Chapter 9
Assessing Studies Based on Multiple Regression

■ Solutions to Exercises

1. As explained in the text, potential threats to external validity arise from differences between the population and setting studied and the population and setting of interest. The statistical results based on New York in the 1970's are likely to apply to Boston in the 1970's but not to Los Angeles in the 1970's. In 1970, New York and Boston had large and widely used public transportation systems. Attitudes about smoking were roughly the same in New York and Boston in the 1970s. In contrast, Los Angeles had a considerably smaller public transportation system in 1970. Most residents of Los Angeles relied on their cars to commute to work, school, and so forth.

   The results from New York in the 1970's are unlikely to apply to New York in 2002. Attitudes towards smoking changed significantly from 1970 to 2002.

2. (a) When Y_i is measured with error, we have $\tilde{Y}_i = Y_i + w_i$, or $Y_i = \tilde{Y}_i - w_i$. Substituting the 2nd equation into the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$ gives $\tilde{Y}_i - w_i = \beta_0 + \beta_1 X_i + u_i$, or $\tilde{Y}_i = \beta_0 + \beta_1 X_i + u_i + w_i$. Thus $v_i = u_i + w_i$.

   (b) (1) The error term v_i has conditional mean zero given X_i:

   $$E(v_i|X_i) = E(u_i + w_i|X_i) = E(u_i|X_i) + E(w_i|X_i) = 0 + 0 = 0.$$

   (2) $\tilde{Y}_i = Y_i + w_i$ is i.i.d since both Y_i and w_i are i.i.d. and mutually independent; X_i and $\tilde{Y}_j (i \neq j)$ are independent since X_i is independent of both Y_j and w_j. Thus, $(X_i, \tilde{Y}_i)$, $i = 1, \ldots, n$ are i.i.d. draws from their joint distribution.

   (3) $v_i = u_i + w_i$ has a finite fourth moment given that both u_i and w_i have finite fourth moments and are mutually independent. So (X_i, v_i) have nonzero finite fourth moments.

   (c) The OLS estimators are consistent because the least squares assumptions hold.

   (d) Because of the validity of the least squares assumptions, we can construct the confidence intervals in the usual way.

   (e) The answer here is the economists' "On the one hand, and on the other hand." On the one hand, the statement is true: i.i.d. measurement error in X means that the OLS estimators are inconsistent and inferences based on OLS are invalid. OLS estimators are consistent and OLS inference is valid when Y has i.i.d. measurement error. On the other hand, even if the measurement error in Y is i.i.d. and independent of Y_i and X_i, it increases the variance of the regression error ($\sigma_v^2 = \sigma_u^2 + \sigma_w^2$), and this will increase the variance of the OLS estimators. Also, measurement error that is not i.i.d. may change these results, although this would need to be studied on a case-by-case basis.

3. The key is that the selected sample contains only employed women. Consider two women, Beth and Julie. Beth has no children; Julie has one child. Beth and Julie are otherwise identical. Both can earn $25,000 per year in the labor market. Each must compare the $25,000 benefit to the costs of working. For Beth, the cost of working is forgone leisure. For Julie, it is forgone leisure and the costs (pecuniary and other) of child care. If Beth is just on the margin between working in the labor market or not, then Julie, who has a higher opportunity cost, will decide not to work in the labor market. Instead, Julie will work in "home production," caring for children, and so forth. Thus, on average, women with children who decide to work are women who earn higher wages in the labor market.

4.

			Estimated Effect of a 10% Increase in Average Income	
State	$\beta_{\ln(Income)}$	**Std. Dev. of Scores**	**In Points**	**In Std. Dev.**
Calif.	11.57 (1.81)	19.1	1.157 (0.18)	0.06 (0.001)
Mass.	16.53 (3.15)	15.1	1.65 (0.31)	0.11 (0.021)

The income effect in Massachusetts is roughly twice as large as the effect in California.

