Panel Data

Dr. Patrick Toche

Textbook:

James H. Stock and Mark W. Watson, *Introduction to Econometrics*, 4th Edition, Pearson.

Other references:

Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach*, 7th Edition, Cengage Learning.
Joshua D. Angrist and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, 1st Edition, Princeton University Press.

The textbook comes with online resources and study guides. Other references will be given from time to time.

In this lesson you will learn ...

- Panel data with two time periods: before/after.
- Entity fixed effects regression.
- Time fixed effects regression.
- Assumptions for causal inference with fixed effects.

Panel Data

- **Panel data:**
  aka longitudinal data — n different entities observed at T different time periods: Y_{it} denotes the variable Y observed for the ith of n entities in the tth of T periods.
- **Balanced panel:**
  Variables are observed for each entity and each time period.
- **Unbalanced panel:**
  Some missing data for at least one time period for at least one entity.
- **Fixed effects panel regression:**
  Controls for omitted variables that vary across entities, but do not change over time.
- **Time fixed effects panel regression:**
  Controls for omitted variables that are constant across entities, but change over time.

Panel Data: Traffic Deaths and Alcohol Taxes

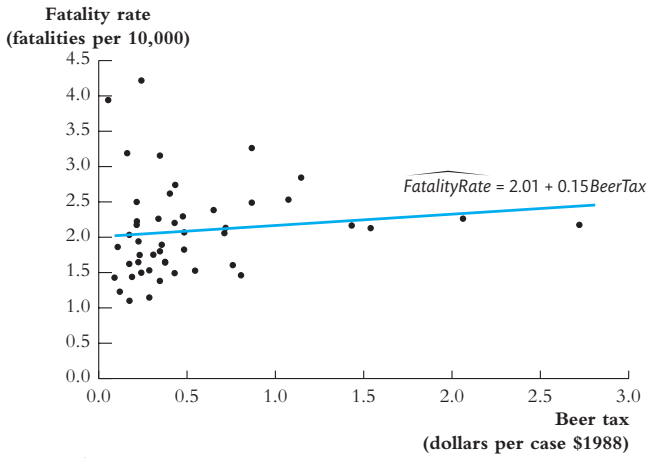
Traffic Deaths and Alcohol Taxes

$$\text{Year 1982:} \quad \widehat{FatalityRate} = \underset{}{2.01} + \underset{(0.15)}{0.15}\ BeerTax \quad \underset{(0.13)}{}$$

$$\text{Year 1988:} \quad \widehat{FatalityRate} = \underset{}{1.86} + \underset{(0.11)}{0.44}\ BeerTax \quad \underset{(0.13)}{}$$

- The coefficient on the real beer tax is statistically significant at the 1% level in 1988, but not so at the 10% level in 1982. The estimated coefficients for the 1982 and the 1988 data are positive! Are higher real beer taxes associated with more traffic fatalities? No! These regressions suffer from omitted variable bias.
- **Dealing with omitted variables:** Factors that affect the fatality rate: cars' safety features, whether state highways are in good repair, density of cars on the road, laws and police controls. Some of these omitted variables would be difficult to measure.
- **These factors can be held constant even when they cannot be measured or observed.**

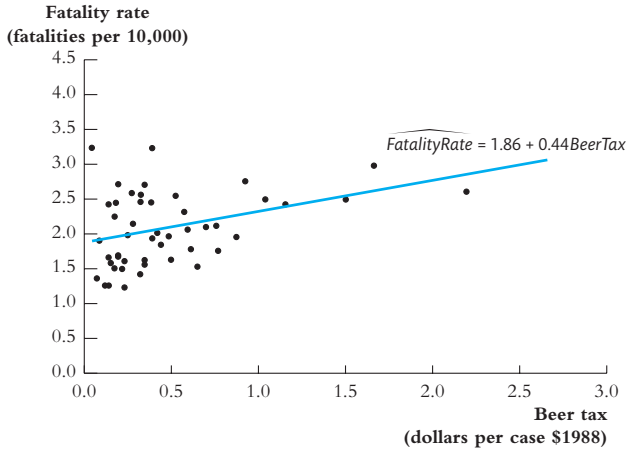
Traffic Deaths and Alcohol Taxes: 1982



(a) 1982 data

Traffic fatality rates and the real tax on a case of beer for 48 states in 1982 (in 1988 dollars).

