6. Regression with panel data

Key feature of this section:

• Up to now, analysis of data on $n$ distinct entities at a given point of time
  (*cross sectional data*)

• Example:
  
  ■ Student-performance data set
  
  ■ Observations on different schooling characteristics in $n = 420$ districts (entities)

• Now, data structure in which each entity is observed at two or more points of time
  
  → *Panel data*
6.1. Structure of panel data sets

Definition 6.1: (Panel data)

Panel data consist of observations on the same $n$ entities at two or more time periods $T$. If the data set contains observations on the independent variables $X_1, X_2, \ldots, X_k$ and the dependent variable $Y$, then we denote the data by

$$(X_{1,it}, X_{2,it}, \ldots, X_{k,it}, Y_{it}), \quad i = 1, \ldots, n \text{ and } t = 1, \ldots, T,$$

where the first subscript, $i$, refers to the entity being observed and the second subscript, $t$, refers to the date at which it is observed.
Selected observations on cigarette sales, prices, and taxes, by state and year for U.S. states, 1985–1995

<table>
<thead>
<tr>
<th>Observation Number</th>
<th>State</th>
<th>Year</th>
<th>Cigarette Sales (packs per capita)</th>
<th>Average Price per Pack (including taxes)</th>
<th>Total Taxes (cigarette excise tax + sales tax)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alabama</td>
<td>1985</td>
<td>116.5</td>
<td>$1.022</td>
<td>$0.333</td>
</tr>
<tr>
<td>2</td>
<td>Arkansas</td>
<td>1985</td>
<td>128.5</td>
<td>$1.015</td>
<td>0.370</td>
</tr>
<tr>
<td>3</td>
<td>Arizona</td>
<td>1985</td>
<td>104.5</td>
<td>$1.086</td>
<td>0.362</td>
</tr>
<tr>
<td>47</td>
<td>West Virginia</td>
<td>1985</td>
<td>112.8</td>
<td>$1.089</td>
<td>0.382</td>
</tr>
<tr>
<td>48</td>
<td>Wyoming</td>
<td>1985</td>
<td>129.4</td>
<td>$0.935</td>
<td>0.240</td>
</tr>
<tr>
<td>49</td>
<td>Alabama</td>
<td>1986</td>
<td>117.2</td>
<td>$1.080</td>
<td>0.334</td>
</tr>
<tr>
<td>96</td>
<td>Wyoming</td>
<td>1986</td>
<td>127.8</td>
<td>$1.007</td>
<td>0.240</td>
</tr>
<tr>
<td>97</td>
<td>Alabama</td>
<td>1987</td>
<td>115.8</td>
<td>$1.135</td>
<td>0.335</td>
</tr>
<tr>
<td>528</td>
<td>Wyoming</td>
<td>1995</td>
<td>112.2</td>
<td>$1.585</td>
<td>0.360</td>
</tr>
</tbody>
</table>

Note: The cigarette consumption data set is described in Appendix 12.1.
Terminology:

- A balanced panel is a panel that has all its observations (focus of this lecture)
- An unbalanced panel is a panel that has some missing data for at least one time period or for at least one entity

Description of example data set:

- Traffic deaths and alcohol taxes (State Traffic Fatality (STF) data set)
- How effective are various government policies designed to discourage drunk driving in reducing traffic deaths?
Description of example data set: [continued]


- Important variables:
  - **FATALITYRATE** is the number of annual traffic deaths per 10000 people in the population in the state
  - **BEERTAX** is the 'real' tax on a case of beer put in 1988 U.S. dollars by adjusting for inflation
  - Various dummy variables indicating state-specific characteristics such as legal drinking age and punishment
Preliminary analysis:

- In a first step we focus on the two years 1982 and 1988 and, for each year, perform an OLS regression of \( \text{FATALITYRATE} \) on \( \text{BEERTAX} \).

