Financial Econometrics
Shangai Jiao Tong University
Summer School – 2025-2026
Problem Set - Lecture 6
With Solutions

Fatih Kansoy

Reference: Stock, J.H. and Watson, M.W., Introduction to Econometrics (4th Edition, Global Edition)

Chapter 10: Regression with Panel Data

Question 1. This exercise refers to the drunk driving panel data regression summarized in Table 10.1 of the textbook.

(a) New Jersey has a population of 8.85 million people. Suppose New Jersey increased the tax on a case of beer by \$2 (in 1988 dollars). Use the results in column (5) to predict the number of lives that would be saved over the next year. Construct a 99% confidence interval for your answer.

(b) The drinking age in New Jersey is 21. Suppose that New Jersey lowered its drinking age to 19. Use the results in column (5) to predict the change in the number of traffic fatalities in the next year. Construct a 95% confidence interval for your answer.

(c) Suppose real income per capita in New Jersey increases by 3% in the next year. Use the results in column (6) to predict the change in the number of traffic fatalities in the next year. Construct a 95% confidence interval for your answer.

(d) How should standard errors be clustered in the regressions in columns (2) through (7)?

(e) How should minimum drinking age be included in the regressions? Should it enter as a continuous variable or as a series of indicator variables?

Solution:

Note: Assume from Table 10.1 that the coefficient on Beer Tax in column (5) is -0.69 *with* $SE = 0.35$*, the coefficient on Minimum Legal Drinking Age is* 0.076 *with* $SE = 0.068$*, and the coefficient on ln(Income per capita) in column (6) is* 1.79 *with* $SE = 0.64$*. The dependent variable is traffic fatality rate per 10,000 people.*

(a) **Effect of $2 beer tax increase:**

Predicted change in fatality rate: $\Delta \hat{FR} = -0.69 \times 2 = -1.38$ per 10,000

With population 8.85 million = 885 units of 10,000:

Lives saved: $1.38 \times 885 = 1,221$

Standard error: $SE = 2 \times 0.35 \times 885 = 619.5$

99% CI: $1221 \pm 2.576 \times 619.5 = 1221 \pm 1596$

$$\boxed{[-375, 2817] \text{ lives saved}}$$

Note: The wide interval (including negative values) reflects substantial uncertainty.

(b) **Effect of lowering drinking age from 21 to 19:**

Change in fatality rate: $\Delta \hat{FR} = 0.076 \times (21 - 19) = 0.076 \times 2 = 0.152$ per 10,000

Additional fatalities: $0.152 \times 885 = 134.5 \approx 135$

SE: $2 \times 0.068 \times 885 = 120.4$

95% CI: $135 \pm 1.96 \times 120 = 135 \pm 235$

$$\boxed{[-100, 370] \text{ additional fatalities}}$$

(c) **Effect of 3% income increase:**

Since the model uses ln(Income), a 3% increase means $\Delta \ln(\text{Income}) \approx 0.03$.

Change in fatality rate: $1.79 \times 0.03 = 0.0537$ per 10,000

Additional fatalities: $0.0537 \times 885 = 47.5$

SE: $0.03 \times 0.64 \times 885 = 17.0$

95% CI: $47.5 \pm 1.96 \times 17 = 47.5 \pm 33.3$

$$\boxed{[14, 81] \text{ additional fatalities}}$$

(d) **Clustering:**

Standard errors should be **clustered by state**.

Rationale:

- Observations within the same state are likely correlated over time (autocorrelation).
- State-specific factors create within-cluster correlation.
- Clustering accounts for this serial correlation and provides valid inference.

(e) **Specification of drinking age:**

Continuous variable: Appropriate if we believe the effect is linear (each year of increase has the same proportional effect).

Indicator variables: Better if the effect is nonlinear (e.g., the jump from 20 to 21 matters more than from 18 to 19).

In practice: Since the continuous coefficient is insignificant ($t = 0.076/0.068 = 1.12$) and a joint F-test on age indicators is also insignificant, drinking age appears to have limited predictive power in this specification.

Question 2. Consider the binary variable version of the fixed effects model in Equation (10.11) except with an additional regressor, D_{1i}; that is, let:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_1 D_{1i} + \gamma_2 D_{2i} + \cdots + \gamma_n D_{ni} + u_{it}$$

(a) Suppose that $n = 3$. Show that the binary regressors and the "constant" regressor are perfectly multicollinear; that is, express one of the variables D_{1i}, D_{2i}, D_{3i}, and $X_{0,it}$ as a perfect linear function of the others, where $X_{0,it} = 1$ for all i, t.

(b) Show the result in (a) for general n.

(c) What will happen if you try to estimate the coefficients of the regression by OLS?

Solution:

(a) **Perfect multicollinearity with $n = 3$:**

The dummy variables are defined as:

$$D_{1i} = \begin{cases} 1 & \text{if } i = 1 \\ 0 & \text{otherwise} \end{cases}$$
$$D_{2i} = \begin{cases} 1 & \text{if } i = 2 \\ 0 & \text{otherwise} \end{cases}$$
$$D_{3i} = \begin{cases} 1 & \text{if } i = 3 \\ 0 & \text{otherwise} \end{cases}$$

For each observation, exactly one dummy equals 1 and the others equal 0.

Therefore:

$$\boxed{D_{1i} + D_{2i} + D_{3i} = 1 = X_{0,it}}$$

This is the **dummy variable trap**: the sum of all entity dummies equals the constant.

(b) **General case with n entities:**

For any observation (i, t):

$$\sum_{j=1}^{n} D_{ji} = 1$$

because exactly one of the n dummies equals 1 for each entity.

Therefore:

$$\boxed{D_{1i} + D_{2i} + \cdots + D_{ni} = X_{0,it} = 1}$$

The constant term is a **perfect linear combination** of the entity dummies.

(c) **Consequence for OLS:**

With perfect multicollinearity, the **OLS estimator does not exist**.

Mathematical reason:

- The matrix $(X'X)$ is singular (not invertible).
- $\hat{\beta} = (X'X)^{-1}X'Y$ cannot be computed.

What happens in practice:

- Most software will either: (1) drop one of the collinear variables automatically, or (2) report an error message.
- The solution is to omit either the constant OR one of the dummy variables.

Standard approach:

Omit D_{1i} (or any one dummy). Then γ_j for $j \geq 2$ measures the effect of being entity j relative to entity 1 (the reference category).