Traffic Deaths and Alcohol Taxes: 1988



(b) 1988 data

Traffic fatality rates and the real tax on a case of beer for 48 states in 1988 (in 1988 dollars).

Before / After Comparisons

Eliminating Fixed Effects

- Let Z_i be a variable that determines the fatality rate in the ith state but does not change over time: Note the absence of the time subscript in Z_i, meaning $Z_{i,1982} = Z_{i,1988}$.

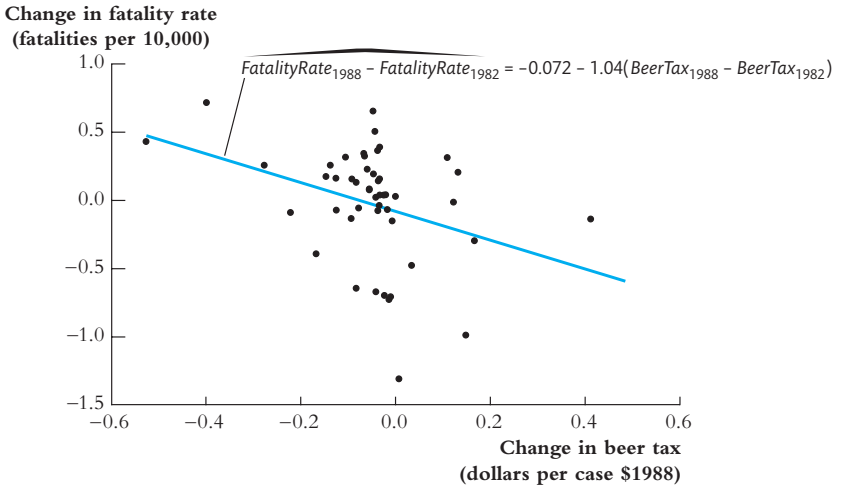
$$FatalityRate_{i,1982} = \beta_0 + \beta_1 BeerTax_{i,1982} + \beta_2 Z_i + u_{i,1982}$$
$$FatalityRate_{i,1988} = \beta_0 + \beta_1 BeerTax_{i,1988} + \beta_2 Z_i + u_{i,1988}$$

- Because Z_i does not change over time, it cannot be the cause of any change in the fatality rate between 1982 and 1988.
- **The influence of Z_i can be eliminated by analyzing the change in the fatality rate between the two periods.** Taking the difference between the above two equations:

$$\begin{aligned} &FatalityRate_{i,1988} - FatalityRate_{i,1982} \\ &= \left(\beta_0 + \beta_1 BeerTax_{i,1988} + \beta_2 Z_i + u_{i,1988}\right) - \left(\beta_0 + \beta_1 BeerTax_{i,1982} + \beta_2 Z_i + u_{i,1982}\right) \\ &= \beta_1 \left(BeerTax_{i,1988} - BeerTax_{i,1982}\right) + u_{i,1988} - u_{i,1982} \end{aligned}$$

where the constants β_0 and Z_i have been cancelled out of the difference.

Tax on Beer and Traffic Fatality Rate: Before/After



There is a negative relationship between changes in the fatality rate and changes in the beer tax between 1982 and 1988.

Tax on Beer and Traffic Fatality Rate: Before/After

Difference Operator Δ

- The difference operator, denoted Δ, is a convenient device to compute a difference between two regression equations with less notation.
- Let ΔX_i denote the difference operator:

$$\Delta X_i = X_{i,a} - X_{i,b} \quad \text{for some variable } X \text{ measured between dates } a \text{ and } b.$$

- Consider the regression equation in year t (where t stands for either date a or date b):

$$Y_{i,t} = \beta_0 + \beta_1 X_{i,t} + \beta_2 Z_i + u_{i,t}$$

- Apply the operator to the regression equation:

$$\Delta Y_i = \beta_1 \Delta X_i + \Delta u_i$$

(note that $\beta_0 + \beta_2 Z_i$ cancels out)

- Thus, let ΔX_i denote the difference operator for years 1988 and 1982:

$$\Delta\, FatalityRate_i = \beta_1 \Delta\, BeerTax_i + \Delta u_i$$

Tax on Beer and Traffic Fatality Rate: Before/After

Interpreting Before-and-After Effects

- If omitted variables such as "cultural attitude" or "law and order" did not change between 1982 and 1988, then they cannot be the cause of any observed change in fatalities in the state over that period.
- After differencing, omitted variables that are constant cancel out of the regression equation: **We do not need data for these omitted variables!**

$$\Delta \widehat{FatalityRate}_i = -0.072 - \underset{(0.065)}{1.04}\ \Delta\, BeerTax_i \qquad (0.36)$$

- The estimated regression suggests that traffic fatalities can be cut in half merely by increasing the real tax on beer by $\$1$ per case.
- This simple "before-and-after" analysis — regression in differences — would be cumbersome to extend to several time periods. Since our dataset contains seven years, we develop a more general approach: fixed effect regression.

