ECON 468

CHAPTER 3: STATISTICAL PROPERTIES OF OLS

In this chapter, we consider the statistical properties of OLS, ones that depend on how the data were actually generated.

Topics covered:

- unbiased estimators of OLS
- consistent estimators of OLS
- covariance matrix of the OLS parameter estimates
- efficiency of OLS estimator
- residuals and error terms
- misspecification of linear regression models

3.1 INTRODUCTION

Linear regression model: $y_t = X_t \beta + u_t, \quad u_t \sim IID(0, \sigma^2)$

- coefficient vector β takes some value in $\mathbb{R}^k$
- variance σ^2 is positive and real
- u_t are statistically independent and all follow the same distribution, $E(u_t) = 0$, $Var(u_t) = \sigma^2$.

Classical normal linear model: $y_t = X_t \beta_0 + u_t, \quad u_t \sim NID(0, \sigma_0^2)$.

- β and σ^2 have specific values β_0 and σ_0^2 respectively,
- u_t are normally, independently and identically distributed

All of the results proved in this chapter apply to the linear regression model with no normality assumption. It is assumed that whatever model we are studying, it is correctly specified, ie true DGP $\in$ model.

3.2 ARE OLS ESTIMATORS UNBIASED?

Suppose that $\hat{\theta}$ is an estimator of some parameter θ, the true value of which is θ_0. The bias of $\hat{\theta}$ is defined $E(\hat{\theta}) - \theta_0$, the expectation of $\hat{\theta}$ minus the true value of θ.

If bias of an estimator is 0 for every admissible value of θ_0, then the estimator is said to be unbiased.

Is the OLS estimator $\hat{\beta} = (X^TX)^{-1}X^Ty$ unbiased?

Replace y by whatever it is equal to under the DGP that is assumed to have generated the data:

$$y = X\beta_0 + u$$

$$\hat{\beta} = (X^TX)^{-1}X^T(X\beta_0 + u)$$

$$= \beta_0 + (X^TX)^{-1}X^Tu$$

$$E(\hat{\beta}) = E(\beta_0 + (X^TX)^{-1}X^Tu)$$

$$= \beta_0 + E((X^TX)^{-1}X^Tu)$$

$\hat{\beta}$ is unbiased iff the second term, $E((X^TX)^{-1}X^Tu) = \underset{\sim}{0}$.

For $E((X^TX)^{-1}X^Tu) = 0$ to hold, we have to make some assumptions:

1. X is non-stochastic / fixed. — strong assumption, but unreasonable
   If X is fixed, then $(X^TX)^{-1}X^T$ is not random and can be determined.
   $E((X^TX)^{-1}X^Tu) = (X^TX)^{-1}X^TE(u)$
   If we further assume that $E(u) = 0$, the second term thus reduces to $\underset{\sim}{0}$.
   However, this assumption is not a reasonable assumption to make; it is more common that X contains some variables that are no less random than y itself!

2. Explanatory variables in X are exogenous — weaker assumption
   Exogeneity of X implies that any randomness in the DGP that generated X is independent of the error terms u in the DGP for y.
   Hence, $E(u|X) = 0$. (exogeneity condition)
   $E((X^TX)^{-1}X^Tu) = E(E((X^TX)^{-1}X^Tu|X))$
   $= E((X^TX)^{-1}X^T\underbrace{E(u|X)}_{=0})$ by Law of It. Exp.
   $= 0$

3. Predeterminedness Condition: $E(u_t|X_t) = 0$.
   Weaker than the exogeneity condition as it only rules out the possibility that it may depend on their values for the current observation.

The OLS estimator $\hat{\beta}$ is unbiased if we assume that X is exogenous.
If we are not prepared to go beyond the predeterminedness assumption, then we will find that $\hat{\beta}$ is, in general, biased.

Many regression models for time series data include one or more lagged variables among the regressors.

$y_t = X_t\beta + \rho y_{t-1} + u_t$, y_{t-1} is the lagged dependent variable.

The predeterminedness condition can be assumed in this model, we are essentially saying that $E(u_t) = 0$ for every possible value of y_{t-1}. y_{t-1} is realized before u_t, and its realized value has no impact on the expectation of u_t. $E(u_t \mid y_{t-1}) = 0$.

But $E(u \mid y_{-1}) \neq 0$ since y_{t-1} depends on u_{t-1}, u_{t-2}, ...; ie the lagged dependent variable is not exogenous.