- The estimated regression equations (neglecting subscripts) for the 1982 and 1988 data along with the standard errors (in brackets) are given by:

\[
\hat{\text{FATALITYRATE}} = 2.01 + 0.15 \cdot \text{BEERTAX} \quad (6.1)
\]

\[
\hat{\text{FATALITYRATE}} = 1.86 + 0.44 \cdot \text{BEERTAX} \quad (6.2)
\]

\[
\begin{array}{c}
(0.15) \\
(0.13)
\end{array}
\]

\[
\begin{array}{c}
(0.11) \\
(0.13)
\end{array}
\]
The traffic fatality rate and the tax on beer

Panel (a) is a scatterplot of traffic fatality rates and the real tax on a case of beer (in 1988 dollars) for 48 states in 1982. Panel (b) shows the data for 1988. Both plots show a positive relationship between the fatality rate and the real beer tax.

Fatality rate
(fatalities per 10,000)

FatalityRate = 2.01 + 0.15BeerTax

Beer tax
(dollars per case $1988)

Fatality rate
(fatalities per 10,000)

FatalityRate = 1.86 + 0.44BeerTax

Beer tax
(dollars per case $1988)
Preliminary analysis: [continued]

- The OLS estimate $\hat{\beta}_1$ for the 1982 data is not significant at the 10% level (the $t$-statistic is $1.15 < 1.64$)

- The OLS estimate $\hat{\beta}_1$ for the 1988 data is significant at the 1% level (the $t$-statistic is $3.43 > 2.58$)

- Both OLS estimates are positive what, taken literally, implies that higher real beer taxes are associated with more (not fewer) traffic fatalities

  → Indication of substantial omitted variable bias
Preliminary analysis: [continued]

- Some potentially neglected state-specific factors:
  - Quality of automobiles driven in the state
  - Quality of state highways
  - Rural versus urban driving
  - Density of cars on the road
  - Cultural acceptance of drinking and driving
Problem:

- Some of these variables (such as the cultural acceptance of drinking and driving) might be hard or even impossible to measure

Possible resort:

- If these factors remain constant over time in a given state, then we make use of the panel data structure to effectively hold these factors constant even though we cannot measure them
  
  \[\rightarrow\text{OLS regression with } \textit{fixed effects}\]
6.2. Panel data with two time periods: 'before-and-after' comparisons

Aim of this section:

- Provision of intuition on how we can exploit the panel data structure to mitigate the omitted-variable-bias problem

Approach:

- We consider a panel with $T = 2$ time periods
- We focus on changes in the dependent variable
- This 'before-and-after' comparison holds constant the unobserved factors that differ from one state to the next but do not change over time within the state
More explicitly:

- Consider the variable $Z_i$ with the following properties:
  - $Z_i$ determines the fatality rate in the $i^{th}$ state
  - $Z_i$ does not change over time
    (no time-subscript $t$)

- For example, $Z_i$ could represent the local cultural attitude towards drinking and driving which changes slowly
  (we consider it to be constant between 1982 and 1988)

- Regression equation:

$$\text{FATALITYRATE}_{it} = \beta_0 + \beta_1 \cdot \text{BEERTAX}_{it} + \beta_2 \cdot Z_i + u_{it} \quad (6.3)$$

with $i = 1, \ldots, n$ and $t = 1, 2$
Now:

- $Z_i$ does not change over time
  
  $\rightarrow$ $Z_i$ does not produce any change in FATALITYRATE between 1982 and 1988

- We eliminate the impact of $Z_i$ by analyzing the change in FATALITYRATE between the two periods

Derivation of the change:

- Regression equations for each time period:

  $$\text{FATALITYRATE}_{i1982} = \beta_0 + \beta_1 \cdot \text{BEERTAX}_{i1982} + \beta_2 \cdot Z_i + u_{i1982}$$
  $$\text{FATALITYRATE}_{i1988} = \beta_0 + \beta_1 \cdot \text{BEERTAX}_{i1988} + \beta_2 \cdot Z_i + u_{i1988}$$
Derivation of the change: [continued]

- Subtraction of both regression equations:

\[
FATALITYRATE_{i,1988} - FATALITYRATE_{i,1982} = \beta_1 \cdot (BEERTAX_{i,1988} - BEERTAX_{i,1982}) + u_{i,1988} - u_{i,1982} \quad (6.4)
\]

Interpretation of Eq. (6.4):

- \( Z_i \) does not change between 1982 and 1988
  \( \rightarrow \) Any changes in traffic fatalities over time must have arisen from other sources

  \( \rightarrow \) These changes are
  - changes in the tax on beer
  - changes in the error terms
    (capturing changes in other factors on traffic deaths)
More precisely:

- Specifying the regression changes in Eq. (6.4) eliminates the effect of the unobserved variables $Z_i$ that are constant over time.