Fixed Effects Regression

- Fixed effects regression is a method for controlling for omitted variables in panel data when the omitted variables vary across the n entities but do not change over the T time periods.
- In a fixed effects regression, there are n different intercepts, one for each entity, which can be represented by a set of binary variables.
- These binary variables absorb the influences of all omitted variables that differ from one entity to the next but are constant over time.
- Unlike a simple regression in differences, a fixed effects regression can be used when there are two or more time observations for each entity.

Fixed Effects Regression

Entity Fixed Effects

- Consider the regression model:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}$$

- Because Z_i varies from one state to the next but is constant over time, the population regression model can be interpreted as a system of n regressions with identical slope coefficients but different intercepts:

$$\begin{aligned} Y_{1t} &= \alpha_1 + \beta_1 X_{1t} + u_{1t} \\ Y_{2t} &= \alpha_2 + \beta_1 X_{2t} + u_{2t} \\ &\vdots \\ Y_{nt} &= \alpha_n + \beta_1 X_{nt} + u_{nt} \end{aligned}$$

where $\alpha_1, \alpha_2, \ldots \alpha_n$ are unknown intercepts to be estimated, one for each state.

- The $\alpha_i, i = 1, \ldots, n$ are entity fixed effects.

Fixed Effects Regression

Entity Fixed Effects as Binary Variables

- Let D_{ji} denote a binary variable for entity j, with $D_{ji} = 1$ if $i = j$ and $D_{ji} = 0$ if $i \neq j$.
- The regression model with entity fixed effects may be written in the equivalent form:

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

  where $\gamma_2, \ldots, \gamma_n$ denote the coefficients on the binary variables.
- The first binary variable, D_{1i} is omitted to avoid the "dummy variable trap."
- With this notation, the estimated regression equation for entity i may be written:

$$\hat{Y}_{it} = \hat{\beta}_0 + \hat{\beta}_1 X_{it} + \hat{\gamma}_i$$

  which maps to the earlier notation as $\hat{\alpha}_i = \hat{\beta}_0 + \hat{\gamma}_i$.
- Fixed effects estimates:

$$\widehat{FatalityRate} = \underset{(0.29)}{-0.66}\ BeerTax + \text{entity fixed effects}$$

Time Fixed Effects Regression

Time Fixed Effects as Binary Variables

- Just as fixed effects for each entity can control for variables that are constant over time but differ across entities, so time fixed effects can control for variables that are constant across entities but change over time.
- Consider the population regression equation:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + \beta_3 S_t + u_{it}$$

  - Z_i **varies across entities but is constant over time.**
  - S_t **varies over time but is constant across entities.**
- The time fixed effects regression model be represented using $T - 1$ binary indicators:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \cancel{\gamma_1 D_{1i}} + \gamma_2 D_{2i} + \ldots + \gamma_n D_{ni} + \cancel{\delta_1 B_{1t}} + \delta_2 B_{2t} + \ldots + \delta_T B_{Tt} + u_{it}$$

  where $B_{i,t} = 1$ if $t = i$ and $B_{i,t} = 0$ if $t \neq i$;
  where $\delta_2, \ldots, \delta_T$ are coefficients to be estimated;
  where $B_{1,t}$ is excluded to avoid the "dummy variable trap."

Time Fixed Effects Regression

Combined Entity and Time Fixed Effects

- Included regressors: The beer tax; 47 binary variables for entity fixed effects; 6 binary variables for time fixed effects; an intercept — or $1 + 47 + 6 + 1 = 55$ right-hand variables.

$$\widehat{FatalityRate} = \underset{(0.36)}{-0.64}\ BeerTax + \text{entity fixed effects} + \text{time fixed effects}$$

- Time effects are significant at the 10% level but not at the 5% level.
- Time effects have little effect on the coefficient on the real beer tax.
- Omitted variables issues:
  Only if there are effects that are neither fixed across entities nor fixed across time!

