ECON 452* -- NOTE 2

Specification Errors in the Selection of Regressors

1. Two Alternative Models

Two Alternative Linear Regression Models for the Dependent Variable Y

Model 1

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i \qquad E\left(u_i \middle| X_{i1}, X_{i2}\right) = 0 \tag{1}$$

OLS estimation of equation (1) yields the OLS sample regression equation

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{u}_i . \tag{1*}$$

Model 2

$$Y_i = \beta_0 + \beta_1 X_{i1} + u_i . \tag{2}$$

OLS estimation of equation (2) yields the OLS sample regression equation

$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 X_{i1} + \tilde{u}_i . \tag{2*}$$

Fact: In practice, we don't know the true model that actually generated the sample data.

Two General Types of Specification Errors in Selecting Regressors

1. *Exclusion* of a *Relevant* Regressor

- The ***true* model** is **Model 1**

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i \qquad \textbf{(1)}$$

- The ***estimated* model** is **Model 2**, which incorrectly excludes from the population regression function the regressor X_{i2}.

$$Y_i = \beta_0 + \beta_1 X_{i1} + u_i\,. \qquad \textbf{(2)}$$

The OLS SRE for Model 2 is:

$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 X_{i1} + \tilde{u}_i\,. \qquad \textbf{(2*)}$$

- ***Question:*** What are the statistical properties of the OLS slope coefficient estimator $\tilde{\beta}_1$ in the misspecified model (2)?

Two General Types of Specification Errors in Selecting Regressors (continued)

2. *Inclusion* of an *Irrelevant* Regressor

- The ***true*** **model** is **Model 2**

$$Y_i = \beta_0 + \beta_1 X_{i1} + u_i \,. \tag{2}$$

- The ***estimated*** **model** is **Model 1**, which incorrectly includes in the population regression function the regressor X_{i2}.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i \tag{1}$$

The OLS SRE for model (1) is:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{u}_i \,. \tag{1*}$$

- ***Question:*** What are the statistical properties of the OLS slope coefficient estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ in the misspecified model (1)?

2. Exclusion of a Relevant Regressor

- The ***true*** **model** is **Model 1 given by PRE (1)**

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i \tag{1}$$

- The ***estimated*** **model** is **Model 2 given by PRE (2)**, which incorrectly excludes from the population regression function the regressor X_{i2}.

$$Y_i = \beta_0 + \beta_1 X_{i1} + u_i\,. \tag{2}$$

The OLS SRE for model (2) is:

$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 X_{i1} + \tilde{u}_i\,. \tag{2*}$$

The OLS SRE for model (2) is:

$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 X_{i1} + \tilde{u}_i . \qquad \textbf{(2*)}$$

- **Formulas**

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{N} x_{i1} y_i}{\sum_{i=1}^{N} x_{i1}^2} = \frac{\sum_{i=1}^{N} x_{i1} Y_i}{\sum_{i=1}^{N} x_{i1}^2} = \frac{\sum_{i=1}^{N} (X_{i1} - \overline{X}_1) Y_i}{\sum_{i=1}^{N} (X_{i1} - \overline{X}_1)^2} = \sum_{i=1}^{N} k_{i1} Y_i$$

$$\mathrm{Var}(\tilde{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{N} x_{i1}^2} = \frac{\sigma^2}{\sum_{i=1}^{N} (X_{i1} - \overline{X}_1)^2} = \frac{\sigma^2}{TSS_1}$$

where

$$x_{i1} = X_{i1} - \overline{X}_1; \qquad y_i = Y_i - \overline{Y}; \qquad k_{i1} = \frac{x_{i1}}{\sum_{i=1}^{N} x_{i1}^2} = \frac{X_{i1} - \overline{X}_1}{\sum_{i=1}^{N} (X_{i1} - \overline{X}_1)^2};$$

$$TSS_1 = \sum_{i=1}^{N} x_{i1}^2 = \sum_{i=1}^{N} (X_{i1} - \overline{X}_1)^2 .$$

- ***Question:*** What are the statistical properties of the OLS slope coefficient estimator $\tilde{\beta}_1$ in model (2) when model (1) is the true model?

- ***Result 1:*** $\tilde{\beta}_1$ **in model (2)** is in general a ***biased*** **(and *inconsistent*) estimator of β_1 when model (1) is the *true* model**.

$E(\tilde{\beta}_1) \neq \beta_1$ when X_{i1} and X_{i2} are correlated

$\text{Bias}(\tilde{\beta}_1) = E(\tilde{\beta}_1) - \beta_1 \neq 0$ when X_{i1} and X_{i2} are correlated

Proof: Consists of deriving the expression for $E(\tilde{\beta}_1)$ when model (1) is the true model.

