Statistical Properties of OLS

Linear regression model:

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + u_i, \quad u_i \sim \text{IID}(0, \sigma^2), \quad i = 1, \ldots, N. \tag{1}$$

We may assume that the data were actually generated by the **data-generating process**, or **DGP**,

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta}_0 + u_i, \quad u_i \sim \text{NID}(0, \sigma_0^2), \quad i = 1, \ldots, N. \tag{2}$$

Meanings of "$\sim$", $\text{IID}(0, \sigma^2)$, $\text{NID}(0, \sigma_0^2)$, and the "0" subscripts.

A **model** is a set of DGPs, perhaps denoted by $\mathbb{M}$. It might consist of all DGPs like (2), where $\boldsymbol{\beta}$ takes some value in $\mathbb{R}^k$, σ^2 is a positive real number, and u_i follows a distribution with mean 0 and variance σ^2.

The set of DGPs (2) defines the **classical normal linear model**.

Bias and Unbiasedness

Let $\hat{\theta}$ be an estimator of some parameter θ, with true value θ_0. Then the **bias** of $\hat{\theta}$ is defined as $\mathrm{E}(\hat{\theta}) - \theta_0$.

If the bias of $\hat{\theta}$ is zero for every admissible value of θ_0, then $\hat{\theta}$ is said to be **unbiased**.

The model (1) can also be written, using matrix notation, as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{IID}(\boldsymbol{0}, \sigma^2 \mathbf{I}), \tag{3}$$

where $\boldsymbol{y}$ and $\boldsymbol{u}$ are N-vectors, $\boldsymbol{X}$ is an $N \times k$ matrix, $\boldsymbol{\beta}$ is a k-vector, and $\mathrm{IID}(\boldsymbol{0}, \sigma^2 \mathbf{I})$ is shorthand.

We can replace $\boldsymbol{y}$ by $\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}$ in $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$. This yields

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}) = \boldsymbol{\beta}_0 + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}. \tag{4}$$

The expectation of (4) is

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}_0 + \mathrm{E}\big((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}\big). \tag{5}$$

Thus $\hat{\boldsymbol{\beta}}$ is unbiased if and only if the second term on the r.h.s. is $\mathbf{0}$.

Unfortunately, few estimators in econometrics are unbiased.

If the matrix $\boldsymbol{X}$ is **fixed** or **nonstochastic**, then

$$\mathrm{E}\big((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}\big) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathrm{E}(\boldsymbol{u}). \tag{6}$$

For example, $\boldsymbol{X}$ might be $\boldsymbol{\iota}$, a vector of 1s.

In this case, $\hat{\boldsymbol{\beta}}$ is unbiased whenever $\mathrm{E}(\boldsymbol{u}) = \mathbf{0}$.

Can we treat the matrix $\boldsymbol{X}$ as fixed? Yes, for experimental data, where the experimenter chose $\boldsymbol{X}$ before nature generated $\boldsymbol{y}$.

But the assumption that $\boldsymbol{X}$ is fixed is often unreasonable. Some of the columns of $\boldsymbol{X}$ typically correspond to variables that are random.

A weaker assumption is that the regressors in $\boldsymbol{X}$ are **exogenous**, so that $\boldsymbol{X}$ is independent of $\boldsymbol{u}$. Therefore,

$$\mathrm{E}(\boldsymbol{u} \mid \boldsymbol{X}) = \mathbf{0}. \tag{7}$$

The expectation of the entire vector $\boldsymbol{u}$, that is, of every one of the u_i, is assumed to be zero conditional on the entire matrix $\boldsymbol{X}$.

Equation (7) is an **exogeneity assumption**. Given (7), it is easy to show that $\hat{\boldsymbol{\beta}}$ is unbiased.

Because $\mathrm{E}(\boldsymbol{X}^\top \mid \boldsymbol{X}) = \boldsymbol{X}^\top$, and $\mathrm{E}(\boldsymbol{u} \mid \boldsymbol{X}) = \mathbf{0}$, it is clear that $\mathrm{E}(\boldsymbol{X}^\top \boldsymbol{u}) = \mathbf{0}$. Similarly,

$$\mathrm{E}\big((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u} \mid \boldsymbol{X}\big) = \mathbf{0}. \tag{8}$$

We apply the Law of Iterated Expectations to obtain (8).

The exogeneity assumption (7) may be reasonable for **cross-section data**, but it makes no sense for **time-series data**.

For time-series data, we index observations by t.

Even if u_t is not related to current and past values of the regressors, it must be related to future values if current values of y_t affect them.

