Introduction to Econometrics (4^{th} Edition)

by

James H. Stock and Mark W. Watson

Solutions to Odd-Numbered End-of-Chapter Exercises: Chapter 9

(This version September 18, 2018)

9.1. 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 2018. Attitudes towards smoking changed significantly from 1970 to 2018.

9.3. 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.

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

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

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

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

(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).

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

(b) True.

9.9. 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.

9.11. There are several 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 for today's market.

9.13. (a) $\hat{\beta}_1 = \frac{\sum_{i=1}^{300}(\tilde{X}_i - \bar{\tilde{X}})(Y_i - \bar{Y})}{\sum_{i=1}^{300}(\tilde{X}_i - \bar{\tilde{X}})^2}$. Because all of the X_i's are used (although some are used for the wrong values of Y_j), $\bar{\tilde{X}} = \bar{X}$, and $\sum_{i=1}^{n}(X_i - \bar{X})^2$. Also, $Y_i - \bar{Y} = \beta_1(X_i - \bar{X}) + u_i - \bar{u}$. Using these expressions:

$$\hat{\ }_1 = {}_1\frac{\sum_{i=1}^{0.8n}(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + {}_1\frac{\sum_{i=0.8n+1}^{n}(\tilde{X}_i - \bar{X})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \frac{\sum_{i=1}^{n}(\tilde{X}_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$= {}_1\frac{\frac{1}{n}\sum_{i=1}^{0.8n}(X_i - \bar{X})^2}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} + {}_1\frac{\frac{1}{n}\sum_{i=0.8n+1}^{n}(\tilde{X}_i - \bar{X})(X_i - \bar{X})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} + \frac{\frac{1}{n}\sum_{i=1}^{n}(\tilde{X}_i - \bar{X})(u_i - \bar{u})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

where n = 300, and the last equality uses an ordering of the observations so that the first 240 observations (= 0.8×n) correspond to the correctly measured observations ($\tilde{X}_i$ = X$_i$).

As is done elsewhere in the book, we interpret n = 300 as a large sample, so we use the approximation of n tending to infinity. The solution provided here thus shows that these expressions are approximately true for n large and hold in the limit that n tends to infinity. Each of the averages in the expression for $\hat{\beta}_1$ have the following probability limits:

$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \xrightarrow{p} \sigma_X^2$,

$\frac{1}{n}\sum_{i=1}^{0.8n}(X_i - \bar{X})^2 \xrightarrow{p} 0.8\sigma_X^2$,

$\frac{1}{n}\sum_{i=1}^{n}(\tilde{X}_i - \bar{X})(u_i - \bar{u}) \xrightarrow{p} 0$, and

$\frac{1}{n}\sum_{i=0.8n+1}^{n}(\tilde{X}_i - \bar{X})(X_i - \bar{X}) \xrightarrow{p} 0$,

(continued on next page)

9.13 (continued)

where the last result follows because $\tilde{X}_i \neq X_i$ for the scrambled observations and X_j is independent of X_i for $i \neq j$. Taken together, these results imply $\hat{\beta}_1 \xrightarrow{p} 0.8\beta_1$.

(b) Because $\hat{\beta}_1 \xrightarrow{p} 0.8\beta_1$, $\hat{\beta}_1 / 0.8 \xrightarrow{p} \beta_1$, so a consistent estimator of β_1 is the OLS estimator divided by 0.8.

(c) Yes, the estimator based on the first 240 observations is better than the adjusted estimator from part (b). Equation (4.21) in Key Concept 4.4 (page 129) implies that the estimator based on the first 240 observations has a variance that is

$$\text{var}(\hat{\beta}_1(240obs)) = \frac{1}{240}\frac{\text{var}\left[(X_i - \mu_X)u_i\right]}{\left[\text{var}(X_i)\right]^2}.$$

From part (a), the OLS estimator based on all of the observations has two sources of sampling error. The first is $\frac{\sum_{i=1}^{300}(\tilde{X}_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{300}(X_i - \bar{X})^2}$ which is the usual source that comes from the omitted factors (u). The second is $\beta_1\frac{\sum_{i=241}^{300}(\tilde{X}_i - \bar{X})(X_i - \bar{X})}{\sum_{i=1}^{300}(X_i - \bar{X})^2}$, which is the source that comes from scrambling the data. These two terms are uncorrelated in large samples, and their respective large-sample variances are:

$$\text{var}\left(\frac{\sum_{i=1}^{300}(\tilde{X}_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{300}(X_i - \bar{X})^2}\right) = \frac{1}{300}\frac{\text{var}\left[(X_i - \mu_X)u_i\right]}{\left[\text{var}(X_i)\right]^2}.$$

and

$$\text{var}\left(\beta_1\frac{\sum_{i=241}^{300}(\tilde{X}_i - \bar{X})(X_i - \bar{X})}{\sum_{i=1}^{300}(X_i - \bar{X})^2}\right) = \beta_1^2\frac{0.2}{300}.$$

(continued on next page)

9.13 (continued)

Thus

$$\operatorname{var}\left(\frac{\hat{\beta}_1(300obs)}{0.8}\right)=\frac{1}{0.64}\left[\frac{1}{300}\frac{\operatorname{var}\left[(X_i-\mu_X)u_i\right]}{\left[\operatorname{var}(X_i)\right]^2}+\beta_1^2\frac{0.2}{300}\right]$$

which is larger than the variance of the estimator that only uses the first 240 observations.