1. Substitute for Y_i in the formula for $\tilde{\beta}_1$ the regression equation for model (1).

   Substitute $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i$ into $\tilde{\beta}_1 = \sum_{i=1}^{N} k_{i1} Y_i$.

$$
\begin{aligned}
\tilde{\beta}_1 &= \sum_{i=1}^{N} k_{i1} Y_i \\
&= \sum_{i=1}^{N} k_{i1} (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i) \\
&= \sum_{i=1}^{N} (\beta_0 k_{i1} + \beta_1 k_{i1} X_{i1} + \beta_2 k_{i1} X_{i2} + k_{i1} u_i) \\
&= \beta_0 \sum_{i=1}^{N} k_{i1} + \beta_1 \sum_{i=1}^{N} k_{i1} X_{i1} + \beta_2 \sum_{i=1}^{N} k_{i1} X_{i2} + \sum_{i=1}^{N} k_{i1} u_i \\
&= \beta_1 + \beta_2 \sum_{i=1}^{N} k_{i1} X_{i2} + \sum_{i=1}^{N} k_{i1} u_i
\end{aligned}
$$

$$\tilde{\beta}_1 = \beta_1 + \beta_2 \sum_{i=1}^{N} k_{i1} X_{i2} + \sum_{i=1}^{N} k_{i1} u_i$$

Note: We have used the computational properties $\sum_{i=1}^{N} k_{i1} = 0$ and $\sum_{i=1}^{N} k_{i1} X_{i1} = 1$.

2. Now take the expectation of $\tilde{\beta}_1$ conditional on X_{i1} and X_{i2}, $E\left(\tilde{\beta}_1 \middle| X_{i1}, X_{i2}\right)$, using the zero conditional mean error assumption $E\left(u_i \middle| X_{i1}, X_{i2}\right) = 0$.

$$\begin{aligned} E\left(\tilde{\beta}_1 \middle| X_{i1}, X_{i2}\right) &= \beta_1 + \beta_2 \sum_{i=1}^{N} k_{i1} X_{i2} + \sum_{i=1}^{N} k_{i1} E\left(u_i \middle| X_{i1}, X_{i2}\right) \\ &= \beta_1 + \beta_2 \sum_{i=1}^{N} k_{i1} X_{i2} \end{aligned}$$

Interpretation of the expression for $E(\tilde{\beta}_1)$:

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2 \sum_{i=1}^{N} k_{i1} X_{i2} .$$

- Suppose we estimate by OLS the auxiliary linear regression equation

$$X_{i2} = \beta_{20} + \beta_{21} X_{i1} + v_i .$$

- The OLS sample regression equation corresponding to this auxiliary regression equation is:

$$X_{i2} = \hat{\beta}_{20} + \hat{\beta}_{21} X_{i1} + \hat{v}_i .$$

- The OLS coefficient estimate $\hat{\beta}_{21}$ is given by the formula

$$\hat{\beta}_{21} = \frac{\sum_{i=1}^{N} x_{i1} x_{i2}}{\sum_{i=1}^{N} x_{i1}^2} = \sum_{i=1}^{N} k_{i1} x_{i2} = \frac{\sum_{i=1}^{N} (X_{i1} - \overline{X}_1)(X_{i2} - \overline{X}_2)}{\sum_{i=1}^{N} (X_{i1} - \overline{X}_1)^2}$$

where $x_{i2} = X_{i2} - \overline{X}_2$ and $k_{i1} = \dfrac{x_{i1}}{\sum_{i=1}^{N} x_{i1}^2} = \dfrac{X_{i1} - \overline{X}_1}{\sum_{i=1}^{N} (X_{i1} - \overline{X}_1)^2}$.

- But $x_{i2} = X_{i2} - \overline{X}_2$ and $\sum_{i=1}^{N} k_{i1} = 0$, so that this formula for $\hat{\beta}_{21}$ can be written as

$$\hat{\beta}_{21} = \sum_{i=1}^{N} k_{i1} x_{i2} = \sum_{i=1}^{N} k_{i1} (X_{i2} - \overline{X}_2) = \sum_{i=1}^{N} k_{i1} X_{i2} - \overline{X}_2 \sum_{i=1}^{N} k_{i1} = \sum_{i=1}^{N} k_{i1} X_{i2} .$$

❑ ***Result:*** The expression for $E(\tilde{\beta}_1)$ can be written as

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2 \sum_{i=1}^{N} k_{i1} X_{i2} = \beta_1 + \beta_2 \hat{\beta}_{21} \qquad \text{since} \qquad \hat{\beta}_{21} = \sum_{i=1}^{N} k_{i1} X_{i2} .$$

Omitted Variables Bias

- The ***omitted variables bias*** **of the OLS estimator $\tilde{\beta}_1$ in model (2)** is:

$$\text{Bias}(\tilde{\beta}_1) = E(\tilde{\beta}_1) - \beta_1 = \beta_2 \hat{\beta}_{21}.$$

- The **OLS estimator $\tilde{\beta}_1$ in model (2)** is a ***biased*** **estimator** of the slope coefficient β_1 if **the sample values of X_1 and X_2 are *correlated***.

