Econometrics notes round 1.

Chapter 3: the statistical property of OLS

Compared to the numeric property of OLS (e.g., it minimizes the MSE), the statistical property of OLS (e.g., unbiasedness, efficiency, consistency etc.) can be hold only when the assumption made is true.

We denote the Data Generating Process (**DGP**) as the mechanism that dominates the reality (e.g., to generate the data). In econometrics, we develop models to describe the DGP or at least to partially describe it. Therefore, we may assume that our regression model is correctly specified, i.e., it embeds the DGP (in practice, however, the real DGP for the economy can be very sophisticated and usually we do not pretend to know every detail of it – that is why 'all the models are wrong, though some are useful'). For example, suppose that we assume that the **DGP** is:

$$y_t = X_t\beta_0 + u_t,\ u_t \sim NID(0,\ \sigma_0^2) \tag{3.02}$$

where X_t represent a column vector for the t^{th} observation. we can develop the following **model**:

$$y_t = X_t\beta + u_t,\ u_t \sim IID(0, \sigma^2) \tag{3.01}$$

(3.02) is embedded in (3.01) as (3.02) is simply (3.01) by taking specific values for the parameters (e.g., β, σ_0) and restricting the distribution which the error term follows to be the Normal distribution.

We assume that our model is correctly specified – only then we can talk about the statistical properties of our models. The challenge is that in practice the DGP may not be generated by (3.02), and the property of the OLS estimator for the model (3.01) becomes questionable.

3.2 Are OLS estimators unbiased?

Following the example which we show in (3.01) and (3.02), we can rewrite our model in (3.01) in a matrix notation as:

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

Or

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \beta + \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}, u_t \sim IID(0, \sigma^2)$$

The OLS estimator can be written as:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

Suppose our model is correctly specified and the true GDP is indeed generated by (3.02). We can insert (3.02) into the OLS estimator, and we have:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta_0 + \boldsymbol{u})$$
$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta_0} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}$$

If we take expectations on both sides, we have:

$$E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}_0 + E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}) \tag{3.06}$$

Thus, the OLS estimator is unbiased when the second component of (3.06) equals to zero.

Sometimes it is reasonable to consider $\boldsymbol{X}$ as fixed or non-stochastic. For example, in an experiment (though this normally happens in the statistical, not econometrical, contexts), we can control the values of $\boldsymbol{X}$. Thus, we have:

$$E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta_0} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{u}) = \boldsymbol{\beta_0}$$

In this case, the OLS estimator is unbiased. However, having fixed values of $\boldsymbol{X}$ is very rarely the case in practice. A weaker assumption is **exogeneity**, which suggests that any randomness in the DGP which generate $\boldsymbol{X}$ is **(mean) independent** of the error term $\boldsymbol{u}$, or alternatively speaking, the expected values of all the error terms are mean independent to the all the independent variables, e.g.,

$$E(\boldsymbol{u}|\boldsymbol{X}) = \boldsymbol{0}$$

We can write the equation in a vector format:

$$E\left(\begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \middle| \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}\right) = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} = \boldsymbol{0}$$

where u_1 and X_1 are the error term and the column vector for the explanatory variables for the 1st observation. The equation suggests that the mean of the entire vector u, e.g., $\begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}$, or alternatively speaking, the mean of every single element of the $\boldsymbol{u}$ (e.g., $u_1, u_2, ..., u_n$) is zero conditional on the entire matrix of $\boldsymbol{X}$, e.g., $\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$. Since by definition $E(\boldsymbol{u}|\boldsymbol{X})$ is a vector of variables (e.g., $E(u_1|\boldsymbol{X})$, $E(u_2|\boldsymbol{X})$, ..., $E(u_n|\boldsymbol{X})$) which depend on the value of $\boldsymbol{X}$ (e.g., all the elements including $X_1, X_2, \ldots X_n$), the equation $E(\boldsymbol{u}|\boldsymbol{X}) = 0$ suggests that the mean of <u>every</u> single element of the $\boldsymbol{u}$ vector would be zero no matter what values any element of the matrix $\boldsymbol{X}$ (e.g., and thus, each of the $X_1, X_2, \ldots X_n$) could possibly take. Or alternatively speaking, the vector variable $\boldsymbol{u}$ (e.g., $\begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}$) is mean independent of the vector variable $\boldsymbol{X}$ (e.g., $\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$).

If the assumption of exogeneity holds, by using the Law of Iterated Expectations, it can be proven that:

$$E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u})) = \boldsymbol{0}$$

Thus, we have:

$$E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}_0 + E\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\right) = \boldsymbol{\beta}_0$$

Thus, if we assume exogeneity, we can conclude that the OLS estimator is unbiased.

