (\(\tilde{b}_2 - 10\)) are close to 0 when \(M = 1000\) \(n = 100\)
We then increase \(\sigma^2\) and look at the effect on (\(\tilde{b}_2 - 2\)) and (\(\tilde{b}_2 - 10\))
We also investigate the effect on variances of the two estimators.
Monte Carlo 2: Stochastic X

\[ y_i = 10 + 2 x_i + \varepsilon_i \]

\[ \varepsilon_i \sim N(0, 0.1^2) \]

homoskedasticity and

no disturbance-correlation

\( x_i \): Stochastic

\( n = 100 \)
We can compare also numerical results from the two Monte Carlos. The estimate of the true expectation of $b_2$:

$$E(\tilde{b}_2) = \frac{\sum_{j=1}^{M} b_{2j}}{M}$$

Results for deterministic and stochastic (model 2) regression model:

- Monte Carlo 1: 2.0005, 0.0005
- Monte Carlo 2: 2.0003, 0.0003
The variance of the OLS estimators in the two experiments $(n = 100)$.

\[
\sqrt{\text{Var}(b_2)} \quad \sqrt{\text{Var}(b_2)}^{\text{Stata}}
\]

*Monte Carlo 1*  \hspace{1cm} 0.0098  \quad \approx 0.0098

*Monte Carlo 2*  \hspace{1cm} 0.0088  \quad \approx 0.0088

\[\sqrt{\text{Var}(b_2)}\] is the Monte Carlo estimate of \[\sqrt{\text{Var}(b_2)}\]. \[\sqrt{\text{Var}(b_2)}\] is the estimate we get from Stata, PcGive or another regression programme.
In our third experiment we keep $x_i$ stochastic and $n = 100$, but we increase $\sigma^2$ from 0.1 to 1, and we reduce $s^2_x$. Call the resulting Monte Carlo 3:

$$ bias \quad \sqrt{Var(b_2)} $$

<table>
<thead>
<tr>
<th></th>
<th>Monte Carlo 2</th>
<th>Monte Carlo 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sqrt{Var(b_2)}$</td>
<td>0.0003</td>
<td>0.0028</td>
</tr>
<tr>
<td>$Stata\sqrt{Var(b_2)}$</td>
<td>0.0088068</td>
<td>0.10017</td>
</tr>
</tbody>
</table>

would again give the correct estimate of the standard error of $b_2$. 

Mis-specification I

Next consider departures from the “classical assumptions”. Let us check what happens if

\[ \text{cov}(\varepsilon_i, \varepsilon_j) \neq 0 \]

The *Monte Carlo* is

\[
y_i = 10 + 2x_i + \varepsilon_i
\]

\[
\varepsilon_i = 0.6\varepsilon_{i-1} + \varepsilon_i' \]

\[
\varepsilon_i' \sim N(0, 0.1^2)
\]

disturbances are correlated

\[
x_i : \text{ Stochastic}
\]

\[
N = 100 \text{ (maximum sample)}
\]
Mis-specification II

In other respects we have the same situation as in *Monte Carlo 2*.

\[
\begin{array}{ccc}
\text{bias} & \sqrt{\text{Var}(b_2)} & \sqrt{\text{Var}(b_2)}^\text{Stata} \\
\hline
\text{Monte Carlo 2} & 0.003 & 0.0098 & \approx 0.0098 \\
\text{Monte Carlo 4} & 0.008 & 0.0142 & 0.010752 \\
\end{array}
\]

This shows that when \(\varepsilon_i\) and \(\varepsilon_{i-1}\) are positively correlated, we have \(E(b_2) = \beta_2\) as in the classical case.

- We also have consistency.
- But note that \(\text{Var}(b_2)\) is underestimated!
- To fix that, can modify the OLS estimator. However we have to leave that for later
Mis-specification III

- Heteroskedasticity leads to similar (limited) problem with the OLS estimator.
- This shows that correctness of specification is important.