The exogeneity assumption does not hold for models that deal with time series data.

3.3 ARE OLS ESTIMATORS CONSISTENT?

A consistent estimator is one for which the estimate tends to the quantity being estimated as the size of the sample tends to infinity. Thus, if the sample size is large enough, we can be confident that the estimate will be close to the true value.

The OLS estimator $\hat{\beta}$ will often be consistent even if it is biased.

In order to say what happens to a stochastic quantity that depends on n as $n \to \infty$, we need to introduce the concept of a probability limit.

Law of Large Numbers (LLN)

Suppose that $\bar{x}$ is the sample mean of x_t, $t = 1, \ldots, n$, a sequence of r.v.'s each with expectation μ. Then, provided that the x_t are independent, a LLN would state that:

$$\operatorname*{plim}_{n\to\infty} \bar{x} = \operatorname*{plim}_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} x_t = \mu.$$

$\bar{x}$ has a nonstochastic plim which is equal to the common expectation of each of the x_t.

As $n \to \infty$, we are collecting more and more information about the mean of the x_t, with each individual observation providing a smaller and smaller fraction of that information. Thus, eventually the randomness in the

individual x_t cancels out, and the sample mean $\bar{x}$ converges to the population mean μ.

For this to happen, we need to make some assumption in order to prevent any one of the x_t from having too much impact on $\bar{x}$. So, we assume that x_t are IID.

Stochastic convergence: convergence of a sequence of r.v.s.
- Almost-sure convergence: P(convergence to expectation) = 1.
- convergence in probability: $\exists N$ s.t. $\forall n > N$, $P(|\frac{1}{n}\sum_{i=1}^{n} z_i - E(z)| > \varepsilon) < \varepsilon$, $\forall \varepsilon > 0$. In words, the probability that the mean differs from the expectation by some tolerance level ε is less than ε.

Is $\hat{\beta}$ consistent?

Recall that $\hat{\beta} = (X^TX)^{-1}X^Ty$

Replacing y, $\hat{\beta} = (X^TX)^{-1}X^T(X\beta_0 + u)$

$= \beta_0 + (X^TX)^{-1}X^Tu$

If $\hat{\beta}$ is consistent, then $\operatorname{plim}_{n\to\infty} (X^TX)^{-1}X^Tu = 0$.

- strongly consistent $\hat{\beta}$: $\lim_{n\to\infty} \hat{\beta} = \beta$
- weakly consistent $\hat{\beta}$: $\operatorname{plim}_{n\to\infty} \hat{\beta} = \beta$

Evaluating $\operatorname{plim}_{n\to\infty} (X^TX)^{-1}X^Tu = 0$:

Since $(X^TX)^{-1}$ and X^Tu both do not have a probability limit, we divide them by n to convert them into quantities that, under reasonable assumptions, will have nonstochastic plims.

$$(\operatorname{plim}_{n\to\infty} \frac{1}{n}X^TX)^{-1}(\operatorname{plim}_{n\to\infty} \frac{1}{n}X^Tu) = (S_{X^TX})^{-1} \operatorname{plim}_{n\to\infty} \frac{1}{n}X^Tu$$

$= 0$, where S_{X^TX} is a nonstochastic matrix with full rank k.

The predetermined regressors condition tells us that $E(X_t^Tu_t \mid X_t) = 0 \Rightarrow E(X_t^Tu_t) = 0$.

Hence $\operatorname{plim}_{n\to\infty} \frac{1}{n}X^Tu = \operatorname{plim}_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} X_t^Tu_t = 0$.

example: $y_t = \alpha + \beta \frac{1}{t} + u_t$, $u_t \sim IID(0, \sigma^2)$

- $\hat{\alpha}$ and $\hat{\beta}$ are unbiased OLS estimators.
- β is not consistent:
  as $n \to \infty$, each observation provides less and less information about β_2, as $\frac{1}{t} \to 0$ and varies less and less across observations as t becomes larger.
- α is consistent:
  as $n \to \infty$, we get an amount of information about α roughly proportional to n.

3.4 COVARIANCE MATRIX OF OLS ESTIMATES

Although it is valuable to know that the OLS estimator $\hat{\beta}$ is either unbiased, or under weaker conditions, consistent, this information by itself is not very useful. We need to know, at least approximately, how $\hat{\beta}$ is distributed.