Usually, it is reasonable to assume exogeneity for cross-sectional data. For example, each observation represents an individual firm, person, or city. We may usually consider the observations to be collected from random samples from the population. Thus, we can assume that the error term for any of the observations/individuals tends to have a mean value (of zero) irrelevant to the values of the independent variable for **all** the observations/individuals (e.g., $X_t, X_2, \ldots X_n$). e.g.,

$$E\left(\begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \middle| \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}\right) = E(\boldsymbol{u}|\boldsymbol{X}) = E(\boldsymbol{u}) = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} = \boldsymbol{0}$$

For example, we may model a person's salary based on his/her working experience. The error term contains other influencing factors to the person's salary but not included in the model, such as the person's intelligence. It may be reasonable to assume that the expected impact by one person's intelligence on the person's salary is mean independent of everyone' working experience (and we assume the impact by the person's intelligence conditional on everyone'

working experience equals to zero, as we assume that the expected impact by one person's intelligence on the person's salary is zero). Cross-sectional data are heavily involved in applied microeconomics and other social sciences.

The assumption of exogeneity does not hold for time series data when there are lagged dependent variables used as independent variables. See next.

The OLS estimator can be biased

The OLS estimator can be biased when there are lagged dependent variables in the model (e.g., we call this type of model as '**autoregressive'** models), e.g.,

$$y_t = \beta y_{t-1} + u_t, u_t \sim IID(0, \sigma^2)$$

This equation can be written in a vector/matrix form:

$$\boldsymbol{y} = \boldsymbol{y_1}\boldsymbol{\beta} + \boldsymbol{u}, \boldsymbol{u} \sim \boldsymbol{IID}(\boldsymbol{0}, \boldsymbol{\sigma^2 I}) \tag{3.11}$$

(3.11) can be written equivalently as (again, suppose we have a sample of size *n*):

$$\begin{bmatrix} y_2 \\ \vdots \\ y_n \\ y_{n+1} \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} u_2 \\ \vdots \\ u_n \\ u_{n+1} \end{bmatrix}$$

Thus, we can assume predeterminedness for (3.11): that is, $E(u_t|x_t) = E(u_t) = 0$, in this case, we have the lagged dependent variable as our independent variable (e.g., y_{t-1}), thus we have the predeterminedness assumption as: $E(u_t|y_{t-1}) = E(u_t) = 0$, where $t = 1, \dots, n+1$, for a sample of size *n*. This assumption could be reasonable because we always realize the value of y_{t-1} before u_t, and we can assume that u_t is mean independent of y_{t-1}, or alternatively speaking we assume that y_{t-1} has no impact on the expected value of u_t.

However, in this case, we cannot assume exogeneity. The assumption of exogeneity is $E(\boldsymbol{u}|\boldsymbol{X}) = 0$. In this case, we have the lagged dependent variable as our independent variable, thus the assumption of exogeneity is $E(\boldsymbol{u}|\boldsymbol{y_1}) = E(\boldsymbol{u})$, which requires that $\boldsymbol{y_1}$ has no impact on the expected value of any element of the vector $\boldsymbol{u}$. This can be written as:

$$E(\boldsymbol{u}|\boldsymbol{y_1}) = E\left(\begin{bmatrix} u_2 \\ \vdots \\ u_n \\ u_{n+1} \end{bmatrix} \middle| \begin{bmatrix} y_1 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix}\right) = E\left(\begin{bmatrix} u_2 \\ \vdots \\ u_n \\ u_{n+1} \end{bmatrix}\right)$$

We already assumed predeterminedness, e.g., $E(u_t|y_{t-1}) = E(u_t) = 0$, thus we have $E(u_t y_{t-1}) = 0$. However, we have:

$$\begin{aligned} E(u_t y_t) = E(u_t(\beta y_{t-1} + u_t)) &= \beta E(u_t y_{t-1} + u_t u_t)) \\ &= \beta E(u_t y_{t-1}) + E(u_t u_t) \\ &= E(u_t u_t) = E(u_t^2) \end{aligned}$$

Thus, unless the error term is always zero, $E(u_t y_t) \neq 0$. However, y_t is the regressor for the error term u_{t+1}. Thus, the exogeneity of $E(\boldsymbol{u}|\boldsymbol{y_1}) = E(\boldsymbol{u})$ cannot hold. Intuitively, this is because y_t depends on y_{t-1}, and thus u_{t-1} (according to the model itself by definition, $y_t = \beta_1 + \beta_2 y_{t-1} + u_t$). Thus, we cannot assume that the error terms are mean independent of the regressor.

Thus, if we have the lagged dependent variables as explanatory variables, the second component corresponding to (3.06) do not equal to zero, and the OLS estimator is biased. It can be proved that, not only the parameter for the lagged dependent variable (as the independent variable) would be biased but all the parameters in the autoregressive models from an OLS estimator will be biased.

3.3 Are OLS estimator consistent?

Consistency means that when the sample size goes to infinity, the estimate tends to be the quantity being estimated. That is, when the sample size is large enough, we will be confident that the estimate will be close to the true value. The OLS estimator will be often consistent even when being biased.

How should we think of the sample size of infinity? For cross-sectional data we can pretend that we draw the sample from a population of infinity size (e.g., unlimited number of firms or persons). We simply draw more and more observations from the population. For time series data we can pretend that the time goes back to unlimited history and goes forward to the unlimited future so that we have unlimited observations, and we simply try to observe for longer and longer periods of time.

Probability limit

The probability limit, or ***plim***, of the function of a vector y^n is defined as:

$$\operatorname*{plim}_{n\to\infty} \alpha(y^n) = \alpha_0$$

y^n is usually written as y, omitting the vector size *n*.

