Biases due to Omitted Variables (Specification Bias Analysis)

Let the true equation be

$$Y_t = a_0 + a_1 X_{1t} + a_2 X_{2t} + U_t$$

where U_t satisfies all assumptions. Suppose we estimate instead:

$$Y_t = a_0' + a_1' X_{1t} + V_t$$

Clearly, $V_t = (a_0 - a_0') + (a_1 - a_1') X_{1t} + a_2 X_{2t} + U_t$. Define the OLS estimator of a_1':

$$\hat{a}_1' = \frac{\sum (X_{1t} - \bar{X}_1) Y_t}{\sum (X_{1t} - \bar{X}_1)^2} = \frac{\sum (X_{1t} - \bar{X}_1)[a_0 + a_1 X_{1t} + a_2 X_{2t} + U_t]}{\sum (X_{1t} - \bar{X}_1)^2}$$

$$= \frac{a_1 \sum (X_{1t} - \bar{X}_1) X_{1t} + a_2 \sum (X_{1t} - \bar{X}_1) X_{2t} + \sum (X_{1t} - \bar{X}_1) U_t}{\sum (X_{1t} - \bar{X}_1)^2}$$

$$= a_1 + a_2 \frac{\sum (X_{1t} - \bar{X}_1) X_{2t}}{\sum (X_{1t} - \bar{X}_1)^2} + \frac{\sum (X_{1t} - \bar{X}_1) U_t}{\sum (X_{1t} - \bar{X}_1)^2}$$

Note that $\sum (X_1 - \bar{X}_1) X_2 / \sum (X_1 - \bar{X}_1)^2$ is the covariance

(see p. 75)

This is so because:

$$\hat{\alpha}_1' = \frac{\sum (y_t - \bar{y})(x_{1t} - \bar{x}_1)}{\sum (x_{1t} - \bar{x}_1)} = \frac{\sum (x_{1t} - \bar{x}_1) y_t}{\sum (x_{1t} - \bar{x}_1)^2} + \frac{\sum (x_{1t} - \bar{x}_1)\bar{y}}{\sum (x_{1t} - \bar{x}_1)}$$

the second term on the r.h.s. is zero as:

$$\frac{\sum (x_{1t} - \bar{x}_1)\bar{y}}{\sum (x_{1t} - \bar{x}_1)^2} = \bar{y} \frac{\sum (x_{1t} - \bar{x}_1)}{\sum (x_{1t} - \bar{x}_1)^2} = 0 \text{ since } \sum (x_{1t} - \bar{x}_1) = 0$$

Substitute the true model:

$$\hat{\alpha}_1' = \frac{\sum (x_{1t} - \bar{x}_1) y_t}{\sum (x_{1t} - \bar{x}_1)^2} = \frac{\sum (x_{1t} - \bar{x}_1)(\alpha_0 + \alpha_1 x_{1t} + \alpha_2 x_{2t} + u_t)}{\sum (x_{1t} - \bar{x}_1)^2}$$

$$= \alpha_1 + \alpha_2 \frac{\sum (x_{1t} - \bar{x}_1) x_{2t}}{\sum (x_{1t} - \bar{x}_1)^2} + \frac{\sum (x_{1t} - \bar{x}_1) u_t}{\sum (x_{1t} - \bar{x}_1)^2}$$

as: (*) $\dfrac{\alpha_0 \sum (x_{1t} - \bar{x}_1)}{\sum (x_{1t} - \bar{x}_1)^2} = 0$ since $\sum (x_{1t} - \bar{x}_1) = 0$

(*) $\alpha_1 \dfrac{\sum (x_{1t} - \bar{x}_1) x_{1t}}{\sum (x_{1t} - \bar{x}_1)^2} = \alpha_1 \dfrac{\sum (x_{1t} - \bar{x}_1)(x_{1t} - \bar{x}_1)}{\sum (x_{1t} - \bar{x}_1)^2} = \alpha_1$

between x_1 and x_2 divided by the variance of x_2

$$E(\hat{a}_1') = a_1 + a_2 \frac{\sigma_{x_1 x_2}}{\sigma^2_{x_1}}$$

since $E(x_{tt} v_t) = 0$.