Fixed Effects: Assumptions for Causal Inference

Assumptions for Causal Inference

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}$$

where β_1 is the causal effect of X on Y.

1. $E[u_{it}|X_{i1}, X_{i2}, \ldots, X_{iT}, \alpha_i]$: The conditional mean of errors is zero.
2. $X_{i1}, X_{i2}, \ldots, X_{iT}, u_{i1}, u_{i2}, \ldots, u_{iT}$ are i.i.d. draws from the joint distribution.
3. (X_{it}, u_{it}) have non-zero finite fourth moments: outliers are unlikely.
4. No perfect multicollinearity.

- **Autocorrelation:**
  If X_{it} is correlated with X_{is} for different values of s and t, then X_{it} is "autocorrelated" or "serially correlated".
- **HAR standard errors:**
  Heteroskedasticity-and Autocorrelation-Robust (HAR) standard errors:
  Standard errors that are valid if u_{it} is (potentially or effectively) heteroskedastic and correlated over time within an entity. Example: **clustered standard errors**.

Drunk Driving Laws, Taxes and Traffic Deaths

- In addition to raising taxes to reduce drunk driving, states can also toughen driving laws.
- Both vehicle use and taxes depend in part on economic conditions (whether drivers have jobs; whether a state budget is strained).
- Omitting state laws and economic conditions could result in omitted variable bias.
- OLS regression of the fatality rate on the real beer tax without state and time fixed effects:
- The coefficient on the real beer tax is positive (0.36):
  Increasing beer taxes increases traffic fatalities!
- With state fixed effects, the coefficient on the real beer tax is now negative -0.66.
- The regression $\bar{R}^2$ jumps from 0.091 to 0.889 when fixed effects are included.
- Time effects have little influence on these estimates.

Dependent variable: traffic fatality rate (deaths per 10,000).

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Beer tax	0.36 (0.05) [0.26, 0.46]	−0.66 (0.29) [−1.23, −0.09]	−0.64 (0.36) [−1.35, 0.07]	−0.45 (0.30) [−1.04, 0.14]	−0.69 (0.35) [−1.38, 0.00]	−0.46 (0.31) [−1.07, 0.15]	−0.93 (0.34) [−1.60, −0.26]
Drinking age 18		0.10		0.03 (0.07) [−0.11, 0.17]	−0.01 (0.08) [−0.17, 0.15]		0.04 (0.10) [−0.16, 0.24]
Drinking age 19				−0.02 (0.05) [−0.12, 0.08]	−0.08 (0.07) [−0.21, 0.06]		−0.07 (0.10) [−0.26, 0.13]
Drinking age 20				0.03 (0.05) [−0.07, 0.13]	−0.10 (0.06) [−0.21, 0.01]		−0.11 (0.13) [−0.36, 0.14]
Drinking age						0.00 (0.02) [−0.05, 0.04]	
Mandatory jail or community service?				0.04 (0.10) [−0.17, 0.25]	0.09 (0.11) [−0.14, 0.31]	0.04 (0.10) [−0.17, 0.25]	0.09 (0.16) [−0.24, 0.42]
Average vehicle miles per driver				0.008 (0.007)	0.017 (0.011)	0.009 (0.007)	0.124 (0.049)
Unemployment rate				−0.063 (0.013)		−0.063 (0.013)	−0.091 (0.021)
Real income per capita (logarithm)				1.82 (0.64)		1.79 (0.64)	1.00 (0.68)
Years	1982–88	1982–88	1982–88	1982–88	1982–88	1982–88	1982 & 1988 only
State effects?	no	yes	yes	yes	yes	yes	yes
Time effects?	no	no	yes	yes	yes	yes	yes
Clustered standard errors?	no	yes	yes	yes	yes	yes	yes
***F*-Statistics and *p*-Values Testing Exclusion of Groups of Variables**							
Time effects = 0			4.22 (0.002)	10.12 (<0.001)	3.48 (0.006)	10.28 (<0.001)	37.49 (<0.001)
Drinking age coefficients = 0				0.35 (0.786)	1.41 (0.253)		0.42 (0.738)
Unemployment rate, income per capita = 0				29.62 (<0.001)		31.96 (<0.001)	25.20 (<0.001)
$\bar{R}^2$	0.091	0.889	0.891	0.926	0.893	0.926	0.899