Thus when $n \to \infty$, the vector $\alpha(y^n)$ approaches the vector of α_0. α_0 could be a vector of random or non-random variables.

Under the definition of ***plim,*** $\alpha(y^n)$ and α_0 have the following relationship, for $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr(\| \ (\alpha(y^n) - \alpha_0 \ \| < \varepsilon) = 1$$

Thus, for a specific tolerance level ε, no matter how small, when $n \to \infty$, the probability that the norm of the discrepancy (e.g., Euclidean distance) between the two vectors $\alpha(y^n)$ and α_0 is less than ε goes to unity.

Some vectors originally do not have ***plim***, but they may have ***plim*** when they are divided by some quantity (e.g., *n* or the power of *n*), e.g., $\boldsymbol{X'X}$ does not have a plim, but we can divide $\boldsymbol{X'X}$ by *n* and we have:

$$\operatorname*{plim}_{n\to\infty} \frac{\boldsymbol{X'X}}{n} = \boldsymbol{S_{X'X}} \tag{3.17}$$

Each element in $\boldsymbol{X'X}$ is a scalar product between two columns in $\boldsymbol{X}$.

Thus, we have $\boldsymbol{S_{X'X}}$ as a finite non-random matrix with a full rank of *k*. This is because each element in $\frac{\boldsymbol{X'X}}{\boldsymbol{n}}$ is an average of *n* numbers:

$$\left(\frac{\boldsymbol{X'X}}{n}\right)_{ij} = \frac{1}{n}\sum_{t=1}^{n} x_{ti}x_{tj}$$

we have made implicit assumptions that multiple variables in $\boldsymbol{X}$ do not have too strong dependence (no perfect multicollinearity).

The OLS estimator is consistent

When the model is correctly specified, we have:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}$$

The OLS estimator is consistent when $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}$ has a plim of zero. e.g., we know that:

$$\operatorname*{plim}_{n\to\infty}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}) = \frac{\operatorname*{plim}_{n\to\infty}\boldsymbol{X}'\boldsymbol{u}}{\operatorname*{plim}_{n\to\infty}\boldsymbol{X}'\boldsymbol{X}} = \frac{\operatorname*{plim}_{n\to\infty}\frac{\boldsymbol{X}'\boldsymbol{u}}{n}}{\operatorname*{plim}_{n\to\infty}\frac{\boldsymbol{X}'\boldsymbol{X}}{n}} = \frac{\operatorname*{plim}_{n\to\infty}\frac{\boldsymbol{X}'\boldsymbol{u}}{n}}{\boldsymbol{S}_{X'X}} = \frac{\operatorname*{plim}_{n\to\infty}\frac{\sum_{t=1}^{n}X_t'u_t}{n}}{\boldsymbol{S}_{X'X}}$$

As we assume predeterminedness, e.g.,

$$E(u_t|X_t) = 0$$

According to LLN, we know that $E(X_t u_t) = 0$, thus the previous equation becomes:

$$\operatorname*{plim}_{n\to\infty}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}) = \frac{\operatorname*{plim}_{n\to\infty}\frac{\sum_{t=1}^{n}X_t'u_t}{n}}{\boldsymbol{S}_{X'X}} = \frac{0}{\boldsymbol{S}_{X'X}} = 0$$

Thus, the OLS estimator is consistent, although it may not be unbiased (e.g., when we have lagged dependent variable as the explanatory variable). In this kind, when we have large sample, we can trust the OLS estimate to be very close to the true value of the parameters.

Unbiasedness and consistency are two different things. Sometimes the estimator is biased but consistent, and sometimes it is unbiased but inconsistent (e.g., either the estimator does not tend to any non-random probability limit or tend to a wrong probability limit). See examples in Davison and Mackinnon (2003).

The covariance matrix of the OLS estimator

Until now, we know that the OLS estimator of the parameters to be unbiased and consistent (with the assumption of exogeneity) or biased and consistent (only with predeterminedness), but that is not enough. We would like to know the distribution of the OLS estimate (e.g., $\widehat{\boldsymbol{\beta}}$). We are usually interested in the central second moment of the distribution of $\widehat{\boldsymbol{\beta}}$. e.g.,

The OLS covariance matrix

Previously we have introduced the variance-covariance of the error term, e.g.,

$$Var(\boldsymbol{u}) = E(\boldsymbol{u}\boldsymbol{u}') = \sigma^2 \boldsymbol{I} \tag{3.26}$$

Note that, this only requires the covariance of the error terms to be zero, this does not require the error terms are independent (e.g., $\boldsymbol{IID}$ is a stronger assumption).

Thus, we would like to calculate $Var(\widehat{\boldsymbol{\beta}})$, as we know that:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta_0} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}$$

Thus,

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}$$

As we assume the OLS estimator is unbiased, based on (3.22), we have the variance matrix of $\widehat{\boldsymbol{\beta}}$ as follows:

$$\begin{aligned} Var(\widehat{\boldsymbol{\beta}}) &= E\left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})'\right) = E(\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}') = E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\right)') \\ &= E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\boldsymbol{u}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) \\ &= E(\boldsymbol{u}\boldsymbol{u}')E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) \\ &= \sigma_0^2 E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) \\ &= \sigma_0^2 E\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\right) \end{aligned} \tag{3.28p}$$

where $Var(\widehat{\boldsymbol{\beta}})$ is random variable depending on $\boldsymbol{X}$.