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2 \hat{\beta}_{21} \neq \beta_1 \quad \text{if} \quad \hat{\beta}_{21} = \sum_{i=1}^{N} k_{i1} x_{i2} = \frac{\sum_{i=1}^{N} x_{i1} x_{i2}}{\sum_{i=1}^{N} x_{i1}^2} \neq 0$$

$$\text{if} \quad \sum_{i=1}^{N} x_{i1} x_{i2} = \sum_{i=1}^{N} (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2) \neq 0$$

if **sample values of X_1 and X_2 have *nonzero covariance***.

if **sample values of X_1 and X_2 are *correlated***.

- **Only if the sample values of X_1 and X_2 are *uncorrelated* is the OLS estimator $\tilde{\beta}_1$ in model (2) an *unbiased* estimator** of the slope coefficient β_1.

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2\hat{\beta}_{21} = \beta_1 \qquad \text{only if} \qquad \hat{\beta}_{21} = 0$$

$$\text{only if} \qquad \sum_{i=1}^{N} x_{i1}x_{i2} = \sum_{i=1}^{N}(X_{i1} - \overline{X}_1)(X_{i2} - \overline{X}_2) = 0$$

only if **sample values of X_1 and X_2 have *zero covariance***.

only if **sample values of X_1 and X_2 are *uncorrelated***.

Note:

The **sample *covariance* of X_1 and X_2** is:

$$\hat{\text{Cov}}(X_1, X_2) = \frac{\sum_{i=1}^{N} x_{i1}x_{i2}}{N-1} = \frac{\sum_{i=1}^{N}(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{N-1}.$$

The **sample *correlation coefficient* of X_1 and X_2** is:

$$\hat{\text{Corr}}(X_1, X_2) = \frac{\sum_{i=1}^{N} x_{i1}x_{i2}}{\sqrt{\sum_{i=1}^{N} x_{i1}^2}\sqrt{\sum_{i=1}^{N} x_{i2}^2}} = \frac{\sum_{i=1}^{N}(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{\sqrt{\sum_{i=1}^{N}(X_{i1} - \bar{X}_1)^2}\sqrt{\sum_{i=1}^{N}(X_{i2} - \bar{X}_2)^2}}.$$

Both the **sample *covariance* of X_1 and X_2** and the **sample *correlation* of X_1 and X_2** are ***zero*** if and only if

$$\sum_{i=1}^{N} x_{i1}x_{i2} = \sum_{i=1}^{N}(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2) = 0$$

- ***Direction of Bias:*** The ***direction* (*sign*)** of the ***omitted variables bias*** of the **OLS estimator** $\tilde{\beta}_1$ in model (2) is determined by the sign of the product $\beta_2 \hat{\beta}_{21}$ in the following expression for $\text{Bias}(\tilde{\beta}_1)$:

$$\text{Bias}(\tilde{\beta}_1) = E(\tilde{\beta}_1) - \beta_1 = \beta_2 \hat{\beta}_{21}.$$

1. If β_2 and $\hat{\beta}_{21}$ have the ***same* sign**, then their product is positive and $\tilde{\beta}_1$ **is an *upward biased* estimator of β_1**:

$$\text{Bias}(\tilde{\beta}_1) = E(\tilde{\beta}_1) - \beta_1 = \beta_2 \hat{\beta}_{21} > 0.$$

2. If β_2 and $\hat{\beta}_{21}$ have ***different* signs**, then their product is negative and $\tilde{\beta}_1$ **is a *downward biased* estimator of β_1**:

$$\text{Bias}(\tilde{\beta}_1) = E(\tilde{\beta}_1) - \beta_1 = \beta_2 \hat{\beta}_{21} < 0.$$

Omitted Variables Bias -- Generalization:

- The ***true*** **model** is **Model 1** as given by the population regression equation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{k-1} X_{ik-1} + \beta_k X_{ik} + u_i \qquad \textbf{(1)}$$

- The ***estimated*** **model** is **Model 2**, which incorrectly excludes from the population regression function the k-th regressor X_{ik}.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{k-1} X_{ik-1} + u_i \qquad \textbf{(2)}$$

  The OLS SRE for Model 2, the misspecified model, is:

$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 X_{i1} + \tilde{\beta}_2 X_{i2} + \cdots + \tilde{\beta}_{k-1} X_{ik-1} + \tilde{u}_i \qquad \textbf{(2)}$$

- ***Question:*** What are the statistical properties of the OLS slope coefficient estimators $\tilde{\beta}_j$ (j = 1, 2, …, k-1) in the misspecified model (2)?

