Chapter 10
Regression with Panel Data

■ Solutions to Exercises

1. (a) With a $1 increase in the beer tax, the expected number of lives that would be saved is 0.45 per 10,000 people. Since New Jersey has a population of 8.1 million, the expected number of lives saved is $0.45 \times 810 = 364.5$. The 95% confidence interval is $(0.45 \pm 1.96 \times 0.22) \times 810 = [15.228, 713.77]$.

   (b) When New Jersey lowers its drinking age from 21 to 18, the expected fatality rate increases by 0.028 deaths per 10,000. The 95% confidence interval for the change in death rate is $0.028 \pm 1.96 \times 0.066 = [-0.1014, 0.1574]$. With a population of 8.1 million, the number of fatalities will increase by $0.028 \times 810 = 22.68$ with a 95% confidence interval $[-0.1014, 0.1574] \times 810 = [-82.134, 127.49]$.

   (c) When real income per capita in new Jersey increases by 1%, the expected fatality rate increases by 1.81 deaths per 10,000. The 90% confidence interval for the change in death rate is $1.81 \pm 1.64 \times 0.47 = [1.04, 2.58]$. With a population of 8.1 million, the number of fatalities will increase by $1.81 \times 810 = 1466.1$ with a 90% confidence interval $[1.04, 2.58] \times 810 = [840, 2092]$.

   (d) The low p-value (or high F-statistic) associated with the F-test on the assumption that time effects are zero suggests that the time effects should be included in the regression.

   (e) The difference in the significance levels arises primarily because the estimated coefficient is higher in (5) than in (4). However, (5) leaves out two variables (unemployment rate and real income per capita) that are statistically significant. Thus, the estimated coefficient on *Beer Tax* in (5) may suffer from omitted variable bias. The results from (4) seem more reliable. In general, statistical significance should be used to measure reliability only if the regression is well-specified (no important omitted variable bias, correct functional form, no simultaneous causality or selection bias, and so forth.)

   (f) Define a binary variable *west* which equals 1 for the western states and 0 for the other states. Include the interaction term between the binary variable *west* and the unemployment rate, *west* × (unemployment rate), in the regression equation corresponding to column (4). Suppose the coefficient associated with unemployment rate is β, and the coefficient associated with west × (unemployment rate) is γ. Then β captures the effect of the unemployment rate in the eastern states, and $\beta + \gamma$ captures the effect of the unemployment rate in the western states. The difference in the effect of the unemployment rate in the western and eastern states is γ. Using the coefficient estimate $(\hat{\gamma})$ and the standard error $\text{SE}(\hat{\gamma})$, you can calculate the t-statistic to test whether γ is statistically significant at a given significance level.

2. (a) For each observation, there is one and only one binary regressor equal to one. That is, $D1_i + D2_i + D3_i = 1 = X_{0,it}$.

   (b) For each observation, there is one and only one binary regressor that equals 1. That is, $D1_i + D2_i + \cdots + Dn_i = 1 = X_{0,it}$.

(c) The inclusion of all the binary regressors and the “constant” regressor causes perfect multicollinearity. The constant regressor is a perfect linear function of the *n* binary regressors. OLS estimators cannot be computed in this case. Your computer program should print out a message to this effect. (Different programs print different messages for this problem. Why not try this, and see what your program says?)

3. The five potential threats to the internal validity of a regression study are: omitted variables, misspecification of the functional form, imprecise measurement of the independent variables, sample selection, and simultaneous causality. You should think about these threats one-by-one. Are there important omitted variables that affect traffic fatalities and that may be correlated with the other variables included in the regression? The most obvious candidates are the safety of roads, weather, and so forth. These variables are essentially constant over the sample period, so their effect is captured by the state fixed effects. You may think of something that we missed. Since most of the variables are binary variables, the largest functional form choice involves the *Beer Tax* variable.
A linear specification is used in the text, which seems generally consistent with the data in Figure 8.2. To check the reliability of the linear specification, it would be useful to consider a log specification or a quadratic. Measurement error does not appear to a problem, as variables like traffic fatalities and taxes are accurately measured. Similarly, sample selection is a not a problem because data were used from all of the states. Simultaneous causality could be a potential problem. That is, states with high fatality rates might decide to increase taxes to reduce consumption. Expert knowledge is required to determine if this is a problem.

4. (a) slope = β_1, intercept = β_0
(b) slope = β_1, intercept = β_0
(c) slope = β_1, intercept = $\beta_0 + \gamma_3$
(d) slope = β_1, intercept = $\beta_0 + \gamma_3$

5. Let $D2_i = 1$ if $i = 2$ and 0 otherwise; $D3_i = 1$ if $i = 3$ and 0 otherwise … $Dn_i = 1$ if $i = n$ and 0 otherwise. Let $B2_t = 1$ if $t = 2$ and 0 otherwise; $B3_t = 1$ if $t = 3$ and 0 otherwise … $BT_t = 1$ if $t = T$ and 0 otherwise. Let $\beta_0 = \alpha_1 + \mu_1$; $\gamma_i = \alpha_i - \alpha_1$ and $\delta_t = \mu_t - \mu_1$.

6. $\tilde{v}_{it} = \tilde{X}_{it}u_{it}$. First note that $E(\tilde{v}_{it}) = E(\tilde{X}_{it}u_{it}) = E[\tilde{X}_{it}E(u_{it}|\tilde{X}_{it})] = 0$ from assumption 1. Thus, $\text{cov}(\tilde{v}_{it}\tilde{v}_{is}) = E(\tilde{v}_{it}\tilde{v}_{is}) = E(\tilde{X}_{it}\tilde{X}_{is}u_{it}u_{is}) = E(\tilde{X}_{it}u_{it})E(\tilde{X}_{is}u_{is})$, where the last equality follows because $(u_{it}, \tilde{X}_{it})$ is independent of $(u_{is}, \tilde{X}_{is})$ from assumption (2). The result then follows from $E(\tilde{X}_{it}u_{it}) = 0$.

7. (a) Average snow fall does not vary over time, and thus will be perfectly collinear with the state fixed effect.
(b) $Snow_{it}$ does vary with time, and so this method can be used along with state fixed effects.

8. There are several ways. Here is one: let $Y_{it} = \beta_0 + \beta_1 X_{1,it} + \beta_2 t + \gamma_2 \text{D}2_i + \cdots + \gamma_n \text{D}n_i + \delta_2(\text{D}2_i \times t) + \cdots + \delta_n(\text{D}n_i \times t) + u_{it}$, where $\text{D}2_i = 1$ if $i = 2$ and 0 otherwise and so forth. The coefficient $\lambda_i = \beta_2 + \delta_i$.

9. This assumption is necessary for the usual formula for SEs to be correct. If it is incorrect, errors are correlated, the usual formula for SEs are wrong and inference is faulty. The appendix includes a discussion of more general formulae for the SEs when Assumption #5 is violated.

10. (a) $\hat{\alpha}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$ which has variance $\frac{\sigma_u^2}{T}$. Because T is not growing, the variance is not getting small. $\hat{\alpha}_i$ is not consistent.
(b) The average in (a) is computed over T observations. In this case T is small ($T = 4$), so the normal approximation from the CLT is not likely to be very good.

11. No, one of the regressors is Y_{it-1}. This depends on Y_{it-1}. This means that assumption (1) is violated.