To Be or Not to Be (in the Model)?

Consider the usual regression equation

$$Y = X\beta + \epsilon \tag{1}$$

with $E(\epsilon \mid X) = 0$ and $\mathrm{Cov}(\epsilon \mid X) = \sigma^2 I$. Suppose that you are trying to decide whether this model is correct; in particular, you are thinking of adding another variable, Z, which would give the model

$$Y = X\gamma + cZ + \delta \tag{2}$$

with $E(\delta \mid X, Z) = 0$ and $\mathrm{Cov}(\delta \mid X, Z) = \tau^2 I$. We also assume that the new design matrix $[X\ Z]$ has full rank.

You are concerned about omitted-variable bias in (1), but about endogeneity bias in (2). See SM 4.5.5, 11, 13, 14 and 5.9.11. Usually you would prefer a correct model to an incorrect one. However, what if both are correct? Recall that an optional exercise in the notes to chapter 4 showed that if (Y, X, Z) are jointly normal, then both (1) and (2) are correct. Here, we are considering only statistical correctness; causal correctness, the topic of chapter 6, is another story. But we discuss briefly how causal issues lead to uncertainty in choosing between (1) and (2), even when both are statistically correct. Suppose we want to know the effect, if any, of X on Y. This might be specified by (1). But should we control for Z, as might be specified by (2)? If X affects Y through Z, then we should not control for Z. But if Z is a confounding variable for X and Y, then we should. We may not know the answers to these questions, just as we may not know whether X causes Y. These questions go beyond the purely statistical issues here.

There are two ways that both (1) and (2) can be correct, so that neither has a bias, depending on whether $c = 0$ or $c \neq 0$. Here, we mean bias in estimating β for (1) and in estimating γ in (2); if $c = 0$, then $\beta = \gamma$. Although both estimates would be unbiased, their SEs may differ. To analyze the SEs, assume that indeed both (1) and (2) are correct. (Note that we are not saying that (2) is correct for two different values of c; the assumptions in (2) preclude that. Indeed, we have seen, for (1) say, that β is determined by the assumptions, namely, $X\beta = E(Y \mid X)$ and X has full rank, so $X\beta$ determines β.)

If $c = 0$, then $\gamma = \beta$, $\delta = \epsilon$, and $\tau = \sigma$. Also, $E(\epsilon \mid X, Z) = 0$. In this case, the SEs of each $\widehat{\beta}_k$ are smaller if we omit Z. Why should we expect something like this? Recall that $\hat{Y} = P_{\mathrm{col}\, X}(Y)$ for (1), so that given X, the random part of $\hat{Y}$ equals $P_{\mathrm{col}\, X}(\epsilon)$. Similarly, the random part for (2) equals $P_{\mathrm{col}[X\ \ Z]}(\delta) = P_{\mathrm{col}[X\ \ Z]}(\epsilon)$. Since the column space of X

and Z together is larger than the column space of X, we are taking more of the error for (2) than for (1). If ϵ is normal, then the fact that taking more of the error leads to more variability can be justified by our knowledge of orthogonal projections applied to normal vectors.

To see that the SEs of each $\widehat{\beta}_k$ are smaller if we omit Z in general, i.e., whether or not ϵ is normal, let $\widehat{\beta}^Z$ be the estimate when we do use Z in OLS. (We are still assuming that $c = 0$.) Let W_k be the space spanned by all the columns of X except the kth. Then

$$\mathrm{SE}(\widehat{\beta}_k \mid X) = \frac{\sigma}{\|P^{\perp}_{W_k} X^{[k]}\|}\,, \tag{3}$$

where we are taking the orthogonal projection of the kth column of X, $X^{[k]}$, to the orthocomplement of W_k, whereas

$$\mathrm{SE}(\widehat{\beta}^Z_k \mid X, Z) = \frac{\sigma}{\|P^{\perp}_{W_k, Z} X^{[k]}\|}\,.$$

Since the span of W_k and Z is larger than W_k, the projection to its orthocomplement is smaller. (See the handout on multicollinearity for these SE formulas.) One can also prove that $\mathrm{SE}(a'\widehat{\beta} \mid X) \le \mathrm{SE}(a'\widehat{\beta}^Z \mid X, Z)$ for every a, just as in the Gauss-Markov theorem. (But that theorem doesn't apply to prove this, since $\widehat{\beta}^Z$ uses Z, while the theorem allows only estimators that depend on Y and X, nothing from outside the model. Briefly, a proof is as follows: $a'X'Xa \le [a\ b]'[X\ Z]'[X\ Z][a\ b]$ for all a and b, whence $a'(X'X)^{-1}a \ge [a\ 0]'\big([X\ Z]'[X\ Z]\big)^{-1}[a\ 0]$ for all a. Another proof proceeds by adding and subtracting $a'X\widehat{\beta}$ in the fitted equation and then using orthogonal projection formulas for the SEs.)

Now assume that $c \neq 0$. In this case, $\epsilon = X(\gamma - \beta) + cZ + \delta$, so $E(Z \mid X) = X(\beta - \gamma)/c$. The SEs of $\hat{\gamma}$ could be larger or smaller than those of $\widehat{\beta}$. For example, we could have $Z = Y$, $c = 1$, $\gamma = 0$, and $\delta = 0$. We would have an exact fit and no error. For a less extreme example, suppose that $\epsilon = Z + \delta$, where $Z \sim \mathrm{N}(0, .99\sigma^2 I)$ and $\delta \sim \mathrm{N}(0, .01\sigma^2 I)$, with Z and δ independent of each other and of X. Then $\gamma = \beta$, $c = 1$, and (2) has almost no error compared to (1). For one more example, suppose that $\epsilon = \eta + \delta$, where $\eta \sim \mathrm{N}(0, .99\sigma^2 I)$ and $\delta \sim \mathrm{N}(0, .01\sigma^2 I)$, with η and δ independent of each other and of X. Now define $Z = X\beta/2 + \eta$. Then $\gamma = \beta/2$ and $c = 1$; again the error in (2) is very small compared to (1). On the other hand, if we interpolate between one of these examples and one where $c = 0$ (such as a Z independent of (X, ϵ)), then we can get an example with $c \neq 0$ but with larger SEs than for (1). So there is no general rule.

Note that in general,

$$\mathrm{SE}(\widehat{\beta}^Z_k \mid X, Z) = \frac{\tau}{\|P^{\perp}_{W_k, Z} X^{[k]}\|}\,. \tag{4}$$

Since $\tau^2 = \mathrm{Var}(Y_1 \mid X, Z) \le \mathrm{Var}(Y_1 \mid X) = \sigma^2$, the numerator of (4) is no larger than that of (3), while the denominator of (4) is also no larger than that of (3).

Thus, the comment on p. 251 for 4.5.14 that "putting another variable into the equation likely reduces the sampling error in the estimates" is not true when $c = 0$, but may be true otherwise. The same holds for 4.5.11 and 5.9.11.