We usually use $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ as an estimate of $E((\boldsymbol{X}'\boldsymbol{X})^{-1})$, and write (3.28p) as:

$$Var(\widehat{\boldsymbol{\beta}}) = \sigma_0^2 E((\boldsymbol{X}'\boldsymbol{X})^{-1}) = \sigma_0^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{3.28}$$

However, note that $E((\boldsymbol{X}'\boldsymbol{X})^{-1})$ is the expected value of a complicated nonlinear function of r.v, and thus it does not equal to the nonlinear function of the expected value. That means, strictly speaking, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is a biased estimate of $E((\boldsymbol{X}'\boldsymbol{X})^{-1})$.

Econometricians thus take alternative approaches to justify the equation in (3.28). First, they may argue that $(\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}}$ is a consistent estimate of $\sigma_0^2 E((\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}})$. Thus, $Var(\hat{\beta}) = \sigma_0^2 E\left((\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}}\right) = \sigma_0^2 (\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}}$ asymptotically.

Or they may assume that $\boldsymbol{X}$ is fixed (e.g., stronger than independence between the regressor and the error terms). This also justify (3.28), though this is a very strong assumption. Thus, strictly speaking, when we have small samples, estimation of the variance (3.28) is biased downward because we ignore the variability coming from the change in the explanatory variable observations over repeated samples – but normally we take it as a compromise.

The precision of the LS estimates

For a scalar parameter, the accuracy of an estimator is often taken to be proportional to the inverse of its variance, which is called the **precision** of the parameter. The **precision matrix** is defined as the inverse of the covariance matrix of the estimator. (3.28) can be rewritten as:

$$Var(\widehat{\boldsymbol{\beta}}) = \frac{\sigma_0^2}{n}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-\mathbf{1}}$$

Thus, we find that $Var(\widehat{\boldsymbol{\beta}})$ is proportional to three things.

First, $Var(\widehat{\boldsymbol{\beta}})$ is proportional to the error variance σ_0^2. The more random variation in the error term, the more random variation in the parameter estimate. Also, as (3.17) suggests that, when the sample size *n* is reasonably large, we have:

$$\operatorname{plim}_{n\to\infty} \frac{\boldsymbol{X}'\boldsymbol{X}}{n} = \boldsymbol{S}_{\boldsymbol{X}'\boldsymbol{X}}$$

Thus $\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-\mathbf{1}}$ tends to $\boldsymbol{S}_{\boldsymbol{X}'\boldsymbol{X}}$ which is a non-random vector variable and does not vary with the sample size *n*. Thus, we find that $Var(\widehat{\boldsymbol{\beta}})$ also tends to be proportional to $\frac{1}{n}$. Thus, if we double the sample size, we expect the variance of the estimate to be halved.

$Var(\widehat{\boldsymbol{\beta}})$ also depends on the matrix $\boldsymbol{X}$. When the explanatory variables contain a lot of information, $\boldsymbol{S}_{\boldsymbol{X}'\boldsymbol{X}}$ tends to be large (e.g., this can be intuitively and analytically shown if we look at an example where we only have one regressor, i.e., $\boldsymbol{S}_{\boldsymbol{X}'\boldsymbol{X}}$ is directly related to the variance of the regressor). Some elements in the $Var(\widehat{\boldsymbol{\beta}})$ matrix tend to be large because some explanatory variables contain overlapped information – which we call multicollinearity (see more details in Davison and Mackinnon, 2003).

We can use $Var(\widehat{\boldsymbol{\beta}})$ to make inference about the OLS estimator $\widehat{\boldsymbol{\beta}}$.

Linear function of parameter estimates

Once we know that covariance matrix of the OLS estimator, e.g., $Var(\widehat{\boldsymbol{\beta}})$, we can then know the variance of a linear combination of the OLS estimator, e.g.,

$$\begin{aligned} Var(\boldsymbol{\omega}'\widehat{\boldsymbol{\beta}}) &= E\left(\boldsymbol{\omega}'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})'\boldsymbol{\omega}|\boldsymbol{X}\right) \\ &= \boldsymbol{\omega}' E\left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})'|\boldsymbol{X}\right)\boldsymbol{\omega} \\ &= \boldsymbol{\omega}'\sigma_0^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\omega} \\ &= \sigma_0^2\boldsymbol{\omega}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\omega} \end{aligned} \tag{3.33}$$