The covariance matrix $Var(\underset{\sim}{b})$ of a random k-vector $\underset{\sim}{b}$, with element b_i, organizes all the central second moments of the b_i into a $k \times k$ symmetric matrix. The i^{th} diagonal element in $Var(\underset{\sim}{b})$ is $Var(b_i)$, the variance of b_i. The ij^{th} off-diagonal element of $Var(b)$ is $Cov(b_i, b_j)$.

$Cov(b_i, b_j) = E[(b_i - E(b_i))(b_j - E(b_j))]$.

In matrix notation: $Var(\underset{\sim}{b}) = E((\underset{\sim}{b} - E(\underset{\sim}{b}))(\underset{\sim}{b} - E(\underset{\sim}{b}))^T)$

If $E(\underset{\sim}{b}) = \underset{\sim}{0}$, $Var(\underset{\sim}{b}) = E(\underset{\sim}{b}\,\underset{\sim}{b}^T)$

If b_i and b_j are statistically independent, $Cov(b_i, b_j) = 0$.

Correlation between b_i and b_j: $Corr(b_i, b_j) = \frac{Cov(b_i, b_j)}{\sqrt{Var(b_i)\,Var(b_j)}}$

- $-1 \leq Corr(b_i, b_j) \leq 1$.
- all elements on the principal diagonal of the correlation matrix is 1.

$Var(\underset{\sim}{b})$ must be a positive semidefinite matrix. In most cases, covariance matrices and correlation matrices are positive definite.

Hilroy

A $k \times k$ symmetric matrix A is said to be **positive definite** if $\forall$ non-zero k-vectors $\underset{\sim}{x}$, $\underset{\sim}{x}^T A \underset{\sim}{x} > 0$. It is **positive semidefinite** if $\underset{\sim}{x}^T A \underset{\sim}{x} \geq 0$.

($\underset{\sim}{x}^T A \underset{\sim}{x}$: scalar known as quadratic form)

Any matrix of the form $B^T B$ is positive semidefinite.

Proof: $B^T B = (B^T B)^T$, it is symmetric

$$\underset{\sim}{x}^T B^T B \underset{\sim}{x} = (B\underset{\sim}{x})^T B\underset{\sim}{x}$$
$$= \| B\underset{\sim}{x} \|^2 \geq 0 \quad \blacksquare$$

example: identity matrix I.

$\underset{\sim}{x}^T I \underset{\sim}{x} = \sum_{i=1}^{k} x_i^2$ which is always positive for all non-zero vectors $\underset{\sim}{x}$.

The inverse of a positive definite matrix always exists.

Proof: If A is singular, $\exists \underset{\sim}{x} \neq \underset{\sim}{0}$ s.t. $A\underset{\sim}{x} = \underset{\sim}{0}$. But if A is positive definite, $\underset{\sim}{x}^T A \underset{\sim}{x} \neq \underset{\sim}{0}$, contradiction. Hence A cannot be singular.

A^{-1} is also positive definite when A is positive definite.

OLS covariance matrix

error covariance matrix: If error terms are IID, they all have the same variance σ^2 and the covariance of any pair of them is 0. Thus, the covariance matrix would have values σ^2 along the main diagonal and 0 everywhere else.

"white noise" disturbance, homoskedastic

$$\text{Var}(\underset{\sim}{u}) = E(\underset{\sim}{u}\underset{\sim}{u}^T) = \sigma^2 I.$$

Notice that this result does not require $\underset{\sim}{u}$ to be independent!

covariance matrix of $\hat{\beta}$: If we assume X is exogenous (ie $E(u|X) = 0$), we can calculate $\text{Var}(\hat{\beta})$ in terms of $\text{Var}(\underset{\sim}{u})$.

Recall that $\hat{\beta} - \beta_0 = (X^T X)^{-1} X^T u$.