The **predeterminedness** condition

$$\mathrm{E}(u_t \mid \mathbf{X}_t) = 0, \quad t = 1, \ldots, T, \tag{9}$$

is much weaker than (7). We can say that the regressors are **predetermined** with respect to the disturbances.

The **first-order autoregressive**, or **AR(1)**, model is

$$y_t = \beta_1 + \beta_2 y_{t-1} + u_t, \quad u_t \sim \mathrm{IID}(0, \sigma^2). \tag{10}$$

In matrix terms, this is

$$\boldsymbol{y} = \beta_1 \boldsymbol{\iota} + \beta_2 \boldsymbol{y}_1 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{IID}(\mathbf{0}, \sigma^2 \mathbf{I}), \tag{11}$$

where $\boldsymbol{y}_1$ has typical element y_{t-1}.

For the AR(1) model, the predeterminedness condition (9) may well hold, but the exogeneity assumption (7) cannot possibly hold.

The AR(1) model is **dynamic** and involves a **lagged dependent variable**. OLS estimates of such models are never unbiased.

For the AR(1) model,

$$\hat{\beta}_2 = (\boldsymbol{y}_1^\top \boldsymbol{M}_\iota \boldsymbol{y}_1)^{-1} \boldsymbol{y}_1^\top \boldsymbol{M}_\iota \boldsymbol{y}, \quad \text{where} \quad \boldsymbol{M}_\iota = \mathbf{I} - \boldsymbol{\iota}(\boldsymbol{\iota}^\top \boldsymbol{\iota})^{-1} \boldsymbol{\iota}^\top. \tag{12}$$

If we replace $\boldsymbol{y}$ by $\beta_{10}\boldsymbol{\iota} + \beta_{20}\boldsymbol{y}_1 + \boldsymbol{u}$, we find that

$$\hat{\beta}_2 = \beta_{20} + (\boldsymbol{y}_1^\top \boldsymbol{M}_\iota \boldsymbol{y}_1)^{-1} \boldsymbol{y}_1^\top \boldsymbol{M}_\iota \boldsymbol{u}. \tag{13}$$

The second term here does not have expectation 0.

Because $\boldsymbol{y}_1$ is stochastic, we cannot simply move the expectations operator and then take the unconditional expectation of $\boldsymbol{u}$.

Because $\mathrm{E}(\boldsymbol{u} \mid \boldsymbol{y}_1) \neq \mathbf{0}$, we cannot take expectations conditional on $\boldsymbol{y}_1$ and then rely on the Law of Iterated Expectations.

Asymptotic Theory and Consistency

Exact results are often unavailable, so we rely on **asymptotic** results.

Asymptotic theory requires an **asymptotic construction**, which specifies how samples of arbitrarily large size are generated.

This is easy with cross-section data. Just draw more and more observations from a population of infinite size.

For pure time-series models, it is common to let the simulations run on forever. But does this make sense?

With an $M \times k$ regressor matrix, stack copies of $\boldsymbol{X}$ to create matrices of dimensions $2M \times k$, $3M \times k$, $4M \times k$, and so on.

In most cases, the parameters are held fixed as $N \to \infty$. However, it is sometimes necessary to let them change systematically with N.

Examples include time-series models with unit roots and cointegration and cross-section models with weak instruments.

Consider the model

$$y_t = \beta_1 + \beta_2 1/t + u_t, \quad u_t \sim \text{IID}(0, \sigma^2). \tag{14}$$

Here $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased, and the values of the second regressor range from 1 to $1/N$.

If we use (14) directly as an asymptotic construction, the regressor gets smaller and smaller as $N \to \infty$.

If we stack copies of the original $\boldsymbol{X}$, it remains bounded.

We discuss two forms of **stochastic convergence**: **Convergence in probability** and **convergence in distribution**.

The first allows us to find the **probability limit**, or **plim**, of a sequence of random variables.

Let $\{Y_N\}$ denote a sequence of random variables Y_N, $N = 1, \ldots, \infty$. If the sequence converges in probability, then

$$\operatorname*{plim}_{N\to\infty} Y_N = Y_\infty. \tag{15}$$

For equation (15) to be true, what we need is that, for all $\epsilon > 0$,

$$\lim_{N\to\infty} \Pr\big(|Y_N - Y_\infty| > \epsilon\big) = 0. \tag{16}$$

The limit here is the ordinary limit of a sequence of real numbers.

For any chosen ϵ, however small, we can find N^* large enough so that, for all $N > N^*$, the probability that $|Y_N - Y_\infty|$ is greater than ϵ is arbitrarily small.