We can use this finding to derive the variance of the forecast error. We denote that y_s and X_s represent the row vector for the dependent variable and the matrix for the explanatory variables for $t = s$ where s is beyond the in-sample data for $t = 1\ to\ n$. We assume that the model is correctly specified, suppose $\hat{y}_s$ is the forecast, therefore we have the forecast error variance as follows:

$$\begin{aligned} E(y_s - \hat{y}_s)^2 &= E\left(y_s - X_s\widehat{\boldsymbol{\beta}}\right)^2 \\ &= E\left(X_s\boldsymbol{\beta_0} + u_s - X_s\widehat{\boldsymbol{\beta}}\right)^2 \\ &= E\left(u_s + X_s\boldsymbol{\beta_0} - X_s\widehat{\boldsymbol{\beta}}\right)^2 \\ &= E\left(u_s^2 + (X_s\boldsymbol{\beta_0} - X_s\widehat{\boldsymbol{\beta}})^2 + 2u_s(X_s\boldsymbol{\beta_0} - X_s\widehat{\boldsymbol{\beta}})\right)^2 \\ &= E\left(u_s^2\right) + E\left(u_s + X_s\boldsymbol{\beta_0} - X_s\widehat{\boldsymbol{\beta}}\right)^2 + 2E(u_s\left(X_s\boldsymbol{\beta_0} - X_s\widehat{\boldsymbol{\beta}}\right)) \end{aligned}$$

If we can assume exogeneity (e.g., no serial correlation), we have:

$$u_s X_s = 0$$

Thus, the previous equation becomes:

$$E\left(y_s - X_s\widehat{\boldsymbol{\beta}}\right)^2 = E(u_s^2) + E\left(X_s\boldsymbol{\beta_0} - X_s\widehat{\boldsymbol{\beta}}\right)^2$$

If we assume $\hat{\beta}$ as unbiased, we have:

$$E\left(y_s - X_s\widehat{\boldsymbol{\beta}}\right)^2 = \sigma_0^2 + Var(\boldsymbol{X}_s\widehat{\boldsymbol{\beta}})$$

This suggests that the forecasting error variance, when we assume the model is correctly specified and exogeneity to be hold, is the error variance σ_0^2 plus a penalty of $Var(X_s\widehat{\boldsymbol{\beta}})$ due to

the fact we do not use the true parameter $\boldsymbol{\beta_0}$ but the estimate, e.g., $\widehat{\boldsymbol{\beta}}$ to generate the forecast.

Note that, since $\hat{y}_s$ is a linear combination of X_s, and we assumed that X_s is uncorrelated with u_s. We know that $\hat{y}_s$ is uncorrelated with u_s.

Efficiency of the OLS estimator

One estimator is said to be more efficient than another when its estimates are more accurate. That is, it uses the information available more efficiently. e.g.,

1. For scalar parameters, one estimator is more efficient than another if its precision is larger than that of the latter.

2. For a vector parameter, e.g., $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ are both unbiased estimators of the vector parameter $\boldsymbol{\beta}$. Thus, $\widehat{\boldsymbol{\beta}}$ is more efficient than $\tilde{\beta}$ if and only if $Var(\widetilde{\boldsymbol{\beta}}) - Var(\widehat{\boldsymbol{\beta}})$ is a nonzero positive semidefinite matrix.

The fact that $\widehat{\boldsymbol{\beta}}$ is more efficient than $\widetilde{\boldsymbol{\beta}}$ also suggests that every element in $\boldsymbol{\beta}$ or any combination of those parameters, is estimated at least as efficiently by using $\widehat{\boldsymbol{\beta}}$ as using $\widetilde{\boldsymbol{\beta}}$.

Proof: suppose $\boldsymbol{\gamma}$ is a linear combination of $\boldsymbol{\beta}$, e.g., $\boldsymbol{\gamma} = \boldsymbol{\omega}'\boldsymbol{\beta}$. Here in a special case, $\boldsymbol{\gamma}$ could be a particular parameter in $\boldsymbol{\beta}$ if the elements in $\boldsymbol{\omega}$ are all zero except for that parameter in $\boldsymbol{\beta}$. As we see in (3.33) that,

$$Var(\boldsymbol{\omega}'\widehat{\boldsymbol{\beta}}) = \boldsymbol{\omega}'(\sigma_0^2\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\omega} = \boldsymbol{\omega}'Var(\widehat{\boldsymbol{\beta}})\boldsymbol{\omega}$$

Accordingly,

$$Var(\boldsymbol{\omega}'\widetilde{\boldsymbol{\beta}}) = \boldsymbol{\omega}'Var(\widetilde{\boldsymbol{\beta}})\boldsymbol{\omega}$$

Thus, if we take the difference, we have:

$$Var(\boldsymbol{\omega}'\widetilde{\boldsymbol{\beta}}) - Var(\boldsymbol{\omega}'\widehat{\boldsymbol{\beta}}) = \boldsymbol{\omega}'(Var(\widetilde{\boldsymbol{\beta}}) - Var(\widehat{\boldsymbol{\beta}}))\boldsymbol{\omega} \tag{3.36}$$

Therefore, if $\widehat{\boldsymbol{\beta}}$ is more efficient than $\widetilde{\boldsymbol{\beta}}$, the RHS of (3.36) will be non-negative, which suggests that the LHS of (3.36) must be non-negative. This suggests that $Var(\boldsymbol{\omega}'\widetilde{\boldsymbol{\beta}}) - Var(\boldsymbol{\omega}'\widehat{\boldsymbol{\beta}}) \geq 0$.

Thus, any combination of parameters in $\boldsymbol{\beta}$ is estimated at least as efficiently by using $\widehat{\boldsymbol{\beta}}$ as using $\widetilde{\boldsymbol{\beta}}$.

