Exercise SM4.5.5

Here's the exercise:

You are using OLS to fit a regression equation with the usual assumptions. True or false, and explain:
(a) If you exclude a variable from the equation, but the excluded variable is orthogonal to the other variables in the equation, you won't bias the estimated coefficients of the remaining variables.
(b) If you exclude a variable from the equation, and the excluded variable isn't orthogonal to the other variables, your estimates are likely to be biased.
(c) If you put an extra variable into the equation, you won't bias the estimated coefficients—as long as the error term remains independent of the explanatory variables.
(d) If you put an extra variable into the equation, you are likely to bias the estimated coefficients—if the error term is dependent on that extra variable.

We are assuming that the model $Y = X\beta + \epsilon$ with $\epsilon \perp\!\!\!\perp X$ and $E(\epsilon) = \mathbf{0}$ is correct. If it were not, then in order to consider bias in $\hat{\beta}$, we would use equation (8) or (9) in the book, which is the equation (N5) or the next one in the proof in the notes of theorem 4.2. Namely, we have

$$E(\hat{\beta} \mid X) = \beta + (X'X)^{-1}X'E(\epsilon \mid X)\,. \tag{9}$$

This shows that the bias is $\mathbf{0}$, i.e., $\hat{\beta}$ is unbiased given X, iff $(X'X)^{-1}X'E(\epsilon \mid X) = \mathbf{0}$. Now $(X'X)^{-1}$ is an invertible matrix; it maps a vector to $\mathbf{0}$ iff the vector is already $\mathbf{0}$. Thus, the bias is $\mathbf{0}$ iff $X'E(\epsilon \mid X) = \mathbf{0}$, i.e., $X \perp E(\epsilon \mid X)$. This could also be written as $E(X'\epsilon \mid X) = \mathbf{0}$. We will use this criterion to examine the changed assumptions in the exercise.

In all these considerations, when the assumptions of OLS do not hold, yet we want to estimate β, we ought to say what β is. After all, there are many ways to write $Y = X\beta + \epsilon$ if we don't say anything about ϵ. There are two answers to this: The one we consider here is that β is determined by another equation for Y where all OLS assumptions do hold. Another answer involves response schedules, the topic of chapter 6, in which β plays a causal role and is assumed to reflect a property of the real world.

If we exclude a variable, let's say it is the last one and call it U. Thus, write $X = [M\ U]$ and correspondingly write $\beta = \begin{bmatrix} \alpha \\ \beta_p \end{bmatrix}$. That is, $\alpha = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$. Since β is the true parameter vector, we are asking in parts (a) and (b) what the effect of excluding U is on the remaining parameters, i.e., on α. We know α is the truth (by our assumption on β), so we want to estimate α. We do not make any new statistical assumptions on the model.

Why do we exclude U? It might be that even though we'd like to include U, we don't have data on U. In practice, we may not know that X is the correct design matrix and instead think that M is correct. Or it might be that we prefer fewer variables (it might reduce sampling error; see the handout "To Be or Not To Be (in the Model)?") and hope that we don't need U. Or it might be that we don't want to control for U. However, in this last case, it could be that there is some γ such that $Y = M\gamma + \delta$ for a parameter vector γ and an error term δ that satisfy all the OLS assumptions. In this case, where we do not want to control for U, we really want to estimate γ, not α, so there is no issue of bias. Instead, the question is whether the model actually holds.

Now we have

$$Y = X\beta + \epsilon = M\alpha + \underbrace{U\beta_p + \epsilon}_{\text{new error}} .$$

The OLS estimate of α using the new design matrix M is

$$(M'M)^{-1}M'Y = (M'M)^{-1}M'(M\alpha + U\beta_p + \epsilon) = \alpha + (M'M)^{-1}M'U\beta_p + (M'M)^{-1}M'\epsilon .$$

Therefore, the bias is

$$\begin{aligned} E\big((M'M)^{-1}M'Y \mid M\big) &= (M'M)^{-1}E(M'U \mid M)\beta_p + (M'M)^{-1}M'E(\epsilon \mid M) \\ &= (M'M)^{-1}E(M'U \mid M)\beta_p \end{aligned}$$

since $E(\epsilon \mid M) = \mathbf{0}$ by assumption (it is zero even if we condition on all of X). Therefore the bias is zero if $\beta_p = 0$, which is saying that U is not really part of the model to begin with (although we might not know that.) But if $\beta_p \neq 0$, then we need to look more closely. In (a), we are assuming that $U \perp M$, so $M'U = \mathbf{0}$ and again the estimate is unbiased. In (b), we are not assuming this. So while it could happen that $M'U \neq \mathbf{0}$ yet $E(M'U \mid M) = \mathbf{0}$, this is unlikely and therefore it is likely that the estimate is biased.

Note that since by default, the new error term is $U\beta_p + \epsilon$ (by excluding U from the design matrix), we might need to treat U as random, depending on what assumptions we make on the original model and on the new model (unless $\beta_p = 0$).