In (15), Y_∞ is often deterministic, but it may be stochastic.

For a sequence of random vectors, or matrices, denoted by $\{\boldsymbol{Y}_N\}$, $\text{plim}_{N\to\infty} \boldsymbol{Y}_N = \boldsymbol{Y}_\infty$ means that

$$\lim_{N\to\infty} \Pr\big(\|\boldsymbol{Y}_N - \boldsymbol{Y}_\infty\| > \epsilon\big) = 0. \tag{17}$$

Here $\|\cdot\|$ denotes the Euclidean norm of a vector. It can also denote a matrix norm.

A second, weaker, form of convergence is **convergence in distribution**, or **convergence in law**, or **weak convergence**.

The random variables themselves no longer converge, but instead the sequence of CDFs of the random variables converges.

If the random variables are IID, then they all have the same CDF, and so the limiting CDF is simply the CDF of each element of the sequence.

In general, a sequence $\{Y_N\}$ of scalar random variables converges in distribution to the distribution characterized by the CDF F if

$$\lim_{N\to\infty} F_N(x) = F(x), \tag{18}$$

where F_N is the CDF of Y_N, for all real x at which F is continuous.

We can write

$$Y_N \xrightarrow{d} F. \tag{19}$$

If a sequence $\{Y_N\}$ converges in probability, it also converges in distribution.

But convergence in distribution of a sequence does not imply convergence in probability.

The two concepts coincide whenever the limiting distribution is **degenerate**. All the probability mass is concentrated on one single point, which is a **nonstochastic plim**.

Suppose Z_t is a random variable equal to 1 if a coin comes up heads, and equal to 0 if it comes up tails. After N tosses, the proportion of heads is just

$$Y_N \equiv \frac{1}{N} \sum_{t=1}^{N} Z_t. \tag{20}$$

If the coin really is unbiased, $\mathrm{E}(Y_N) = 1/2$. Thus, if Y_N converges to anything, it must be to $1/2$.

Laws of Large Numbers

Suppose $\bar{y}_N$ is the sample mean of y_i, $i = 1, \ldots, N$, a sequence of random variables, each with expectation μ_y.

Provided the y_i are independent (or at least, not too dependent), a law of large numbers, or LLN, would state that

$$\operatorname*{plim}_{N\to\infty} \bar{y}_N = \operatorname*{plim}_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} y_i = \mu_y. \tag{21}$$

$\bar{y}_N$ has a nonstochastic plim equal to the common expectation of each of the y_i.

If the y_i are IID, with variance σ^2,

$$\mathrm{E}(\bar{y}_N) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{E}(y_i) = \frac{1}{N} \sum_{i=1}^{N} \mu_y = \mu_y. \tag{22}$$

It is easy to see that

$$\text{Var}(\bar{y}_N) = \frac{1}{N^2} \sum_{i=1}^{N} \sigma^2 = \frac{1}{N} \sigma^2. \tag{23}$$

Thus $\bar{y}_N$ has mean μ_y and a variance which tends to zero as $N \to \infty$.

In the limit, $\bar{y}_N$ becomes a nonstochastic quantity equal to μ_y.

As $N \to \infty$, we are collecting more information about $\text{E}(y_i)$, with each y_i providing a smaller fraction of that information.

Eventually, the random components of the individual y_i cancel out, and the sample mean $\bar{y}_N$ converges to the population mean μ_y.

We need assumptions to prevent any of the y_i from having too much impact on $\bar{y}_N$. Thus $\text{Var}(y_i)$ must be bounded from above.

We also need to assume that there is not too much dependence among the y_i, to ensure that the random components $y_i - \mu_y$ cancel out.

The assumption that the y_i are $\text{IID}(0, \sigma^2)$ is sufficient.

Convenient Properties of Plims

Suppose that $\{Y_N\}$, $N = 1, \ldots, \infty$, is a sequence of random variables with nonstochastic plim Y_∞, and $\eta(Y_N)$ is a smooth function of Y_N. Then $\operatorname{plim} \eta(Y_N) = \eta(Y_\infty)$.

If $\{Y_N\}$ and $\{Z_N\}$ converge in probability, then

$$\operatorname{plim} Y_N Z_N = \operatorname{plim} Y_N \operatorname{plim} Z_N. \tag{24}$$

These features of plims are not shared by expectations. When $\eta(\cdot)$ is nonlinear,

$$\mathrm{E}\big(\eta(Y)\big) \neq \eta\big(\mathrm{E}(Y)\big). \tag{25}$$

Also, $\mathrm{E}(YZ) \neq \mathrm{E}(Y)\,\mathrm{E}(Z)$ unless Y and Z are independent.