The OLS estimator is more efficient than other linear unbiased estimators

We say an estimator to be a 'linear' estimator if we can write the estimator as a linear function of the vector of observations $\boldsymbol{y}$ for the dependent variable. e.g., We know that the OLS estimator is:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

We can denote that $\tilde{\beta}$ to be a linear estimator but is NOT the OLS estimator, e.g.,

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{C}\boldsymbol{y}$$

where $\boldsymbol{A}$ and $\boldsymbol{C}$ are *k* x *n* matrices that depend on $\boldsymbol{X}$. Based on the equation above, we have:

$$\boldsymbol{A}\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{C}\boldsymbol{y}$$
$$\boldsymbol{C}\boldsymbol{y} = \boldsymbol{A}\boldsymbol{y} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$
$$\boldsymbol{C} = \boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

The **Gauss-Markov Theorem**:

If we can assume that $E(\boldsymbol{u}|\boldsymbol{X}) = 0$ (exogeneity) and $E(\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{X}) = \sigma^2\boldsymbol{I}$ (homoskedasticity) in the linear regression model, then $\hat{\beta}$ is **BLUE** (best linear unbiased estimator). That is, the OLS estimator is more efficient than any other linear unbiased estimator. Note that the OLS estimator may not be more efficient compared to other biased estimators or nonlinear estimators and it only has the BLUE property when the assumptions (e.g., exogeneity and homoskedasticity) are met.

We can prove that if:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

And

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{C}\boldsymbol{y}$$

If we assume $\tilde{\beta}$ is unbiased, we will have a strong constraint on $\boldsymbol{C}$, which makes $Cov(\hat{\beta}, \boldsymbol{C}\boldsymbol{y}) = 0$, (see more details in Davison and Mackinnon, 2003). Thus, we have:

$$Var(\tilde{\beta}) = Var(\hat{\beta} + (\tilde{\beta} - \hat{\beta}))$$
$$= Var(\hat{\beta} + \boldsymbol{Cy})$$
$$= Var(\hat{\beta}) + Var(\boldsymbol{Cy}) - Cov(\hat{\beta}, \boldsymbol{Cy})$$
$$= Var(\hat{\beta}) + Var(\boldsymbol{Cy})$$

Therefore, we have:

$$Var(\tilde{\beta}) - Var(\hat{\beta}) = Var(\boldsymbol{Cy})$$

Since $Var(\boldsymbol{Cy})$ is a covariance matrix, $Var(\tilde{\beta}) - Var(\hat{\beta})$ must be positive semidefinite. This proves the **Gauss-Markov Theorem.**

Residual and the error term

Once we have the parameter estimate $\widehat{\boldsymbol{\beta}}$, we can calculate the estimate of error term, e.g., $\widehat{\boldsymbol{u}}$:

$$\widehat{\boldsymbol{u}} \equiv \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

According to the numerical property of OLS, $\widehat{\boldsymbol{u}}$ is orthogonal to $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ and every vector in the $S(\boldsymbol{X})$. We need $\widehat{\boldsymbol{u}}$ to estimate the error variance σ^2 so that we can calculate the variance matrix of $\widehat{\boldsymbol{\beta}}$:

$$Var(\widehat{\boldsymbol{\beta}}) = \frac{\sigma_0^2}{n}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

We know that $\widehat{\boldsymbol{\beta}}$ is consistent, thus, $\widehat{\boldsymbol{u}} \to \boldsymbol{u}$ when $n \to \infty$. however, the finite property of $\widehat{\boldsymbol{u}}$ differs from those of $\boldsymbol{u}$.

If the model is correctly specified, we have:

$$\widehat{\boldsymbol{u}} = \boldsymbol{M_X y} = \boldsymbol{M_X X \beta_0} + \boldsymbol{M_X u}$$
$$\widehat{\boldsymbol{u}} = \boldsymbol{M_X y} = \boldsymbol{M_X u}$$
$$\widehat{\boldsymbol{u}} = \boldsymbol{M_X u}$$

Each residual in $\widehat{\boldsymbol{u}}$ is a linear combination of every element in $\boldsymbol{u}$. Consider a row of the matrix product of $\widehat{\boldsymbol{u}} = \boldsymbol{M_X u}$,

$$\hat{u}_t = u_t - \boldsymbol{X_t}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}$$
$$\hat{u}_t = u_t - \sum_{s=1}^{n} \boldsymbol{X_t}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_s'\boldsymbol{u_s}$$

(3.42)

Thus, even we may assume that u_t is independent of all the other error terms (e.g., u_s), $\hat{u}_t$ is not. There is some dependency between each pair of $\hat{u}_t$ (e.g., in (3.42), all the $\hat{u}_t$ depend on $\sum_{s=1}^{n} \boldsymbol{X_t}(\boldsymbol{X'X})^{-1}\boldsymbol{X'_s u_s}$)) though this dependency diminishes as sample size *n* increase.

If we assume exogeneity, e.g., $E(\boldsymbol{u}|\boldsymbol{X}) = 0$. This assumption suggests that $E(u_t|\boldsymbol{X}) = 0$. Also, we know from (3.42) that $\hat{u}_t$ is a linear combination of all the u_t. Therefore, this means the $E(\hat{u}_t|\boldsymbol{X}) = 0$. In this respect, $\hat{u}_t$(or $\widehat{\boldsymbol{u}}$) just behaves like u_t(or $\boldsymbol{u}$).