We assume that $\hat{\beta}$ is unbiased, ie $E(\hat{\beta} - \beta_0) = 0 \Rightarrow E(\hat{\beta}) = \beta_0$

Thus, $\text{Var}(\hat{\beta}) = E((\hat{\beta} - \beta_0)(\hat{\beta} - \beta_0)^T)$

$$= E[(X^T X)^{-1} X^T u (u^T X (X^T X)^{-1})]$$
$$= (X^T X)^{-1} X^T E(uu^T) X (X^T X)^{-1}$$
$$= (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1}$$
$$= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1}$$
$$= \sigma^2 (X^T X)^{-1}$$

$\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$ under the assumption that it is a linear regression model and $\hat{\beta}$ is unbiased.

example: Covariance matrix of $\hat{u}$

recall that $\hat{u} = M_x u$, $E(\hat{u}) = 0$

$$\text{Var}(\hat{u}) = E(\hat{u}\hat{u}^T)$$
$$= M_x E(uu^T) M_x$$
$$= \sigma^2 M_x$$

↕ M_x is idempotent.

Precision of OLS estimates

What determines the precision of OLS estimator $\hat{\beta}$?

1. True variance of the error terms, σ_0^2

   $\text{Var}(\hat{\beta})$ is proportional to σ_0^2. The more random variation there is in the error terms, the more random variation there is in the parameter estimates.

2. Sample size, n

   $\text{Var}(\hat{\beta})$ can be rewritten as $\text{Var}(\hat{\beta}) = \frac{1}{n}\sigma_0^2 \left(\frac{1}{n} X^T X\right)^{-1}$

3. The matrix X.

   Suppose $X = [x_1 \; X_2] \Rightarrow y = x_1 \beta_1 + X_2 \beta_2 + u$.

   By FWL, $M_2 y = M_2 x_1 \hat{\beta}_1 + M_2 u$

$$\hat{\beta}_1 = \frac{x_1^T M_2 y}{x_1^T M_2 x_1}$$

$$= \frac{x_1^T (M_2 x_1 \hat{\beta}_1 + M_2 u)}{x_1^T M_2 x_1}$$

$$= \beta_1 + \frac{x_1^T M_2 u}{x_1^T M_2 x_1}$$

$$\text{Var}(\hat{\beta}_1) = \frac{x_1^T M_2 E(uu^T) M_2 x_1}{(x_1^T M_2 x_1)^2}$$

$$= \sigma_0^2 \cdot \frac{1}{x_1^T M_2 x_1}$$

$$= \frac{\sigma_0^2}{\|M_2 x_1\|^2}$$

$\text{Var}(\hat{\beta})$ is proportional to the inverse of the sum of squared residuals from the regression $x_1 = X_2 c + \text{residuals}$.

Hilroy

When x_1 is well explained by other columns of X (x_1 is collinear to X), the SSR is small and $\text{Var}(\hat{\beta}_1)$ will be large. If x_1 is not well-explained, then SSR is large and $\text{Var}(\hat{\beta}_1)$ is small.

$\text{Var}(\hat{\beta}) = \sigma_0^2 (X^T X)^{-1}$ is rarely known, but can be estimated using an estimate of σ_0^2. $\widehat{\text{Var}}(\hat{\beta}) = s^2 (X^T X)^{-1}$.

Linear functions of parameter estimates

$\text{Var}(\hat{\gamma}) = ?$ where $\gamma = w^T \beta$, $\hat{\gamma} = w^T \hat{\beta}$, w is a k-vector of known coefficients.

$\text{Var}(\hat{\gamma}) = w^T \text{Var}(\hat{\beta}) w = \sigma^2 w^T (X^T X)^{-1} w$

Proof: $\text{Var}(\hat{\gamma}) = \text{Var}(w^T \hat{\beta})$

$= E\left[(w^T \hat{\beta} - E(w^T \hat{\beta}))(w^T \hat{\beta} - E(w^T \hat{\beta}))^T\right]$

$= E\left[(w^T(\hat{\beta} - \beta_0))(\hat{\beta} - \beta_0)^T w\right)$ since $E(\hat{\beta}) = \beta_0$ (unbiased)

$= w^T E(\hat{\beta} - \beta_0)(\hat{\beta} - \beta_0)^T w$

$= w^T \text{Var}(\hat{\beta}) w$

$= w^T (\sigma_0^2 (X^T X)^{-1}) w$

$= \sigma_0^2 w^T (X^T X)^{-1} w$ ∎

3.5 Efficiency of OLS estimator

An estimator is said to be more efficient than another, if, on average, the former yields more accurate estimates than the latter.

Efficiency in terms of precision: estimator $\hat{\beta}$ is said to be more efficient than another estimator $\tilde{\beta}$ iff $\text{Precision}(\hat{\beta}) - \text{Precision}(\tilde{\beta})$ is a nonzero positive semidefinite matrix.