Drunk Driving Laws, Taxes and Traffic Deaths

Main regression results:

1. Including the additional variables reduces the estimated effect of the beer tax from -0.64 in column (3) to -0.45 in column (4). The estimated effect of a $\$0.50$ increase in the beer tax is to decrease the expected fatality rate by $0.45 \times 0.50 = 0.23$ deaths per $10,000$. Since the average fatality rate is 2 deaths per $10,000$, a reduction of 0.23 corresponds to reducing traffic deaths by nearly one-eighth. However, the 95% confidence interval for this effect is quite large:

$$-0.45 \times 0.50 \pm 1.96 \times 0.30 \times 0.50 = (-0.52, 0.08)$$

2. The minimum legal drinking age is precisely estimated to have a small effect on traffic fatalities. The 95% confidence interval for the increase in the fatality rate in a state with a minimum legal drinking age of 18, relative to age 21, is $(-0.11, 0.17)$.

Drunk Driving Laws, Taxes and Traffic Deaths

Main regression results (continued):

3. The coefficient on the first offense punishment variable is also estimated to be small and is not significantly different from 0 at the 10% significance level.
4. The economic variables have considerable explanatory power for traffic fatalities.
   - High unemployment rates are associated with fewer fatalities:

     A one-percentage-point increase in the unemployment rate is expected to reduce traffic fatalities by 0.063 deaths per $10,000$.
   - High values of real per capita income are associated with high fatalities:

     A one-percentage increase in real per capita income is expected to decrease traffic fatalities by 0.0182 deaths per $10,000$.

Drunk Driving Laws, Taxes and Traffic Deaths

- Including fixed effects reduces the risk that omitted variable could bias least-squares estimates.
- Entity fixed effects eliminate the bias caused by unobserved variables that do not change over time, like cultural attitudes toward drinking and driving.
- Time fixed effects eliminate the bias caused by unobserved variables that do not vary across entities, like safety innovations and federal safety regulations.
- Omitted variable bias:
  - Changes in the real tax on beer could be correlated with other alcohol taxes, so the estimated effect could be driven by changes in other alcohol taxes. In this case, the real tax on beer would act as a "proxy" for a wider range of taxes.
  - Increases in the real beer tax could be associated with public education campaigns, so the estimated effect of the beer tax would also reflect the effect of a broader campaign to reduce drunk driving. In this case, the effect of the real tax on beer would be overestimated.
- Conclusion: Punishments and increases in the minimum legal drinking age do not have important effects on fatalities, while alcohol taxes do reduce traffic deaths.

Summary

- Panel data consist of observations on multiple units — households, countries — where each entity is observed at two or more time periods.
- Regression with fixed effects controls for unobserved variables that differ from one unit to the next but remain constant over time.
- When there are two time periods, fixed effects regression can be estimated by a "before and after" regression of the change in Y from the first period to the second on the corresponding change in X.
- Time fixed effects control for unobserved variables that are the same across units but vary over time.
- Standard errors need to allow both for serial correlation and for heteroskedasticity, and one way to do so is to use clustered standard errors.

Problems and Applications

Stock & Watson, Introduction (4th), Chapter 10, Exercise 1.

This exercise refers to the drunk driving panel data regressions summarized in Table 10.1.

1. New Jersey has a population of 8.1 million people. Suppose New Jersey increased the tax on a case of beer by $\$1$ (in 1988 dollars). Use the results in column (4) to predict the number of lives that would be saved over the next year. Construct a 95% confidence interval for your answer.
2. The drinking age in New Jersey is 21. Suppose New Jersey lowered its drinking age to 18. Use the results in column (4) to predict the change in the number of traffic fatalities in the next year. Construct a 95% confidence interval for your answer.
3. Should time effects be included in the regression? Why or why not?
4. A researcher conjectures that the unemployment rate has a different effect on traffic fatalities in the western states than in the other states. How would you test this hypothesis? (Be specific about the specification of the regression and the statistical test you would use.)

Keywords

balanced panel unbalanced panel fixed effects regression model entity fixed effects time fixed effects autocorrelated serially correlated heteroskedasticity-and autocorrelation-robust (HAR) standard errors clustered standard errors