In other respects, $\hat{u}_t$(or $\widehat{\boldsymbol{u}}$) does not behave like u_t(or $\boldsymbol{u}$). e.g., the Euclidean length of the vector of the LS residuals $\widehat{\boldsymbol{u}}$ is always smaller than that of the vector of the residuals evaluated at any other value. This means, $\widehat{\boldsymbol{u}}$ must be shorter than the vector of error terms $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{\beta_0})$

We can calculate the variance of $\widehat{\boldsymbol{u}}$:

$$\begin{gathered} Var(\widehat{\boldsymbol{u}}) = Var(\boldsymbol{M_X u}) = \boldsymbol{E}(\boldsymbol{M_X u u' M_X}) \\ = \boldsymbol{E}\left(\boldsymbol{M_X u u'^{M_X}}\right) = \boldsymbol{M_X E}(\boldsymbol{u u'})\boldsymbol{M_X} = \boldsymbol{M_X} Var(\boldsymbol{u})\boldsymbol{M_X} \\ = \boldsymbol{M_X}(\sigma^2 \boldsymbol{I})\boldsymbol{M_X} = \sigma^2 \boldsymbol{M_X M_X} = \sigma^2 \boldsymbol{M_X} \end{gathered}$$

(3.43)

The second equation holds because the expected value of $\boldsymbol{M_X u}$ is zero. (3.43) suggests that in general the variance-covariance of $\widehat{\boldsymbol{u}}$ is not a diagonal matrix. This suggests that the off-diagonal elements in the matrix are not zero, e.g., $E(\hat{u}_t \hat{s}_t) \neq 0$. Thus, even we assume the original errors are uncorrelated, the residuals are not uncorrelated.

(3.43) also suggests that the residuals do not have a constant variance, and the residual variance for every $\hat{u}_t$ must be smaller than σ_0^2 (see details in Davison and Mackinnon, 2003).

Estimating the variance of the error terms

The method of LS estimate β but not σ^2. We can estimate σ^2 using MM based on the moment of the sample.

Ideally, we have:

$$\hat{\sigma}^2 = Var(u_t) = \frac{1}{n}\sum_{t=1}^{n} u_t^2$$

However, we do not observe u_t but only $\hat{u}_t$, thus we have:

$$\hat{\sigma}^2 = Var(\hat{u}_t) = \frac{1}{n}\sum_{t=1}^{n} \hat{u}_t^2$$

(3.46)

It can be shown that (3.46) is consistent. However, because each $\hat{u}_t^2$ is smaller than σ^2, (3.46) is biased downward. It can be shown that we can make some adjustment to obtain an unbiased estimator for σ^2, e.g.,

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

All regression programs report s^2. We can therefore update the estimate variance of the OLS parameter from:

$$Var(\hat{\beta}) = \frac{\sigma_0^2}{n}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

To:

$$\widehat{Var}(\hat{\beta}) = s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

It can be proven that s^2 is the best quadratic unbiased estimator of σ^2 in the CLR model and is the best unbiased estimator of σ^2 in the CNLR model.

3.7 misspecification of linear regression models

Overspecification

Suppose that the DGP is generated by:

$$\boldsymbol{y} = \boldsymbol{X\beta_0} + \boldsymbol{u}, \boldsymbol{u} \sim \boldsymbol{IID}(\boldsymbol{0}, \sigma_0^2\boldsymbol{I})$$

(3.52)

If we have a correctly specified model, e.g.,

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

We will have the OLS estimate as:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

However, suppose we have an over-specified model:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{u}, \boldsymbol{u} \sim \boldsymbol{IID}(\boldsymbol{0}, \boldsymbol{\sigma^2 I})$$

(3.51)

We will have the OLS estimate as:

$$\widetilde{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{M_Z X})^{-1}\boldsymbol{X}'\boldsymbol{M_Z y}$$

Overspecification is strictly speaking not a misspecification because (3.52) is still embedded in (3.51). Indeed, if we replace $\boldsymbol{y}$ by $\boldsymbol{X\beta_0} + \boldsymbol{u}$, we have:

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta_0} + (\boldsymbol{X}'\boldsymbol{M_Z X})^{-1}\boldsymbol{X}'\boldsymbol{M_Z u}$$

It is obvious that both $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ are unbiased under the assumption of exogeneity (or fixed regressor). However, we could find that $\widehat{\boldsymbol{\beta}}$ is more efficient than $\widetilde{\boldsymbol{\beta}}$.

Proof:
e.g., we have:

$$Var(\widetilde{\boldsymbol{\beta}}) = \boldsymbol{E}((\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta_0})(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta_0})') = \sigma_0^2(\boldsymbol{X}'\boldsymbol{M_Z X})^{-1}$$

$$Var(\widehat{\boldsymbol{\beta}}) = \sigma_0^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Thus, we have:

$$Var(\widehat{\boldsymbol{\beta}})^{-1} - Var(\widetilde{\boldsymbol{\beta}})^{-1} = \frac{1}{\sigma_0^2}(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}'\boldsymbol{M_Z X}) = \frac{1}{\sigma_0^2}(\boldsymbol{P_Z X})'\boldsymbol{P_Z X}$$

(3.58)

(3.58) is a matrix of the form $\boldsymbol{B}'\boldsymbol{B}$, it must be a positive semidefinite matrix. Thus, this is equivalent to saying that: $Var(\widetilde{\boldsymbol{\beta}}) - Var(\widehat{\boldsymbol{\beta}})$ is a positive semidefinite matrix.