$\text{Precision}(\hat{\beta}) - \text{precision}(\tilde{\beta}) = \text{Var}(\hat{\beta})^{-1} - \text{Var}(\tilde{\beta})^{-1}$

$\text{Var}(\hat{\beta})^{-1} - \text{Var}(\tilde{\beta})^{-1}$ positive semi definite $\equiv \text{Var}(\tilde{\beta}) - \text{Var}(\hat{\beta})$ positive semidefinite.

* This is because if A and B are positive definite matrices of the same dimensions, then $A - B$ is positive semidefinite iff $B^{-1} - A^{-1}$ is positive semi definite.

Thus, the efficiency condition can be stated as $\hat{\beta}$ is more efficient than $\tilde{\beta}$ iff $\text{Var}(\tilde{\beta}) - \text{Var}(\hat{\beta})$ is a nonzero positive semi definite matrix.

The OLS estimator $\hat{\beta} = (X^TX)^{-1}X^Ty$ is "BLUE" (best linear unbiased estimator), according to the Gauss-Markov Theorem.

Gauss-Markov Theorem: If it is assumed that $E(u|X) = 0$ (X exogenous) and $E(uu^T|X) = \sigma^2 I$ (errors are homoskedastic) in the linear regression model $y = X\beta + u$, then the OLS estimator $\hat{\beta}$ is more efficient than any other linear unbiased estimator $\tilde{\beta}$, in the sense that $Var(\tilde{\beta}) - Var(\hat{\beta})$ is a positive semidefinite matrix.

Proof:

Let $\tilde{\beta}$ denote any linear estimator other than the OLS estimator.

$\tilde{\beta} = Ay = (X^TX)^{-1}X^Ty + Cy$, $C \equiv A - (X^TX)^{-1}X^T$

Assuming the model is correctly specified, replace $y = X\beta_0 + u$.

$\tilde{\beta} = A(X\beta_0 + u) = AX\beta_0 + Au$

We want $\tilde{\beta}$ to be an unbiased estimator $\Rightarrow E(\tilde{\beta}) = \beta_0$

$E(AX\beta_0 + Au \mid X) = E(AX\beta_0 \mid X) = \beta_0$ since $E(Au|X) = 0$.

$E(AX\beta_0) = \beta_0$ is true iff $AX = I$.

$AX = I \Rightarrow CX = AX - (X^TX)^{-1}X^TX = I - I = 0$.

The condition that $CX = 0 \Rightarrow Cy = Cu$

$Cy = \tilde{\beta} - \hat{\beta}$, so $\tilde{\beta} - \hat{\beta}$ has conditional mean of 0.

The unbiasedness condition also implies that the covariance matrix of $\tilde{\beta} - \hat{\beta}$ and $\hat{\beta}$ is a zero matrix.

Consequently, $\tilde{\beta} = \hat{\beta} + Cy$, Cy has mean 0 and is uncorrelated with $\hat{\beta}$. This makes it clear that $\hat{\beta}$ is more efficient than $\tilde{\beta}$.

To complete the proof, we note that:

$$\begin{aligned} Var(\tilde{\beta}) &= Var(\hat{\beta} + (\tilde{\beta} - \hat{\beta})) \\ &= Var(\hat{\beta} + Cy) \\ &= Var(\hat{\beta}) + Var(Cy) \quad \text{because } Cov(\hat{\beta}, Cy) = 0. \end{aligned}$$

The difference between $Var(\tilde{\beta})$ and $Var(\hat{\beta})$ is $Var(Cy)$. As it is a covariance matrix, it is positive semidefinite. Thus, $Var(\tilde{\beta}) - Var(\hat{\beta})$ is positive semidefinite. ▨

~~example of unbiased estimator of β_0, $\hat{\beta}_0$, [illegible]~~

~~[illegible]~~

Hilroy

3.6 Residuals and Error Terms

$\hat{u} = y - X\hat{\beta}$ is easily calculated once we have $\hat{\beta}$.
Here, we consider the statistical properties of $\hat{u}$ as an estimator of u.