Now suppose that instead of excluding a variable, we include a new one, call it Z. Call its coefficient c. Then we are considering the model

$$Y = X\gamma + cZ + \delta ,$$

but perhaps the assumption $E(\delta \mid X, Z) = \mathbf{0}$ doesn't hold. Recall that we are still assuming the original model is correct:

$$Y = X\beta + \epsilon$$

with $E(\epsilon \mid X) = \mathbf{0}$. Let $\hat{\gamma}$ be the OLS estimate of γ. We will look at whether $\hat{\gamma}$ is an unbiased estimate of β, i.e., whether $E(\hat{\gamma} \mid X, Z) = \beta$.

Why would we include Z and still want to estimate β, rather than γ? Again, it might be that we do not know in practice that X is the correct design matrix and we think it should be $[X\ Z]$. Perhaps we think we should control for Z when we should not. It might be that β has a causal interpretation (see chapter 6), whereas γ does not. Also simultaneous equations models (chapter 9) lead to such types of equations. It might be that by including Z we hope to reduce sampling error (again, see "To Be or Not to Be (in the Model)?").

Now there is an ambiguity in the statement of the exercise: does "the error term" refer to the original one, ϵ, or the new one, δ?

Suppose that $c = 0$. Then $\gamma = \beta$ and $\delta = \epsilon$, so both error terms are the same (again, we might not know whether $c = 0$, which might be why we are uncertain whether to include Z). In (c), we are assuming, therefore, that $\delta \perp\!\!\!\perp X, Z$ and $E(\delta) = E(\epsilon) = \mathbf{0}$, whence $E(\delta \mid X, Z) = \mathbf{0}$ holds, and this implies that the estimate is unbiased by theorem 4.2. In (d), however, we are assuming that $\delta \not\perp\!\!\!\perp Z$, so it is likely that $E(\delta \mid X, Z) \neq \mathbf{0}$, whence $\hat{\gamma}$ as an estimate of γ is likely biased. In fact, the bias, conditional on X and Z, is $\mathbf{0}$ iff $[X\ Z]' E(\delta \mid X, Z) = \mathbf{0}$. This is not likely to hold without some assumption. However, we are asking about using $\hat{\gamma}$ as an estimate of β. Now by (9) above, we have that

$$E(\hat{\gamma} \mid X, Z) = \gamma + \big([X\ Z]'[X\ Z]\big)^{-1}[X\ Z]' E(\delta \mid X, Z)\,.$$

Since $Y = X\hat{\gamma} + \hat{c}Z + d$ for some $d \perp X, Z$, we have

$$Y = X\beta + X(\hat{\gamma} - \beta) + \hat{c}X(X'X)^{-1}X'Z + e = X\beta + X\big(\hat{\gamma} - \beta + \hat{c}(X'X)^{-1}X'Z\big) + e\,,$$

where $e \perp X$. Thus, the bias in $\hat{\gamma}$ as an estimator of β is

$$\begin{aligned} E(\hat{\gamma} - \beta + \hat{c}(X'X)^{-1}X'Z \mid X, Z) &= E(\hat{\gamma} \mid X, Z) - \beta + E(\hat{c} \mid X, Z)(X'X)^{-1}X'Z \\ &= \gamma - \beta + E(\hat{c} \mid X, Z)(X'X)^{-1}X'Z + \big([X\ Z]'[X\ Z]\big)^{-1}[X\ Z]' E(\delta \mid X, Z)\,. \end{aligned}$$

Nothing leads us to think that this mess is $\mathbf{0}$.

The answer in the back of the book assumes that $c = 0$. This is perhaps because the new model does hold with $c = 0$ and $\delta = \epsilon$. However, it can be that $c \neq 0$ and still the model holds, though now with $\delta \neq \epsilon$. 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.

If $c \neq 0$, however, then we should interpret "error term" as the new error, δ. This is because it makes more sense to ask about the new error term when the model changes, as it does when $c \neq 0$.

Now in (c), theorem 4.2 applies by assumption, which means that $\hat{\gamma}$ is an unbiased estimator of γ. Thus, $\hat{\gamma}$ is an unbiased estimator of β iff $\gamma = \beta$. This is unlikely to be true since $c \neq 0$, so now the answer to (c) becomes no. In (d), the situation is even worse, with the bias written out as above for the analysis when $c = 0$, so the answer is still no.

One more aspect of (a) is worth noting: the estimates don't actually change—so of course they are still unbiased. In other words, we claim that $\hat{\alpha} = \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix}$. The reason is that

$$M\hat{\alpha} = P_{\mathrm{col}(M)}(Y) = P_{\mathrm{col}(M)}\big(P_{\mathrm{col}(X)}(Y)\big)$$

because $\mathrm{col}(M) \subset \mathrm{col}(X)$ (see the linear algebra review). Now

$$P_{\mathrm{col}(X)}(Y) = X\hat{\beta} = M \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix} + U\hat{\beta}_p \,.$$

The assumption that $U \perp M$ now gives that the orthogonal projection of this onto $\mathrm{col}(M)$ is $M \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix}$, so $M\hat{\alpha} = M \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix}$. Since M has full rank, our claim follows.