(3.58) suggests that $\widehat{\boldsymbol{\beta}}$ is more efficient than $\widetilde{\boldsymbol{\beta}}$, and $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ are equivalent only when $\boldsymbol{P_Z X} = 0$ which means $\boldsymbol{X}$ and $\boldsymbol{Z}$ are mutually orthogonal. In general, including regressors that do not belong in a model increase the variance of the estimates of coefficients on the regressors that do belong, and the increase can be very great in many cases.

Under-specification

Suppose we have the following DGP:

$$\boldsymbol{y} = \boldsymbol{X\beta_0} + \boldsymbol{Z\gamma_0} + \boldsymbol{u}, \boldsymbol{u} \sim \boldsymbol{IID}(\boldsymbol{0}, \sigma_0^2\boldsymbol{I})$$

(3.59)

And we have an underspecified model:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}, \boldsymbol{u} \sim \boldsymbol{IID}(\boldsymbol{0}, \boldsymbol{\sigma^2 I}) \tag{3.55}$$

Under-specification is a misspecification because (3.59) is NOT embedded in (3.51). Now we need to realize that the OLS estimate based on (3.55) is biased in general. e.g.,

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}) &= E((\boldsymbol{X'X})^{-1}\boldsymbol{X'y}) \\ E(\widehat{\boldsymbol{\beta}}) &= E((\boldsymbol{X'X})^{-1}\boldsymbol{X'}(\boldsymbol{X\beta_0} + \boldsymbol{Z\gamma_0} + \boldsymbol{u})) \\ &= \boldsymbol{\beta_0} + (\boldsymbol{X'X})^{-1}\boldsymbol{X'Z\gamma_0} \end{aligned} \tag{3.60}$$

(3.60) equals to zero only when $\boldsymbol{X'Z} = 0$ (e.g., $\boldsymbol{X}$ and $\boldsymbol{Z}$ are mutually orthogonal). The magnitude of the bias depends on $\boldsymbol{X}, \boldsymbol{Z}, \text{and } \boldsymbol{\gamma_0}$. Also, because this bias does not vanish as *n* increase, thus $\hat{\beta}$ is not consistent.

Since $\hat{\beta}$ is biased, we cannot calculate the variance-covariance matrix to evaluate its accuracy. Instead, we use the mean square error (**MSE**) matrix:

$$MSE(\widehat{\boldsymbol{\beta}}) = E((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})')$$

Note that $MSE(\widehat{\boldsymbol{\beta}})$ looks similar to $Var(\widehat{\boldsymbol{\beta}})$ but is different because the expected value of $\widehat{\boldsymbol{\beta}}$ is not $\boldsymbol{\beta_0}$ (see (3.60). The MSE matrix is equivalent to $Var(\widehat{\boldsymbol{\beta}})$ when $\widehat{\boldsymbol{\beta}}$ is unbiased, though not otherwise. Since we know that:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0} &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'}(\boldsymbol{X\beta_0} + \boldsymbol{Z\gamma_0} + \boldsymbol{u}) - \boldsymbol{\beta_0} \\ &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'Z\gamma_0} - (\boldsymbol{X'X})^{-1}\boldsymbol{X'u} \end{aligned}$$

Thus, we can derive the MSE matrix as:

$$\begin{aligned} MSE(\widehat{\boldsymbol{\beta}}) &= E\left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})'\right) \\ MSE(\widehat{\boldsymbol{\beta}}) &= \sigma_0^2(\boldsymbol{X'X})^{-1} + (\boldsymbol{X'X})^{-1}\boldsymbol{X'Z\gamma_0\gamma_0'Z'X}(\boldsymbol{X'X})^{-1} \end{aligned} \tag{3.62}$$

Suppose that $\widetilde{\boldsymbol{\beta}}$ is the OLS estimate based on the correctly specified model. Note that:

$$\begin{aligned} Var(\widetilde{\boldsymbol{\beta}}) &= \boldsymbol{E}((\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta_0})(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta_0})') \\ &= \sigma_0^2(\boldsymbol{X'M_Z X})^{-1} \end{aligned}$$

Thus, we would like to compare $MSE(\widehat{\boldsymbol{\beta}})$ with $Var(\widetilde{\boldsymbol{\beta}})$ (which is what it should be if the model is correctly specified). However, no unambiguous comparison is possible. It is possible that

some parameters may be estimated more efficiently by $\widehat{\boldsymbol{\beta}}$ and others more efficiently by $\widetilde{\boldsymbol{\beta}}$. Nevertheless, the covariance matrix of $\widehat{\boldsymbol{\beta}}$ calculated by programs will be incorrect as it only calculates the first component of (3.62).

Therefore, we conclude that under-specification will lead the OLS estimate to be **biased** and **inconsistent** and the variance-covariance matrix of the parameters to be **misleading**.