Many stochastic quantities do not have plims unless we divide them by N or, perhaps, by some power of N.

Consider the matrix $\boldsymbol{X}^\top \boldsymbol{X}$. Each element of this matrix is a sum of N products of, say, X_{ij} with X_{il}.

As $N \to \infty$, such a sum would tend to infinity as well. Therefore, the matrix $\boldsymbol{X}^\top \boldsymbol{X}$ does not generally have a plim. Instead, we assume that

$$\operatorname*{plim}_{N\to\infty} \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}, \tag{26}$$

where $\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}$ is a finite nonstochastic matrix.

Each element of the matrix $N^{-1}\boldsymbol{X}^\top \boldsymbol{X}$ is now an average of N numbers:

$$\left(\frac{1}{N} \boldsymbol{X}^\top \boldsymbol{X} \right)_{jl} = \frac{1}{N} \sum_{i=1}^{N} x_{ij} x_{il}. \tag{27}$$

We are implicitly assuming that an LLN holds for the sequences generated by the squares of the regressors and their cross-products.

The Same-Order Notation

Many quantities in econometrics change systematically as the sample size changes. They can do so at different rates.

The **same-order relation** provides a very convenient way to deal with such quantities.

Let $f(N)$ be a real-valued function of the integer $N > 0$, and r be a rational number.

Then $f(N)$ is of the same order as N^r if there exists a constant K, independent of N, and a positive integer N^* such that

$$\left| \frac{f(N)}{N^r} \right| < K \text{ for all } N > N^*. \tag{28}$$

When $f(N)$ is of the same order as N^r, we can write $f(N) = O(N^r)$.

Many quantities in econometrics are stochastic. For them, we need the **stochastic same-order relation**.

Let $\{a_N\}$ be a sequence of random variables indexed by N. Then a_N is of order N^r in probability if, for all $\epsilon > 0$, there exist a constant K and a positive integer N^* such that

$$\Pr\left(\left|\frac{a_N}{N^r}\right| > K\right) < \epsilon \text{ for all } N > N^*. \tag{29}$$

When a_N is of order N^r in probability, we can write $a_N = O_p(N^r)$.

The deterministic and stochastic same-order relations have the same properties. ETM often uses "$O(\cdot)$" when it should be using "$O_p(\cdot)$."

We can manipulate the same-order relations as if they were simply powers of N. This makes them extremely useful.

Suppose $f(N)$ and $g(N)$ are $O(N^r)$ and $O(N^q)$, respectively. Then

$$f(N)g(N) = O(N^r)O(N^q) = O(N^{r+q}) \tag{30}$$

and

$$f(N) + g(N) = O(N^r) + O(N^q) = O(N^{\max(r,q)}). \tag{31}$$

We have assumed that $N^{-1}\boldsymbol{X}^\top\boldsymbol{X}$ has a probability limit of $\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}$, which is a finite, positive definite, deterministic matrix.

Because each element of $N^{-1}\boldsymbol{X}^\top\boldsymbol{X}$ is $O_p(1)$ (or maybe $O(1)$ if the regressors are not stochastic), we can write $\boldsymbol{X}^\top\boldsymbol{X} = O_p(N)$.

Since $O(1) = O(N^0)$, the same-order relationship is still about N even when N does not explicitly appear.

In general, sums of random variables that *do not* have mean zero are $O_p(N)$. But sums of random variables that *do* have mean zero are $O_p(N^{1/2}) = O_p(\sqrt{N})$.

If we assume that $\boldsymbol{X}^\top \boldsymbol{y} = O_p(N)$, we see that

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = O_p(N^{-1}) O_p(N) = O_p(1). \tag{32}$$

Equation (32) says that $\hat{\boldsymbol{\beta}}$ does not systematically get larger or smaller as $N \to \infty$.

Econometricians mainly use same-order notation to make asymptotic theory easier. For example, suppose that $\hat{\theta} = (A + B)/\sqrt{C + D}$.

If we can show that $A = O_p(N^{1/2})$, $B = O_p(1)$, $C = O_p(N)$, and $D = O_p(N^{1/2})$, then it must be the case that

$$\hat{\theta} \overset{a}{=} \frac{A}{\sqrt{C}} = O_p(1). \tag{33}$$

The asymptotic approximation is likely to be good if $A - B \cong A$ and $D \ll C$. But it might be very poor if either or both of these conditions does not hold.

Consistency

Even when the OLS estimator is biased, it may turn out to be **consistent**.