<u>Omitted Variables Bias – General Result</u>:

- The expression for $E(\tilde{\beta}_j \mid X)$, the conditional expectation of $\tilde{\beta}_j$ for any given regressor matrix X, can be shown to be

  $$E(\tilde{\beta}_j \mid X) = \beta_j + \beta_k \hat{\beta}_{kj} \qquad \text{(for j = 1, 2, …, k-1)}$$

  where $\hat{\beta}_{kj}$ is OLS slope coefficient estimate in the following **auxiliary OLS regression** of the omitted regressor X_{ik} on all the included regressors X_{ij} (j = 1, 2, …, k-1) and an intercept constant:

  $$X_{ik} = \hat{\beta}_{k0} + \hat{\beta}_{k1}X_{i1} + \cdots + \hat{\beta}_{kj}X_{ij} + \cdots + \hat{\beta}_{k,k-1}X_{ik-1} + \hat{v}_{ik} \qquad \textbf{(3)}$$

- If the ***included*** **regressor X_{ij} is partially *correlated* with the *omitted* regressor X_{ik}** – i.e., if $\hat{\beta}_{kj} \neq 0$ in the auxiliary OLS regression equation (3) – then the OLS slope coefficient estimator $\tilde{\beta}_j$ of X_{ij} in the misspecified model (2) is a **biased** (and inconsistent) estimator of the slope coefficient β_j in the true population regression equation (1).

- Only if the ***included*** **regressor X_{ij} is partially *uncorrelated* with the *omitted* regressor X_{ik}** – i.e., if $\hat{\beta}_{kj} = 0$ in the auxiliary OLS regression equation (3) – will the OLS slope coefficient estimator $\tilde{\beta}_j$ of X_{ij} in the misspecified model (2) be an **unbiased** (and consistent) estimator of the slope coefficient β_j in the true population regression equation (1).

❑ ***<u>Result 2</u>:*** The **variance of $\tilde{\beta}_1$ in model (2) is *less than* (or equal to)** the **variance of $\hat{\beta}_1$ in model (1)**, regardless of whether model (1) or model (2) is the true model: i.e., $\text{Var}(\tilde{\beta}_1) \le \text{Var}(\hat{\beta}_1)$.

- The **variance of $\tilde{\beta}_1$ in model (2)** is given by the formula

$$\text{Var}(\tilde{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{N} x_{i1}^2} = \frac{\sigma^2}{\sum_{i=1}^{N} (X_{i1} - \bar{X}_1)^2} = \frac{\sigma^2}{TSS_1}$$

where:

$\sigma^2 = \text{Var}(u_i) =$ the constant variance of the error terms u_i;

$TSS_1 = \sum_{i=1}^{N} (X_{i1} - \bar{X}_1)^2 = \sum_{i=1}^{N} x_{i1}^2 =$ total sum-of-squares for X_1.

- The **variance of $\hat{\beta}_1$ in model (1)** is given by the formula

$$\operatorname{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{N} x_{i1}^2 (1 - R_1^2)} = \frac{\sigma^2}{\sum_{i=1}^{N} (X_{i1} - \bar{X}_1)^2 (1 - R_1^2)} = \frac{\sigma^2}{TSS_1 (1 - R_1^2)}$$

where:

R_1^2 = the R^2 from OLS regression of X_1 on X_2 and an intercept constant

= the R^2 from OLS estimation of $X_{i1} = \beta_{10} + \beta_{12} X_{i2} + \varepsilon_i$

= the R^2 from the OLS SRE $X_{i1} = \hat{\beta}_{10} + \hat{\beta}_{12} X_{i2} + \hat{\varepsilon}_i$.