- Analyzing changes in $Y$ and $X$ has the effect of controlling for variables that are constant over time thereby eliminating this source of omitted variable bias.

- Consider the change in the fatality rate between 1982 and 1988 against the change in the real beer tax between 1982 and 1988 for the 48 U.S. states.
Changes in fatality rates and beer taxes, 1982–1988

This is a scatterplot of the *change* in the traffic fatality rate and the *change* in real beer taxes between 1982 and 1988 for 48 states. There is a negative relationship between changes in the fatality rate and changes in the beer tax.

\[
\text{FatalityRate}_{1988} - \text{FatalityRate}_{1982} = -0.072 - 1.04(\text{BeerTax}_{1988} - \text{BeerTax}_{1982})
\]
Empirical results:

• OLS estimation results:

\[
\hat{\text{FATALITYRATE}}_{i1988} - \hat{\text{FATALITYRATE}}_{i1982} = -0.072 - 1.04 \cdot (\hat{\text{BEERTAX}}_{i1988} - \hat{\text{BEERTAX}}_{i1982}) \tag{6.5}
\]

\[(0.065) (0.36)\]

• Intercept in Eq. (6.5) allows for the possibility that the mean change in the fatality rate, in the absence of a change in the real beer tax, is nonzero

• The negative intercept (-0.072) could reflect improvements in auto safety from 1982 to 1988 that reduced the average fatality rate
Empirical results: [continued]

- Estimated effect of a change in the real beer tax is negative (as predicted by economic theory)

- OLS slope coefficient of $-1.04$ is significant at the 1% level (the absolute value of the $t$ statistic is $2.89 > 2.58$)

  ➔ Increase in the real beer tax by $1$ per case reduces the traffic fatality rate by $1.04$ deaths per 10000 people (substantial effect)

Remarks:

- The regression Eq. (6.5) controls for fixed factors such as cultural attitudes towards drinking and driving

- There are other factors influencing traffic safety
Remarks: [continued]

- If these factors change over time and are correlated with the real beer tax, then their omission will produce omitted variable bias
  
  → More careful analysis in Section 6.5

- Transference of the ideas valid for \( T = 2 \) to more than 2 time periods \( (T > 2) \)
  
  → Method of fixed effects regression
6.3. Fixed effects regression

Now:

- Method for controlling for omitted variables in panel data when the omitted variables vary across entities but do not change over time
- The fixed effects regression model has $n$ different intercepts, one for each entity
- These intercepts 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
More explicitly:

- Consider the regression model (6.3) from Slide 161:
  \[ Y_{it} = \beta_0 + \beta_1 \cdot X_{it} + \beta_2 \cdot Z_i + u_{it}, \]  
  (6.6)

where \( Z_i \) is an unobserved variable that varies from one state to the next but does not change over time (for example, \( Z_i \) represents cultural attitudes toward drinking and driving)

- We aim at estimating \( \beta_1 \), the effect on \( Y \) of \( X \) holding constant the unobserved state characteristic \( Z \)

- We can interpret Eq. (6.6) as having \( n \) intercepts, one for each entity
More explicitly: [continued]

- Specifically, define $\alpha_i \equiv \beta_0 + \beta_2 \cdot Z_i$, so that Eq. (6.6) becomes

$$Y_{it} = \beta_1 \cdot X_{it} + \alpha_i + u_{it}$$ (6.7)

- $\alpha_1, \ldots, \alpha_n$ are treated as state-specific intercepts to be estimated

- Population regression line for the $i^{th}$ state: $Y_{it} = \alpha_i + \beta_1 \cdot X_{it}$