Consistency of $\hat{\beta}$ implies that $\hat{u} \to u$ as $n \to \infty$. But finite-sample properties of $\hat{u}$ differ from u.

$$y = X\beta + u \longrightarrow M_X y = M_X X \beta_0 + M_X u$$
$$= M_X u \quad \text{as } M_X X = 0$$

$\hat{u} = M_X u$ when the model is correctly specified.

$$\hat{u}_t = M_X u_t = (I - P_X) u_t$$
$$= u_t - X_t (X^T X)^{-1} X^T u_t$$
$$= u_t - \sum_{s=1}^{n} X_t (X^T X)^{-1} X_s^T u_s .$$

Thus, even if each error term u_t is independent of all other error terms, each of the $\hat{u}_t$ will not be independent of all other residuals.

Now assume that $E(u|X) = 0$, ie $E(u_t|X) = 0$ for all t. Since $\hat{u}_t$ is a linear combination of all u_t, it follows that $E(\hat{u}_t|X) = 0$ too.

$$\text{Var}(\hat{u}_t) = E(\hat{u}_t^2) - [E(\hat{u}_t)]^2$$
$$= E(\hat{u}_t^2)$$

The vector of least squares residuals $\hat{u}$ is always smaller than that of the vector of residuals $u(\beta)$. In particular, $\hat{u}$ must be shorter than $u(\beta_0)$, the vector of error terms.

$$\|\hat{u}\|^2 \leq \|u\|^2 \Rightarrow E(\|\hat{u}\|^2) \leq E(\|u\|^2).$$

Assuming that $\text{Var}(u) = \sigma^2$ under the true DGP,

$$\sum_{t=1}^{n} \text{Var}(\hat{u}_t) = \sum_{t=1}^{n} E(\hat{u}_t^2)$$
$$= E\left(\sum_{t=1}^{n} \hat{u}_t^2\right)$$
$$= E(\|\hat{u}\|^2) \leq E(\|u\|^2) = E\left(\sum_{t=1}^{n} u_t^2\right)$$
$$= \sum_{t=1}^{n} E(u_t^2)$$
$$= \sum_{t=1}^{n} \text{Var}(u)$$
$$= n\sigma_0^2$$

Hence, $\text{Var}(\hat{u}_t) < \sigma_0^2$.

Calculating $\text{Var}(\hat{u}_t)$ by calculating $\text{Var}(\hat{u})$:

$$\text{Var}(\hat{u}) = \text{Var}(M_X u)$$
$$= E[M_X u u^T M_X^T) \quad \text{since } E(M_X u) = 0$$
$$= M_X E(uu^T) M_X^T$$
$$= M_X \text{Var}(u) M_X^T$$
$$= \sigma_0^2 M_X \quad \text{as } M_X \text{ is symmetric and idempotent.}$$

The residuals will not have constant variance, and thus variance will always be smaller than σ_0^2.

$$\text{Var}(\hat{u}_t) = E(\hat{u}_t^2)$$
$$= (1 - h_t)\sigma_0^2$$

↖ the t^{th} diagonal element in M_X. $0 < 1 - h_t < 1$

Estimating the variance of error terms

The method of least squares provides estimates of the regression coefficients, but it does not directly provide an estimate of σ^2, the variance of the error terms. The method of moments suggests that we can estimate σ_0 by using the corresponding sample moment.

The sample moment would be $\frac{1}{n}\sum_{t=1}^{n} u_t^2$ if u was observable. Only $\hat{u}_t$ is observable, so the simplest possible MM estimator is:

$$\hat{\sigma}^2 \equiv \frac{1}{n}\sum_{t=1}^{n} \hat{u}_t^2, \text{ the average of } n \text{ squared residuals}$$

Bias of $\hat{\sigma}^2$: $E(\hat{\sigma}^2) = \frac{1}{n}\sum_{t=1}^{n} E(\hat{u}_t^2)$

$$= \frac{1}{n}\sum_{t=1}^{n} (1 - h_t)\sigma_0^2$$
$$= \frac{\sigma_0^2}{n}\sum_{t=1}^{n} 1 - h_t$$
$$= \frac{\sigma_0^2}{n}(n - k) \quad \text{since } \sum_{t=1}^{n}(1 - h_t) = \text{tr}(M_X)$$
$$= \frac{n-k}{n}\sigma_0^2$$

$$E(\hat{\sigma}_0^2) - \sigma_0^2 = \frac{n-k}{n}\sigma_0^2 - \sigma_0^2$$
$$= -\frac{k}{n}\sigma_0^2$$