- **Compare** $\text{Var}(\tilde{\beta}_1)$ and $\text{Var}(\hat{\beta}_1)$:

$$\text{Var}(\tilde{\beta}_1) = \frac{\sigma^2}{\text{TSS}_1} \qquad \text{and} \qquad \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\text{TSS}_1(1-R_1^2)}.$$

Since $0 \le R_1^2 < 1$ implies that $0 < 1 - R_1^2 \le 1$, and $\text{TSS}_1 > 0$, it follows that

$$\text{TSS}_1(1-R_1^2) \le \text{TSS}_1 \quad \Rightarrow \quad \frac{1}{\text{TSS}_1(1-R_1^2)} \ge \frac{1}{\text{TSS}_1}$$

$$\Rightarrow \quad \frac{\sigma^2}{\text{TSS}_1(1-R_1^2)} \ge \frac{\sigma^2}{\text{TSS}_1} \quad \text{since } \sigma^2 > 0$$

$$\Rightarrow \quad \text{Var}(\hat{\beta}_1) \ge \text{Var}(\tilde{\beta}_1)$$

- ***Result 2:*** The **variance of the OLS estimator $\hat{\beta}_1$** in model (1) is ***greater than*** **or** ***equal to*** the **variance of the OLS estimator $\tilde{\beta}_1$** in model (2):

$$\mathrm{Var}(\hat{\beta}_1) \geq \mathrm{Var}(\tilde{\beta}_1)$$

where:

$$\mathrm{Var}(\tilde{\beta}_1) = \frac{\sigma^2}{TSS_1} \qquad \text{and} \qquad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{TSS_1(1 - R_1^2)}.$$

Special Case: $\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}(\tilde{\beta}_1)$ if and only if R_1^2, the R^2 from OLS regression of X_1 on X_2 and an intercept constant, equals zero -- i.e., if and only if the included regressor X_1 is uncorrelated in the sample with the excluded regressor X_2:

$$R_1^2 = 0 \qquad \Rightarrow \qquad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{TSS_1(1 - R_1^2)} = \frac{\sigma^2}{TSS_1} = \mathrm{Var}(\tilde{\beta}_1).$$

Otherwise, $\mathrm{Var}(\hat{\beta}_1) > \mathrm{Var}(\tilde{\beta}_1)$:

$$0 < R_1^2 < 1 \qquad \Rightarrow \qquad \mathrm{Var}(\hat{\beta}_1) > \mathrm{Var}(\tilde{\beta}_1).$$

- ***Result 3:*** The **estimator of the error variance from OLS estimation of model (2)** is a ***biased*** **(and *inconsistent*) estimator of the error variance σ^2** in the true model, model (1).

- The OLS estimator of the error variance σ^2 from model (2) is given by the usual formula:

$$\tilde{\sigma}^2 = \frac{RSS_{(2)}}{N-2} = \frac{\sum_{i=1}^{N}\tilde{u}_i^2}{N-2} = \frac{\sum_{i=1}^{N}(Y_i - \tilde{\beta}_0 - \tilde{\beta}_1 X_{i1})^2}{N-2}$$

where

$$RSS_{(2)} = \sum_{i=1}^{N}\tilde{u}_i^2 = \sum_{i=1}^{N}(Y_i - \tilde{\beta}_0 - \tilde{\beta}_1 X_{i1})^2$$

= the residual sum-of-squares from OLS estimation of model (2)

- ***Direction of Bias:*** More specifically, it can be shown that the **error variance estimator $\tilde{\sigma}^2$ from model (2)** is an ***upward biased*** **estimator of the true error variance σ^2**.

$$E(\tilde{\sigma}^2) > \sigma^2$$

Moreover, this upward bias of $\tilde{\sigma}^2$ does not vanish even in the special case when X_1 and X_2 are uncorrelated.

- ***Result 4:*** The **OLS estimators of the *variances* of the *coefficient estimates* from misspecified model (2)** are ***biased* (and *inconsistent*)**.

- The OLS estimator of the variance of $\tilde{\beta}_1$ from model (2) is given by the usual formula:

$$\hat{\text{Var}}(\tilde{\beta}_1) = \frac{\tilde{\sigma}^2}{\text{TSS}_1} = \frac{\tilde{\sigma}^2}{\sum_{i=1}^{N}(X_{i1} - \overline{X}_1)^2}$$

where $$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{N}\tilde{u}_i^2}{N-2} = \frac{\sum_{i=1}^{N}(Y_i - \tilde{\beta}_0 - \tilde{\beta}_1 X_{i1})^2}{N-2}.$$

- But $\hat{\text{Var}}(\tilde{\beta}_1)$ is a biased and inconsistent estimator of $\text{Var}(\hat{\beta}_1)$ in the true model, model (1), for two reasons.

  1. $\hat{\text{Var}}(\tilde{\beta}_1)$ uses an incorrect formula for $\text{Var}(\hat{\beta}_1)$.

     The correct formula for $\text{Var}(\hat{\beta}_1)$ is: $\text{Var}(\hat{\beta}_1) = \dfrac{\sigma^2}{TSS_1(1-R_1^2)}$.