- The slope coefficient $\beta_1$ is the same for all states, but the intercept varies from one state to the next

- The intercept $\alpha_i$ can be thought of as the 'effect' of being in entity $i$
More explicitly: [continued]

- The terms $\alpha_1, \ldots, \alpha_n$ are known as *entity fixed effects*
- The variation in the entity fixed effects comes from omitted variables (like $Z_i$ in Eq. (6.6)) that vary across entities but not over time
- Eq. (6.7) is known as the *fixed effects regression model*

Representation with dummy variables:

- Consider the $n - 1$ dummy variables

\[
D_{2,i} = \begin{cases} 
1 & \text{when } i = 2 \\
0 & \text{otherwise}
\end{cases}, \ldots, D_{n,i} = \begin{cases} 
1 & \text{when } i = n \\
0 & \text{otherwise}
\end{cases}
\]
Then, the fixed effects regression model (6.7) can be equivalently expressed as

\[ Y_{it} = \beta_0 + \beta_1 \cdot X_{it} + \gamma_2 \cdot D_{2,i} + \ldots + \gamma_n \cdot D_{n,i} + u_{it}, \quad (6.8) \]

where \( \beta_0, \beta_1, \gamma_2, \ldots, \gamma_n \) are coefficients to be estimated.

- Relationships between parameters in Eqs. (6.7) and (6.8):

\[ \alpha_1 = \beta_0, \alpha_2 = \beta_0 + \gamma_2, \ldots, \alpha_n = \beta_0 + \gamma_n \]

- The entity-specific intercepts in Eq. (6.7) and the binary regressors in Eq. (6.8) have the same source, namely the unobserved variable \( Z_i \) that varies across entities but not over time.
Now:

- Extension to multiple $X$-regressors

**Definition 6.2: (Fixed effects regression model)**

The **fixed effects regression model** is

$$Y_{it} = \beta_1 \cdot X_{1,it} + \ldots + \beta_k \cdot X_{k,it} + \alpha_i + u_{it}, \quad (6.9)$$

where $i = 1, \ldots, n$ and $t = 1, \ldots, T$ and $\alpha_1, \ldots, \alpha_n$ are the entity-specific intercepts. Equivalently, the fixed effects regression model can be written in terms of a common intercept, the $X$-regressors and the $n - 1$ dummy variables defined on Slide 172:

$$Y_{it} = \beta_0 + \beta_1 \cdot X_{1,it} + \ldots + \beta_k \cdot X_{k,it} + \gamma_2 \cdot D_{2,i} + \ldots + \gamma_n \cdot D_{n,i} + u_{it}. \quad (6.10)$$
Estimation and inference:

- In principle, the binary variable specification (6.10) can be estimated via OLS

- However, specification (6.10) requires estimation of $k + n$ parameters which becomes problematic if the number of entities $n$ is large

  \[\rightarrow\] Use of special routines for OLS estimation of fixed effects regressions

- (Two-step) entity-demeaned OLS algorithm
  - Subtract the entity-specific averages from each variable
  - Perform OLS regression using the entity-demeaned variables
Estimation and inference: [continued]

- Example:
  - Consider the (single-regressor) fixed effects model (6.7)
  - Taking (time) averages on both sides of (6.7) yields
    \[
    \bar{Y}_i = \beta_1 \cdot \bar{X}_i + \alpha_i + \bar{u}_i
    \]
    with \( \bar{Y}_i = \frac{1}{T} \sum_{t=1}^{T} Y_{it} \) and \( \bar{X}_i \) and \( \bar{u}_i \) similarly defined
  - It follows from Eq. (6.7) that
    \[
    \frac{Y_{it} - \bar{Y}_i}{\equiv \tilde{Y}_{it}} = \beta_1 \cdot X_{it} + \alpha_i + u_{it} - \beta_1 \cdot \bar{X}_i - \alpha_i - \bar{u}_i
    \]
    \[
    = \beta_1 \cdot (X_{it} - \bar{X}_i) + (u_{it} - \bar{u}_i)
    \equiv \tilde{X}_{it} + \tilde{u}_{it}
    \]
    \[
    = \beta_1 \cdot \tilde{X}_{it} + \tilde{u}_{it}
    \]
    (6.11)
Estimation and inference: [continued]