Consider the auxiliary regression of the omitted variable against included variables:

$$x_{2t} = p_0 + p_1 x_{1t} + \zeta_t$$

The OLS estimator is:

$$\hat{p}_1 = \frac{\sigma_{x_1 x_2}}{\sigma^2_{x_1}}$$

We write:

$$E(\hat{a}_1') = a_1 + a_2 p_1$$

Hence: the bias $E(\hat{a}_1') - a_1 = a_2 p_1$ is the product of the true coefficient on excluded variable and the auxiliary coefficient.

Note: The estimator $\hat{a}_1'$ is unbiased if either

i) $a_2 = 0$

ii) $\rho_{1,2} = \sigma_{x_1 x_2} / \sigma^2_{x_1} = 0$

Applications to Bivariate Case

1. Production function estimation

Suppose true p.f. is:

$$q_t = \alpha l_t + \beta c_t + u_t$$

where small letters denote logs. Suppose true measure of labor L is the measured labor L^* times a quality index Z:

$$L_t = Z_t L^*_t$$

Then $l_t = z_t + l^*_t$. Suppose we estimate:

$$q_t = \alpha l^*_t + \beta c_t + v_t$$

The true p.f. in terms of observables is

$$q_t = \alpha l_t^0 + \alpha z_t + \beta c_t + v_t$$

and we have effectively omitted z_t. Then

$$E(\hat{\alpha}) = \alpha + \alpha p_{zl^0}$$

$$E(\hat{\beta}) = \beta + \alpha p_{zc}$$

where the auxiliary regression is

$$z_t = p_{zc} c_t + p_{zl^0} l_t^0 + \xi_t$$

Can we say something about probable signs of p_{zc} and p_{zl^0}? Seems plausible that $p_{zc} > 0$ and $p_{zl^0} > 0$. So we expect that both α and β to be overestimated here.

Note: Both $\hat{\alpha}$ and $\hat{\beta}$ are generally biased

Error in Measurement and Specification Bias

Let true model be:

$$y_t = a_0 + a_1 x_t + u_t \quad ; \quad u_t \text{ Gauss-Markov.}$$

1. Suppose we observe the independent variable with random error

$$x_t^\circ = x_t + \epsilon_t \qquad \text{where} \quad \epsilon_t \sim N(0, \sigma_\epsilon^2)$$
$$\text{and} \quad E(\epsilon_t, x_t) = 0.$$

What happens if we estimate using x_t° instead of x_t?
First write the true model in terms of observed variable,

$$y_t = a_0 + a_1 (x_t^\circ - \epsilon_t) + u_t$$

or $$y_t = a_0 + a_1 x_t^\circ + u_t - a_1 \epsilon_t .$$

In a sense, we estimate this model but omit the (random) variable ϵ_t, which is not observable. By specification bias analysis we obtain:

$$E(\hat{a}_1) = a_1 - a_1 p_1$$

where auxiliary regression is $\epsilon_t = p_0 + p_1 x_t^\circ + \zeta_t$.

But here we can say more about ρ_1. Because:

$$\rho_1 = \frac{\sigma_{\epsilon x^\circ}}{\sigma^2_{x^\circ}} = \frac{\sigma^2_\epsilon}{\sigma^2_{x^\circ}}$$

where second equality uses fact that $x^\circ_t = x_t + \epsilon_t$ and $E(\epsilon_t x_t) = 0$.

Then we have:

$$E(\hat{a}_1) = a_1 - a_1 \frac{\sigma^2_\epsilon}{\sigma^2_{x^\circ}} = a_1 \left(1 - \frac{\sigma^2_\epsilon}{\sigma^2_{x^\circ}}\right).$$

Since $\sigma^2_{x^\circ} = \sigma^2_x + \sigma^2_\epsilon$, $0 \leq \sigma^2_\epsilon / \sigma^2_{x^\circ} \leq 1$. Therefore, we conclude that a random error in measurement in the independent variable biases the coefficient toward zero.