     But $\hat{\text{Var}}(\tilde{\beta}_1)$ uses the formula: $\text{Var}(\tilde{\beta}_1) = \dfrac{\sigma^2}{TSS_1}$.

  2. $\hat{\text{Var}}(\tilde{\beta}_1)$ uses the upward biased estimator $\tilde{\sigma}^2$ of the error variance σ^2.

     So even if the formula for $\hat{\text{Var}}(\tilde{\beta}_1)$ was correct, it would still be a biased and inconsistent estimator of the variance of the OLS estimator of β_1.

- ***<u>Result 5</u>:*** The usual **procedures for statistical inference** -- hypothesis testing and confidence interval estimation -- based on OLS estimation of the misspecified model, model (2), **are in general *invalid***. They are likely to lead to incorrect and misleading inferences respecting the statistical significance of the OLS coefficient estimates $\tilde{\beta}_0$ and $\tilde{\beta}_1$ in model (2).

3. Inclusion of an Irrelevant Regressor

- The ***true* model** is **Model 2 given by PRE (2)**:

$$Y_i = \beta_0 + \beta_1 X_{i1} + u_i . \qquad \textbf{(2)}$$

- The ***estimated* model** is **Model 1 given by PRE (1)**, which incorrectly includes in the population regression function the regressor X_{i2}.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i \qquad \textbf{(1)}$$

The OLS SRE for model (1) is:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{u}_i . \qquad \textbf{(1*)}$$

- **Formulas for OLS Coefficient Estimators in Model (1)**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^N x_{i2}^2 \sum_{i=1}^N x_{i1} y_i - \sum_{i=1}^N x_{i1} x_{i2} \sum_{i=1}^N x_{i2} y_i}{\sum_{i=1}^N x_{i1}^2 \sum_{i=1}^N x_{i2}^2 - (\sum_{i=1}^N x_{i1} x_{i2})^2} = \frac{\sum_{i=1}^N x_{i2}^2 \sum_{i=1}^N x_{i1} Y_i - \sum_{i=1}^N x_{i1} x_{i2} \sum_{i=1}^N x_{i2} Y_i}{\sum_{i=1}^N x_{i1}^2 \sum_{i=1}^N x_{i2}^2 - (\sum_{i=1}^N x_{i1} x_{i2})^2}$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^N x_{i1}^2 \sum_{i=1}^N x_{i2} y_i - \sum_{i=1}^N x_{i1} x_{i2} \sum_{i=1}^N x_{i1} y_i}{\sum_{i=1}^N x_{i1}^2 \sum_{i=1}^N x_{i2}^2 - (\sum_{i=1}^N x_{i1} x_{i2})^2} = \frac{\sum_{i=1}^N x_{i1}^2 \sum_{i=1}^N x_{i2} Y_i - \sum_{i=1}^N x_{i1} x_{i2} \sum_{i=1}^N x_{i1} Y_i}{\sum_{i=1}^N x_{i1}^2 \sum_{i=1}^N x_{i2}^2 - (\sum_{i=1}^N x_{i1} x_{i2})^2}$$

where:

$$y_i = Y_i - \bar{Y}; \qquad x_{i1} = X_{i1} - \bar{X}_1; \qquad x_{i2} = X_{i2} - \bar{X}_2;$$

$$\sum_{i=1}^{N} x_{i1} y_i = \sum_{i=1}^{N} x_{i1}(Y_i - \bar{Y}) = \sum_{i=1}^{N} x_{i1} Y_i - \bar{Y}\sum_{i=1}^{N} x_{i1} = \sum_{i=1}^{N} x_{i1} Y_i \quad \text{b/c} \quad \sum_{i=1}^{N} x_{i1} = 0$$

$$\sum_{i=1}^{N} x_{i2} y_i = \sum_{i=1}^{N} x_{i2}(Y_i - \bar{Y}) = \sum_{i=1}^{N} x_{i2} Y_i - \bar{Y}\sum_{i=1}^{N} x_{i2} = \sum_{i=1}^{N} x_{i2} Y_i \quad \text{b/c} \quad \sum_{i=1}^{N} x_{i2} = 0$$

- ***Question:*** What are the statistical properties of the OLS slope coefficient estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ in the misspecified model (1)?