• Example: [continued]
  ■ Estimation of $\beta_1$ in Eq. (6.11) via OLS

• Under certain assumptions stated on Slide 187 (the so-called fixed effects regression assumptions)
  ■ the sampling distribution of the OLS estimator is normal in large samples
  ■ the variance and the standard error of the sampling distribution can be estimated from the data

→ Hypothesis testing (based on $t$- and $F$-statistics) and construction of confidence intervals in exactly the same way as in multiple regressions with cross-sectional data
Application to traffic deaths:

• OLS estimate of the fixed effects regression based on all 
  \( T = 7 \) years of data (observations) is

\[
\hat{\text{FATALITYRATE}} = -0.66 \cdot \text{BEERTAX} + \text{StateFixedEffects} \\
(0.29)
\]

• The sign of \( \hat{\beta}_1 \) is negative and the coefficient is significant 
  at the 5\% level

• Including state fixed effects avoids omitted variable bias arising 
  from omitted factors that vary across states but are constant 
  over time

• What about the effects of omitted factors that evolve over 
  time but are the same for all states? 
  (for example, overall automobile safety improvements)

\( \rightarrow \) Regression with \textit{time fixed effects}
6.4. Regression with time fixed effects

Now:

- We aim at controlling for variables that are constant across entities but evolve over time (such as overall safety improvements in new cars)
- To this end, we augment our regression Eq. (6.6) from Slide 170 to take the form

\[ Y_{it} = \beta_0 + \beta_1 \cdot X_{it} + \beta_2 \cdot Z_i + \beta_3 \cdot S_t + u_{it}, \]  

(6.12)

where \( S_t \) is an unobserved variable (representing automobile safety) that changes over time but is constant across states.

- Note that omitting \( S_t \) from the regression may lead to omitted variable bias.
Time effects only:

• Let us consider for the moment that the variables $Z_i$ are not present, so that Eq. (6.12) becomes

$$Y_{it} = \beta_0 + \beta_1 \cdot X_{it} + \beta_3 \cdot S_t + u_{it} \quad (6.13)$$

• Similar to the entity fixed effects model, it is possible to eliminate $S_t$ from Eq. (6.13)

• Specifically, we set $\lambda_t = \beta_0 + \beta_3 \cdot S_t$ to obtain

$$Y_{it} = \beta_1 \cdot X_{it} + \lambda_t + u_{it} \quad (6.14)$$

• This model has a different intercept, $\lambda_t$, for each time period which can be thought of as the effect on $Y$ of time period $t$

• $\lambda_1, \ldots, \lambda_T$ are known as time fixed effects whose variation stems from omitted variables (like $S_t$) that vary over time but not across entities
Time effects only: [continued]

- Considering the $T - 1$ binary variables

$$B_{2,t} = \begin{cases} 1 & \text{when } t = 2 \\ 0 & \text{otherwise} \end{cases}, \ldots, B_{T,t} = \begin{cases} 1 & \text{when } t = T \\ 0 & \text{otherwise} \end{cases}$$

we can equivalently express model (6.14) as

$$Y_{it} = \beta_0 + \beta_1 \cdot X_{it} + \delta_2 \cdot B_{2,t} + \ldots + \delta_T \cdot B_{T,t} + u_{it}, \quad (6.15)$$

where $\beta_0, \beta_1, \delta_2, \ldots, \delta_T$ are coefficients to be estimated.