The bias is negative, $\hat{\sigma}^2$ is biased downwards

Consistency of $\hat{\sigma}^2$: $E(\hat{\sigma}^2) = \frac{n-k}{n}\sigma_0^2 \to \sigma_0^2$ as $n \to \infty$, $\hat{\sigma}^2$ is consistent

Another MM estimator that is unbiased is:

$$s^2 \equiv \frac{1}{n-k} \sum_{t=1}^{n} \hat{u}_t^2$$

Proof: $E(s^2) = \frac{n}{n-k} E(\hat{\sigma}_u^2)$

$= \frac{n}{n-k} \left(\frac{n-k}{n} \sigma_0^2 \right)$

$= \sigma_0^2$.

$\therefore E(s^2) - \sigma_0^2 = 0$.

3.7 Misspecification of Linear Regression Models

Up to this point, we have assumed that the model is correctly specified, ie the DGP $y = X\beta_0 + u$ belongs to the model $y = X\beta + u$.
What are the statistical properties of $\hat{\beta}$ when the model is not correctly specified? We consider the case of underspecification, when the true DGP is not contained in the model.

In order to understand underspecification better, we begin by discussing its opposite: overspecification.
A model is said to be overspecified if some variables that rightly belong to the information set, but do not appear in the DGP, are mistakenly included in the model. Note that overspecification is not technically a misspecification!

Suppose we estimated the model $y = X\beta + Z\gamma + u$, where the data is actually generated by $y = X\beta_0 + u$.
We can obtain the correct specification of the model with $\beta = \beta_0$ and $\gamma = 0$.
Applying FWL, $M_Z y = M_Z X \beta + \text{residuals}$, where $M_Z = I - Z(Z^T Z)^{-1} Z^T$

$\therefore \tilde{\beta} = (X^T M_Z X)^{-1} X^T M_Z y$

Replacing y by $X\beta_0 + u$,

$\tilde{\beta} = (X^T M_Z X)^{-1} X^T M_Z (X\beta_0 + u)$

$= \beta_0 + (X^T M_Z X)^{-1} X^T M_Z u$

$E(\tilde{\beta}) = E(\beta_0 + (X^T M_Z X)^{-1} X^T M_Z u)$ cond on Z and X

$= \beta_0$ ($M_Z u = u$, $E(u) = 0$)

Hence $\tilde{\beta}$ is unbiased.

$\text{Var}(\tilde{\beta}) = \sigma_0^2 (X^T M_Z X)^{-1}$

$\text{Var}(\tilde{\beta}) - \text{Var}(\hat{\beta})$ is a positive semi-definite matrix; $\tilde{\beta}$ is at most as efficient as $\hat{\beta}$ (when the model is restricted to $y = X\beta + u$).

Hence, overspecification leads to less accurate results — it increases the variance of the estimates of the coefficients on the regressors that belong to the DGP, and the increase can be very great in many cases.

Underspecification

A model where some variables that actually do appear in the DGP are omitted is an underspecified model.

Suppose we have a model $y = X\beta + u$, but the DGP is actually $y = X\beta_0 + Z\gamma_0 + u$.

$\hat{\beta}$ is biased. $E(\hat{\beta}) = E\left((X^TX)^{-1}X^Ty\right)$

$= E\left((X^TX)^{-1}X^T(X\beta_0 + Z\gamma_0 + u)\right)$

$= \beta_0 + (X^TX)^{-1}X^TZ\gamma_0 + \underbrace{E\left((X^TX)^{-1}X^Tu\right)}_{=0}$

$= \beta_0 + \underbrace{(X^TX)^{-1}X^TZ\gamma_0}_{=0}$ if X and Z are orthogonal, ie $X^TZ = 0$, or $\gamma_0 = 0$. (model not underspecified)

Therefore, $\hat{\beta}$ is generally biased. The magnitude of the bias depends on the parameter vector γ_0 and X and Z matrices.
As the bias does not vanish as $n \to \infty$, $\hat{\beta}$ is inconsistent.

In conclusion, underspecification leads to biased estimates and an overestimated covariance matrix that may be misleading, while overspecification leads to a loss in efficiency. Therefore, it would seem that we should err on the side of overspecification. This makes sense in sufficiently large samples; however, in samples of modest size, the gain in efficiency from omitting some variables may be very large relative to the bias caused in their omission.

Hilroy