  The OLS SRE for model (1) is:

  $$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{u}_i . \qquad \textbf{(1*)}$$

❑ ***<u>Result 1</u>:*** Under the zero mean error assumption A2, which implies that $E(u_i) = 0$, it can be shown that $\hat{\beta}_1$ is **an *unbiased* (and *consistent*) estimator of the slope coefficient β_1.**

$$E(\hat{\beta}_1) = \beta_1 \qquad \Rightarrow \qquad Bias(\hat{\beta}_1) = E(\hat{\beta}_1) - \beta_1 = \beta_1 - \beta_1 = 0 .$$

❑ ***<u>Result 2</u>:*** Under the zero mean error assumption A2, which implies that $E(u_i) = 0$, it can be shown that $\hat{\beta}_2$ is **an *unbiased* (and *consistent*) estimator of $\beta_2 = 0$**, where 0 is the true value of β_2 in true model (2).

$$E(\hat{\beta}_2) = 0 = \beta_2 \qquad \Rightarrow \qquad Bias(\hat{\beta}_2) = E(\hat{\beta}_2) - \beta_2 = 0 - 0 = 0 .$$

- ***Result 3:*** The **OLS estimator of the *error variance* from model (1)** is an ***unbiased* (and *consistent*) estimator of the error variance σ^2** in the true model, model (2).

- The OLS estimator of the error variance σ^2 from model (1), denoted as $\hat{\sigma}^2$, is given by the usual formula:

$$\hat{\sigma}^2 = \frac{RSS_{(1)}}{N-3} = \frac{\sum_{i=1}^{N}\hat{u}_i^2}{N-3} = \frac{\sum_{i=1}^{N}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2})^2}{N-3}$$

where

$$RSS_{(1)} = \sum_{i=1}^{N}\hat{u}_i^2 = \sum_{i=1}^{N}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2})^2$$

= the residual sum-of-squares from OLS estimation of model (1)

❑ ***Result 4*:** The **OLS estimators of the *variances* of the *coefficient estimates* from model (1)** are ***unbiased* (and *consistent*)**.

- The OLS estimators of the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ from model (1) are given by the usual formulas:

$$\hat{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{TSS_1(1-R_1^2)} \qquad \text{and} \qquad \hat{Var}(\hat{\beta}_2) = \frac{\hat{\sigma}^2}{TSS_2(1-R_2^2)}$$

where:

$\hat{\sigma}^2$ = the OLS estimator of the error variance from model (1)

$$TSS_1 = \sum_{i=1}^{N}(X_{i1} - \overline{X}_1)^2 = \sum_{i=1}^{N} x_{i1}^2 = \text{total sum-of-squares for } X_1$$

R_1^2 = the R^2 from OLS regression of X_1 on X_2 and an intercept constant

= the R^2 from OLS estimation of $X_{i1} = \beta_{10} + \beta_{12}X_{i2} + \varepsilon_i$

= the R^2 from the OLS SRE $X_{i1} = \hat{\beta}_{10} + \hat{\beta}_{12}X_{i2} + \hat{\varepsilon}_i$.

$$TSS_2 = \sum_{i=1}^{N}(X_{i2} - \overline{X}_2)^2 = \sum_{i=1}^{N} x_{i2}^2 = \text{total sum-of-squares for } X_2$$

R_2^2 = the R^2 from OLS regression of X_2 on X_1 and an intercept constant

= the R^2 from OLS estimation of $X_{i2} = \beta_{20} + \beta_{21}X_{i1} + v_i$

= the R^2 from the OLS SRE $X_{i2} = \hat{\beta}_{20} + \hat{\beta}_{21}X_{i1} + \hat{v}_i$.

- ***Result 5:*** The usual **procedures for statistical inference --** hypothesis testing and confidence interval estimation **-- remain *valid***.

- ***Result 6:*** The variances of the OLS coefficient estimators in Model (1) are generally larger than the variances of the corresponding OLS coefficient estimators in the true model, Model (2).

- The ***variance* of $\hat{\beta}_1$ from Model 1** is: $\mathrm{Var}(\hat{\beta}_1) = \dfrac{\sigma^2}{\mathrm{TSS}_1(1 - R_1^2)}$.

- The ***variance* of $\tilde{\beta}_1$ from Model 2** is: $\mathrm{Var}(\tilde{\beta}_1) = \dfrac{\sigma^2}{\mathrm{TSS}_1}$.

- As previously demonstrated, the variance of $\hat{\beta}_1$ from model (1) is generally ***greater than*** the variance of $\tilde{\beta}_1$ from model (2):

$$\mathrm{Var}(\hat{\beta}_1) \geq \mathrm{Var}(\tilde{\beta}_1)$$

- ***Implications:***

$\hat{\beta}_1$ **from model (1)** is ***inefficient*** relative to $\tilde{\beta}_1$ **from model (2)**.

$\hat{\beta}_1$ **from model (1)** is ***less precise*** than $\tilde{\beta}_1$ **from model (2)**.