Note that measurement error can be analyzed as a special case of specification bias analysis, where we can say something definite about the auxiliary regression coefficient.

2. Error in Measurement : Expectations Case

Suppose wage rates, w, are a function of the current productivity level, y, and the expected price level, x:

$$w_t = \alpha y_t + \beta x_t + u_t$$

We do not observe the expected price level, x_t, but only the actual level, x_t^o. Suppose:

$$x_t^o = x_t + v_t \qquad v_t \sim N(0, \sigma_v^2)$$

$$E(v_t x_t) = 0$$

i.e. that expectations are on average correct but there is a random noise component, v_t, which determines the actual level of x_t, x_t^o. In terms of observables, the true model is:

$$w_t = \alpha y_t + \beta (x_t^o - v_t) + u_t$$

i.e.

$$w_t = \alpha y_t + \beta x_t^o - \beta v_t + u_t$$

Hence, the error in measurement is like an omitted variable, v_t. We analyze as before:

$$E(\hat{\alpha}) = \alpha - \beta \rho_{vy}$$

$$E(\hat{\beta}) = \beta - \beta \rho_{vx^\circ}$$

where

$$v_t = \rho_{vy} y_t + \rho_{vx^\circ} x_t^\circ + \xi_t$$

However, in this case we can get a more exact statement of ρ_{vx°. From

$$x_t^\circ = x_t + v_t$$

it follows:

$$\rho_{vx^\circ} = \frac{\sigma_{vx^\circ}}{\sigma_{x^\circ}^2} = \frac{\sigma_v^2}{\sigma_x^2 + \sigma_v^2}$$

Note $0 \leq \rho_{vx^\circ} \leq 1$

Hence

$$E(\hat{\beta}) = \beta (1 - \rho_{vx^\circ})$$

which we also write as:

$$E(\hat{\beta}) = \beta\left(1 - \frac{\sigma_v^2}{\sigma_{x^*}^2 + \sigma_v^2}\right)$$

or $$E(\hat{\beta}) = \beta\left(\frac{\sigma_x^2}{\sigma_{x^*}^2}\right) < \beta$$

The ratio $\sigma_v^2/\sigma_{x^*}^2$, is called the noise ratio. We conclude that an error in measurement biases the coefficient of the mismeasured variable toward zero.

A macroeconomic example: consumption function

Permanent Income Hypothesis

$$C^P_t = \alpha Y^P_t + U_t$$

However, we observe not permanent consumption and income but the sum of permanent and transitory movements. Define:

$$C^o_t = C^P_t + C^T_t$$

$$Y^o_t = Y^P_t + Y^T_t$$

and assume:

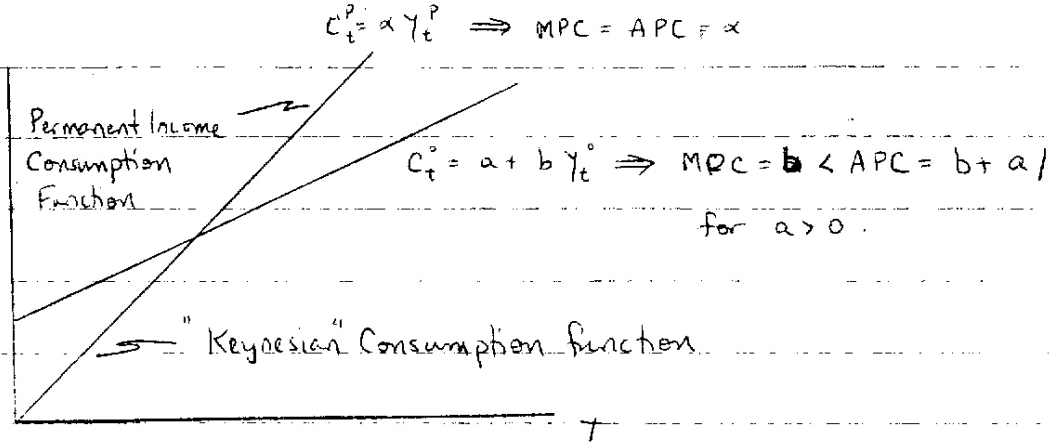
$$E(Y^T_t, Y^T_s) = 0 \qquad t \neq s$$

$$E(C^T_t, C^T_s) = 0 \qquad t \neq s$$

$$E(Y^T_t, Y^P_t) = E(C^T_t, C^P_t) = E(Y^T_t, C^T_t) = 0$$

$$E(Y^P_t, U_t) = 0$$

$$E(C^T_t) = E(Y^T_t) = 0$$


C
$C_t^P = \alpha Y_t^P \Rightarrow MPC = APC = \alpha$
Permanent Income Consumption Function
$C_t^\circ = a + b Y_t^\circ \Rightarrow MPC = b < APC = b + a/$
for $a > 0$.
"Keynesian" Consumption function
Y

In terms of observables, the model is

$$C_t^o - C_t^T = \alpha (Y_t^o - Y_t^T) + U_t$$

or $$C_t^o = \alpha Y_t^o + (U_t - \alpha Y_t^T + C_t^T)$$

The standard approach is to estimate:

$$C_t^o = a + b Y_t^o + \varepsilon_t$$

The OLS estimator of b is:

$$\hat{b} = \frac{\Sigma (C_t^o - \bar{C}^o)(Y_t^o - \bar{Y}_t^o)}{\Sigma (Y_t^o - \bar{Y}^o)^2}$$

Substituting for C_t^o from the true model:

$$\hat{b} = \frac{\Sigma [\alpha (Y_t^o - \bar{Y}^o) + (U_t - \alpha Y_t^T + C_t^T)](Y_t^o - \bar{Y}^o)}{\Sigma (Y_t^o - \bar{Y}^o)^2}$$

$$= \alpha + \frac{\Sigma (C_t^T - \alpha Y_t^T)(Y_t^o - \bar{Y}^o)}{\Sigma (Y_t^o - \bar{Y}^o)^2}$$

But note $y_t^\circ - \bar{y}^\circ = y_t^T$

Hence

$$\sum c_t^T (y_t^\circ - \bar{y}^\circ) = 0$$

and $$\alpha \sum y_t^T (y_t^\circ - \bar{y}^\circ) = \alpha \sigma^2_{y^T}$$

whence

$$\hat{b} = \alpha \left(1 - \frac{\sigma^2_{y^T}}{\sigma^2_{y^\circ}}\right) < \alpha$$

Since $\hat{b} < \alpha$, we would also conclude that

$$\hat{a} > 0$$

Conclusion: Failure to consider transitory components leads to a downward bias in the estimate of the MPC and a spurious finding of a positive intercept.

Instrumental Variable (IV) Estimation

Consider an equation:

$$Y_t = b x_t^* + u_t \qquad u_t \text{ Gauss Markov.}$$

$$x_t = x_t^* + \epsilon_t \qquad E(\epsilon_t x_t^*) = E(\epsilon_t) = 0$$
$$E(\epsilon_t^2) = \sigma_\epsilon^2 .$$

Then:

$$Y_t = b x_t + (u_t - b\epsilon_t) = b x_t + v_t .$$

Note:

$$E(x_t v_t) = E[x_t(u_t - b\epsilon_t)] = E x_t u_t - b E x_t \epsilon_t$$
$$= -b E(x_t^* + \epsilon_t)\epsilon_t = -b\sigma_\epsilon^2$$

Then OLS estimation yields:

$$\hat{b}_{OLS} = \frac{\sum x_t Y_t}{\sum x_t^2} = b + \frac{\sum x_t u_t}{\sum x_t^2} - b\frac{\sum x_t \epsilon_t}{\sum x_t^2}$$

$$\text{plim } \hat{b}_{OLS} = b - b\frac{\sigma_\epsilon^2}{\sigma_x^2} = b\left(1 - \frac{\sigma_\epsilon^2}{\sigma_x^2}\right) < b .$$

How to derive a consistent estimator of b?