5 (a) $$Q = \frac{\gamma_1\beta_0 - \gamma_0\beta_1}{\gamma_1 - \beta_1} + \frac{\gamma_1 u - \beta_1 v}{\gamma_1 - \beta_1}.$$

and $$P = \frac{\beta_0 - \gamma_0}{\gamma_1 - \beta_1} + \frac{u - v}{\gamma_1 - \beta_1}.$$

(b) $$E(Q) = \frac{\gamma_1\beta_0 - \gamma_0\beta_1}{\gamma_1 - \beta_1}, \quad E(P) = \frac{\beta_0 - \gamma_0}{\gamma_1 - \beta_1}$$

(c)

$$Var(Q) = \left(\frac{1}{\gamma_1 - \beta_1}\right)^2 (\gamma_1^2\sigma_u^2 + \beta_1^2\sigma_v^2), \; Var(P) = \left(\frac{1}{\gamma_1 - \beta_1}\right)^2 (\sigma_u^2 + \sigma_v^2), \text{ and}$$

$$Cov(P, Q) = \left(\frac{1}{\gamma_1 - \beta_1}\right)^2 (\gamma_1\sigma_u^2 + \beta_1\sigma_V^2)$$

(d) (i) $$\hat{\beta}_1 \xrightarrow{p} \frac{Cov(Q, P)}{Var(P)} = \frac{\gamma_1\sigma_u^2 + \beta_1\sigma_V^2}{\sigma_u^2 + \sigma_V^2}, \quad \hat{\beta}_0 \xrightarrow{p} E(Q) - E(P)\frac{Cov(P, Q)}{Var(P)}$$

(ii) $\hat{\beta}_1 - \beta_1 \xrightarrow{p} \frac{\sigma_u^2(\gamma_1 - \beta_1)}{\sigma_u^2 + \sigma_V^2} > 0,$ using the fact that $\gamma_1 > 0$ (supply curves slope up) and $\beta_1 < 0$ (demand curves slope down).

6. (a) The parameter estimates do not change. Nor does the the R^2. The sum of squared residuals from the 100 observation regression is $SER_{200} = (100-2)\times 15.1^2 = 22,344.98,$ and the sum of squared residuals from the 200 observation regression is twice this value: $SSR_{200} = 2\times 22,344.98.$ Thus, the *SER* from the 200 observation regression is $SER_{200} = \sqrt{\frac{1}{200-2} SSR_{200}} = 15.02.$ The standard errors for the regression coefficients are now computed using equation (5.4) where $\sum_{i=1}^{200}(X_i - \overline{X})^2 \hat{u}_i^2$ and $\sum_{i=1}^{200}(X_i - \overline{X})^2$ are twice their value from the 100 observation regression. Thus the standard errors for the 200 observation regression are the standard errors in the 100 observation regression multiplied by $\sqrt{\frac{100-2}{200-2}} = 0.704.$ In summary, the results for the 200 observation regression are

$$\hat{Y} = 32.1 + 66.8X,\ SER = 15.02,\ R^2 = 0.81$$
$$(10.63)\ (8.59)$$

   (b) The observations are not *i.i.d.*: half of the observations are identical to the other half, so that the observations are not *independent*.

7. (a) True. Correlation between regressors and error terms means that the OLS estimator is inconsistent.
   (b) True.

8. No, for two reasons. First, test scores in California and Massachusetts are for different tests and have different means and variances. [However, converting (9.5) into units for Massachusetts yields the implied regression to *TestScore*(*MA units*) = 740.9 − 1.80 × *STR*, which is similar to the regression using Massachusetts data shown in Column 1 of Table 9.2.] Second, the regression in Column 1 of Table 9.2 has a low R^2 suggesting that it will not provide a accurate forecast of test scores.

9. Both regressions suffer from omitted variable bias so that they will not provide reliable estimates of the causal effect of income on test scores. However, the nonlinear regression in (8.18) fits the data well, so that it could be used for forecasting.

10. There are several reasons for concern. Here are a few.

    Internal consistency: omitted variable bias as explained in the last paragraph of the box.

    Internal consistency: sample selection may be a problem as the sample used are full-time workers.

    External consistency: Returns to education may change over time because of the relative demands and supplies of skilled and unskilled workers in the economy. To the extent that this is important, the results shown in the box (based on 2004 data) may not accurately estimate the returns to education in 2008.

11. Again, there are reasons for concern. Here are a few.

    Internal consistency: To the extent that price is affected by demand, there may be simultaneous equation bias.

    External consistency: The internet and introduction of "*E*-journals" may induce important changes in the market for academic journals so that the results for 2000 may not be relevant in 2008.