4. Choosing Between Model 1 and Model 2

In practice, one does not know whether Model 1 or Model 2 is the true model.

***Model 1*:** $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i$ **(1)**

***Model 2*:** $Y_i = \beta_0 + \beta_1 X_{i1} + u_i$ **(2)**

Question: How do applied researchers choose between Model 1 and Model 2?

Answer: After estimating Model 1, perform a ***two-tail*** **t-test or F-test** of the one **exclusion restriction** $\boldsymbol{\beta_2 = 0}$, which is the coefficient restriction that Model 2 imposes on Model 1.

- **Null and Alternative Hypotheses**

  H_0: $\beta_2 = 0$ $\Rightarrow$ Model 2 is the true model

  H_1: $\beta_2 \neq 0$ $\Rightarrow$ Model 1 is the true model

  A test of H_0 against H_1 is not only a test of Model 2 against Model 1, it is also a means of choosing between two alternative estimators of the slope coefficient β_1.

- The two alternative OLS estimators of the coefficient β_1 are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} x_{i2}^2 \sum_{i=1}^{N} x_{i1} y_i - \sum_{i=1}^{N} x_{i1} x_{i2} \sum_{i=1}^{N} x_{i2} y_i}{\sum_{i=1}^{N} x_{i1}^2 \sum_{i=1}^{N} x_{i2}^2 - (\sum_{i=1}^{N} x_{i1} x_{i2})^2} = \frac{\sum_{i=1}^{N} x_{i2}^2 \sum_{i=1}^{N} x_{i1} Y_i - \sum_{i=1}^{N} x_{i1} x_{i2} \sum_{i=1}^{N} x_{i2} Y_i}{\sum_{i=1}^{N} x_{i1}^2 \sum_{i=1}^{N} x_{i2}^2 - (\sum_{i=1}^{N} x_{i1} x_{i2})^2}$$

= the **OLS estimator of β_1 in Model 1**

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{N} x_{i1} y_i}{\sum_{i=1}^{N} x_{i1}^2} = \frac{\sum_{i=1}^{N} x_{i1} Y_i}{\sum_{i=1}^{N} x_{i1}^2} = \frac{\sum_{i=1}^{N} (X_{i1} - \overline{X}_1) Y_i}{\sum_{i=1}^{N} (X_{i1} - \overline{X}_1)^2}$$

= the **OLS estimator of β_1 in Model 2**

- **Compare the statistical properties of the two alternative estimators of $\boldsymbol{\beta_1}$** under the null hypothesis H_0 and under the alternative hypothesis H_1.

Under the null hypothesis H_0 -- if H_0 is true, meaning if $\beta_2 = 0$:

- Both $\tilde{\beta}_1$ and $\hat{\beta}_1$ are ***unbiased* and *consistent*** for β_1.
- But $\tilde{\beta}_1$ is ***efficient*** relative to $\hat{\beta}_1$; $\tilde{\beta}_1$ has ***smaller variance*** than $\hat{\beta}_1$.

Result: $\tilde{\beta}_1$ is preferred to $\hat{\beta}_1$ if the null hypothesis $\beta_2 = 0$ is true.

Under the alternative hypothesis H_1 -- if H_1 is true, meaning if $\beta_2 \neq 0$:

- $\tilde{\beta}_1$ is ***biased* and *inconsistent*** for β_1, but has smaller variance than $\hat{\beta}_1$.
- $\hat{\beta}_1$ is ***unbiased* and *consistent*** for β_1, but has larger variance than $\tilde{\beta}_1$.

Result: $\hat{\beta}_1$ is preferred to $\tilde{\beta}_1$ if the alternative hypothesis $\beta_2 \neq 0$ is true.

- **Choice Between Model 1 and Model 2**

The choice between Model 1 and Model 2 depends on which of two possible test outcomes is obtained.

1. If the null hypothesis H_0: $\beta_2 = 0$ is ***retained***, **choose *Model 2***.

Nonrejection of H_0: $\beta_2 = 0$ constitutes sample evidence that Model 2 is the true model.

If Model 2 is the true model, both $\hat{\beta}_1$ and $\tilde{\beta}_1$ are unbiased and consistent for β_1, but $\tilde{\beta}_1$ is efficient relative to $\hat{\beta}_1$.

2. If the null hypothesis H_0: $\beta_2 = 0$ is ***rejected***, **choose *Model 1***.

Rejection of H_0: $\beta_2 = 0$ constitutes sample evidence that Model 2 is *not* the true model -- that Model 1 is the true model.

If Model 1 is the true model, $\hat{\beta}_1$ is unbiased and consistent for β_1, whereas $\tilde{\beta}_1$ is biased and inconsistent for β_1.