- Relationships between parameters in Eqs. (6.14) and (6.15):

$$\lambda_1 = \beta_0, \lambda_2 = \beta_0 + \delta_2, \ldots, \lambda_T = \beta_0 + \delta_T$$

(see Eqs. (6.7) and (6.8) on Slides 171, 173)
Now:
- Combination of entity and time fixed effects

**Definition 6.3:** (Entity and time fixed effects regression model)

The **fixed effects regression model** is

$$Y_{it} = \beta_1 \cdot X_{1,it} + \ldots + \beta_k \cdot X_{k,it} + \alpha_i + \lambda_t + u_{it}, \quad (6.16)$$

where $\alpha_1, \ldots, \alpha_n$ are the entity fixed and $\lambda_1, \ldots, \lambda_T$ time fixed effects. Equivalently, the entity and time fixed effects regression model can be written in terms of a common intercept, the $X$-regressors and the $n - 1$ and $T - 1$ dummy variables defined on Slides 172, 181:

$$Y_{it} = \beta_0 + \beta_1 \cdot X_{1,it} + \ldots + \beta_k \cdot X_{k,it}$$
$$+ \gamma_2 \cdot D_{2,i} + \ldots + \gamma_n \cdot D_{n,i}$$
$$+ \delta_2 \cdot B_{2,t} + \ldots + \delta_T \cdot B_{T,t} + u_{it}. \quad (6.17)$$
Remark:

- The combined entity and time fixed effects regression model eliminates omitted variables bias arising both from unobserved variables that are constant over time and from variables that are constant across states.

Parameter estimation:

- The full model (6.17) can in principle be estimated by OLS.
- Most software packages implement a two-step algorithm using entity and time-period demeaned $Y$ and $X$-variables.
Application to traffic deaths:

- OLS estimate of the entity and time fixed effects regression:
  \[
  \hat{FATALITYRATE} = -0.64 \cdot BEERTAX + SF_{\text{Effects}} + TF_{\text{Effects}} (0.36)
  \]

- This specification includes
  - 47 state binary variables (state fixed effects, not reported)
  - 6 single-year binary variables (time fixed effects, not reported)
  - the variable BEERTAX
  - the intercept (not reported)
Application to traffic deaths: [continued]

- Time fixed effects have little impact on beer tax coefficient (cf. regression estimation on Slide 178)

- Coefficient is significant at the 10% level (but not at the 5% level; \( t \)-statistic is \(-0.64/0.36 = -1.78\))

- Estimation is immune to omitted variable bias from variables that are constant either over time or across states

- However, other relevant but omitted variables may vary both across states and over time
  \( \rightarrow \) Specification might still be subject to omitted variable bias

- More careful analysis of the dataset
  \( \rightarrow \) see class
6.5. The fixed effects regression assumptions and standard errors for fixed effects regression

Aim of this section:

• Formulation of OLS assumptions of the fixed effects regression model so that Theorem 2.4 on Slide 19 holds for the involved OLS estimators (especially the asymptotic normal distribution when \( n \) is large)

• Some comments on the standard errors for fixed effects regressions
**Definition 6.4:** (Fixed effects regression assumptions)

We consider the *fixed effects regression model*

\[ Y_{it} = \beta_1 \cdot X_{it} + \alpha_i + u_{it}, \quad i = 1, \ldots, n, t = 1, \ldots, T. \]

The following are called the *fixed effects regression assumptions*:

1. \( u_{it} \) has conditional mean zero:
   \[ E(u_{it} | X_{i1}, X_{i2}, \ldots, X_{iT}, \alpha_i) = 0. \]

2. \((X_{i1}, X_{i2}, \ldots, X_{iT}, u_{i1}, u_{i2}, \ldots, u_{iT})\), \( i = 1, \ldots, n \), are i.i.d. draws from their joint distribution.

3. Large outliers are unlikely: \( X_{it} \) and \( u_{it} \) have nonzero finite fourth moments.

4. There is no perfect multicollinearity.

For multiple regressors, \( X_{it} \) should be replaced by the full list \( X_{1,it}, X_{2,it}, \ldots, X_{k,it} \).
Remarks:

- Definition 6.4 focuses on entity fixed effects regressions neglecting time effects
- An extension for including time fixed effects is straightforward

Standard errors for fixed effects regression:

- Autocorrelated errors are a pervasive phenomenon in data with a time component (see Section 3.1.2. on Slides 48, 49)
- In the case of autocorrelated errors standard errors should be computed using the HAC estimator of the variance
- One type of HAC errors are *clustered errors* used in the traffic-fatality dataset