Define: <u>Instrumental variable</u> z_t

i) $E(z_t x_t^*) \neq 0$.

ii) $E(z_t \epsilon_t) = 0$.

Consider IV estimator

$$\hat{b}_{IV} = \frac{\sum y_t z_t}{\sum x_t z_t} = \frac{b \sum x_t z_t + \sum u_t z_t - b \sum \epsilon_t z_t}{\sum x_t z_t}$$

$$= b + \frac{\sum u_t z_t}{\sum x_t z_t} - b \frac{\sum \epsilon_t z_t}{\sum x_t z_t}$$

$$\text{plim } \hat{b}_{IV} = b .$$

<u>IV Estimator</u>: Covariance between y_t and z_t divided by covariance between x_t and z_t.

Another interpretation : Two Stage Least Squares (2SLS)

Consider two step procedure :

i) Regress x_t against z_t and take fitted value $\hat{x}_t$.

ii) Regress y_t against $\hat{x}_t$.

Proposition: This two stage estimator is equivalent to IV estimator

Proof.

$$x_t = A z_t + \xi_t$$

$$\hat{A} = \frac{\sigma_{xz}}{\sigma_z^2}$$

$$\hat{x}_t = \hat{A} z_t = \frac{\sigma_{xz}}{\sigma_z^2} z_t$$

Then take

$$y_t = b \hat{x}_t + v_t$$

$$\hat{b} = \frac{\sigma_{y\hat{x}}}{\sigma^2_{\hat{x}}} = \frac{\sigma(y,\ \sigma_{xz}/\sigma_z^2 \cdot z)}{\sigma^2(\sigma_{xz}/\sigma_z^2 \cdot z)}$$

$$= \frac{(\sigma_{xz}/\sigma_z^2)\,\sigma_{yz}}{(\sigma_{xz}^2/\sigma_z^2)\,\sigma_z^2} = \frac{\sigma_{yz}}{\sigma_{xz}} = \hat{b}_{IV}$$

qed.

Inclusion of Irrelevant Variables

Consider the ~~true~~ model

i) $Y_t = a_1 X_{1t} + a_2 X_{2t} + U_t$

and the estimated version

ii) $Y_t = a_1 X_{1t} + a_2 X_{2t} + a_3 X_{3t} + U_t$

Clearly, provided $E(U_t X_{3t}) = 0$, all estimators will be unbiased. But are they BLUE?

Consider the variances of say $\hat{a}_1$:

From (i) $\sigma^2_{\hat{a}_1} = \sigma^2_U / \Sigma \hat{v}_1^2$

where $\hat{v}_1$ is defined by

$$X_{1t} = \hat{c}_0 + \hat{c}_1 X_{2t} + \hat{v}_{1t}$$

From (ii) $\sigma^2_{\hat{a}_1} = \sigma^2_U / \Sigma \hat{w}_{1t}^2$

where $\hat{w}_1$ is defined by

$$X_{1t} = \hat{c}_0 + \hat{c}_1 X_{2t} + \hat{c}_2 X_{3t} + \hat{w}_{1t}$$

Clearly, provided $\hat{c}_2 \neq 0$, $\sum \hat{v}_{1t}^2 > \sum \hat{w}_{1t}^2$. Hence

$$\sigma^{2\,(i)}_{\hat{\alpha}_1} = \frac{\sigma_u^2}{\sum \hat{v}_{1t}^2} < \sigma^{2\,(ii)}_{\hat{\alpha}_1} = \frac{\sigma_u^2}{\sum \hat{w}_{1t}^2}$$

Conclusion: Including irrelevant variables in general leaves unbiasedness property intact. However, unless the included irrelevant variable is uncorrelated with true exogenous variables, the variances of estimated coefficients are increased.