Given a model $\mathbb{M}$, an estimator $\hat{\boldsymbol{\beta}}$, and an asymptotic construction that allows $\hat{\boldsymbol{\beta}}$ to be defined for arbitrary sample size N, $\hat{\boldsymbol{\beta}}$ is consistent if

$$\operatorname*{plim}_{N\to\infty} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0, \tag{34}$$

where $\boldsymbol{\beta}_0$ is the value of $\boldsymbol{\beta}$ associated with the actual DGP, which must belong to $\mathbb{M}$.

When the DGP is a special case of a regression model with $\boldsymbol{\beta} = \boldsymbol{\beta}_0$,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}. \tag{35}$$

To prove that $\hat{\boldsymbol{\beta}}$ is consistent, we need to show that the plim of the second term is zero.

Neither $\boldsymbol{X}^\top \boldsymbol{X}$ nor $\boldsymbol{X}^\top \boldsymbol{u}$ has a plim. However, we can divide both of them by N. Then the plim of the second term in (35) is

$$\left(\operatorname*{plim}_{N\to\infty} \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{X} \right)^{-1} \operatorname*{plim}_{N\to\infty} \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{u} = \left(\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}} \right)^{-1} \operatorname*{plim}_{N\to\infty} \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{u} = \boldsymbol{0}. \tag{36}$$

Here we assume that $\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}$ is nonsingular. Note that, even if $\boldsymbol{X}^\top \boldsymbol{X}$ is nonsingular for any finite $N \geq k$, $\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}$ can be singular.

The second equality in (36) depends on the assumption that $\mathrm{E}(\boldsymbol{X}_i^\top u_i \mid \boldsymbol{X}_i) = \boldsymbol{0}$. Then the Law of Iterated Expectations tells us that $\mathrm{E}(\boldsymbol{X}_i^\top u_i) = \boldsymbol{0}$.

Assuming that we can apply a law of large numbers,

$$\operatorname*{plim}_{N\to\infty} \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{u} = \operatorname*{plim}_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i^\top u_i = \boldsymbol{0}. \tag{37}$$

Consistency is neither weaker nor stronger than unbiasedness.

Consider again the model

$$y_t = \beta_1 + \beta_2 1/t + u_t, \quad u_t \sim \text{IID}(0, \sigma^2). \tag{38}$$

$\hat{\beta}_1$ and $\hat{\beta}_2$ are evidently unbiased.

But $\hat{\beta}_2$ is not consistent if we use the, apparently natural, asymptotic construction in which $t \to \infty$.

As $N \to \infty$, each observation provides less and less information about β_2, which causes $\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}$ to be singular.

Therefore, (36) does not hold, and the second term on the r.h.s. of (35) does not have a plim of zero.

Note that $\hat{\beta}_1$ is consistent even though $\hat{\beta}_2$ is not.

However, $\hat{\beta}_2$ is consistent under the alternative asymptotic construction in which we stack copies of the original $\boldsymbol{X}$ matrix.

There are two types of **inconsistent estimators**:

- An estimator, possibly unbiased, that does not tend to any nonstochastic plim.
- An estimator that tends to the wrong nonstochastic plim.

Suppose we estimate the population mean, μ, from a sample $y_i, 1 = 1, \ldots, N$. The sample mean $\bar{y}$ is unbiased and consistent under reasonable assumptions.

Now consider three (not very sensible) estimators:

$$\hat{\mu}_1 = \frac{1}{N+1} \sum_{i=1}^{N} y_i. \tag{39}$$

$\hat{\mu}_1$ is biased but consistent. It equals $N/(N+1)$ times $\bar{y}$. Thus its mean is $\big(N/(N+1)\big)\mu$, which tends to μ as $N \to \infty$, and it is consistent whenever $\bar{y}$ is.

$$\hat{\mu}_2 = \frac{1.01}{N} \sum_{i=1}^{N} y_i. \tag{40}$$

$\hat{\mu}_2$ is clearly biased and inconsistent. Its mean is 1.01μ, and it actually tends to a plim of 1.01μ as $N \to \infty$.

$$\hat{\mu}_3 = 0.01 y_1 + \frac{0.99}{N-1} \sum_{i=2}^{N} y_i. \tag{41}$$

$\hat{\mu}_3$ is clearly unbiased, since it is a weighted average of two estimators, y_1 and the average of y_2 through y_N, each of which is unbiased.

But it is not consistent, because it does not converge to a nonstochastic plim. Instead, it converges to the random quantity $0.99\mu + 0.01 y_1$.

Some types of inconsistency are unavoidable. For example, individual fixed effects generally cannot be estimated consistently.