ECON4150 - Introductory Econometrics

Lecture 10: Panel data

Monique de Haan
(moniqued@econ.uio.no)

Stock and Watson Chapter 10
OLS: The Least Squares Assumptions

\[ Y_i = \beta_0 + \beta_1 X_i + u_i \]

Assumption 1: conditional mean zero assumption: \( E[u_i|X_i] = 0 \)
Assumption 2: \((X_i, Y_i)\) are i.i.d. draws from joint distribution
Assumption 3: Large outliers are unlikely

- Under these three assumption the OLS estimators are unbiased, consistent and normally distributed in large samples.

- Last week we discussed threats to internal validity

- In this lecture we discuss a method we can use in case of omitted variables
  - Omitted variable is a determinant of the outcome \( Y_i \)
  - Omitted variable is correlated with regressor of interest \( X_i \)
Multiple regression model was introduced to mitigate omitted variables problem of simple regression

\[ Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i \]

Even with multiple regression there is threat of omitted variables:

- some factors are difficult to measure
- sometimes we are simply ignorant about relevant factors

Multiple regression based on panel data may mitigate detrimental effect of omitted variables without actually observing them.
Panel data

Cross-sectional data:
A sample of individuals observed in 1 time period

Panel data: same sample of individuals observed in multiple time periods
• Panel data consist of observations on \( n \) entities (cross-sectional units) and \( T \) time periods

• Particular observation denoted with two subscripts (\( i \) and \( t \))

\[
Y_{it} = \beta_0 + \beta_1 X_{it} + u_{it}
\]

• \( Y_{it} \) outcome variable for individual \( i \) in year \( t \)

• For balanced panel this results in \( nT \) observations
Advantages of panel data

- More control over omitted variables.
- More observations.
- Many research questions typically involve a time component.
The effect of alcohol taxes on traffic deaths

- About 40,000 traffic fatalities each year in the U.S.
- Approximately 25% of fatal crashes involve drivers who drank alcohol.
- Government wants to reduce traffic fatalities.
- One potential policy: increase the tax on alcoholic beverages.
- We have data on traffic fatality rate and tax on beer for 48 U.S. states in 1982 and 1988.
- What is the effect of increasing the tax on beer on the traffic fatality rate?
Traffic deaths and alcohol taxes in 1982

\[ \text{Fatality Rate}_{i,1982} = 2.01 + 0.15 \text{ BeerTax}_{i,1982} \]

\( (0.14) \quad (0.18) \)
Data from 1988

Traffic deaths and alcohol taxes in 1988

\[ \text{FatalityRate}_{i,1988} = 1.86 + 0.44 \times \text{BeerTax}_{i,1988} \]
Panel data: before-after analysis

- Both regression using data from 1982 & 1988 likely suffer from omitted variable bias
- We can use data from 1982 and 1988 together as panel data
- Panel data with $T = 2$
- Observed are $Y_{i1}, Y_{i2}$ and $X_{i1}, X_{i2}$
- Suppose model is

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

  and we assume $E(u_{it}|X_{i1}, X_{i2}, Z_i) = 0$
- $Z_i$ are (unobserved) variables that vary between states but not over time
  - (such as local cultural attitude towards drinking and driving)
- Parameter of interest is $\beta_1$
Panel data

<table>
<thead>
<tr>
<th>state</th>
<th>year</th>
<th>beertax</th>
<th>fatalityrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AL</td>
<td>1982</td>
<td>1.539379</td>
</tr>
<tr>
<td>7</td>
<td>AL</td>
<td>1988</td>
<td>1.501444</td>
</tr>
<tr>
<td>8</td>
<td>AZ</td>
<td>1982</td>
<td>0.2147971</td>
</tr>
<tr>
<td>14</td>
<td>AZ</td>
<td>1988</td>
<td>0.346487</td>
</tr>
<tr>
<td>15</td>
<td>AR</td>
<td>1982</td>
<td>0.650358</td>
</tr>
<tr>
<td>21</td>
<td>AR</td>
<td>1988</td>
<td>0.5245429</td>
</tr>
<tr>
<td>22</td>
<td>CA</td>
<td>1982</td>
<td>0.1073986</td>
</tr>
<tr>
<td>28</td>
<td>CA</td>
<td>1988</td>
<td>0.0866218</td>
</tr>
<tr>
<td>29</td>
<td>CO</td>
<td>1982</td>
<td>0.2147971</td>
</tr>
<tr>
<td>35</td>
<td>CO</td>
<td>1988</td>
<td>0.1732435</td>
</tr>
<tr>
<td>36</td>
<td>CT</td>
<td>1982</td>
<td>0.2243437</td>
</tr>
<tr>
<td>42</td>
<td>CT</td>
<td>1988</td>
<td>0.2172185</td>
</tr>
</tbody>
</table>
Panel data: before

- Consider cross-sectional regression for first period ($t = 1$):

  \[ Y_{i1} = \beta_0 + \beta_1 X_{i1} + \beta_2 Z_i + u_{i1} \quad E[u_i|X_{i1}, Z_i] = 0 \]

- $Z_i$ observed: multiple regression of $Y_{i1}$ on constant, $X_{i1}$ and $Z_i$ leads to unbiased and consistent estimator of $\beta_1$

- $Z_i$ not observed: regression of $Y_{i1}$ on constant and $X_{i1}$ only results in unbiased estimator of $\beta_1$ when $\text{Cov}(X_{i1}, Z_i) = 0$

- What can we do if we don’t observe $Z_i$?
• We also observe $Y_{i2}$ and $X_{i2}$, hence model for second period is:

$$Y_{i2} = \beta_0 + \beta_1 X_{i2} + \beta_2 Z_i + u_{i2}$$

• Similar to argument before cross-sectional analysis for period 2 might fail

• Problem is again the unobserved heterogeneity embodied in $Z_i$
Before-after analysis (first differences)

- We have
  \[ Y_{i1} = \beta_0 + \beta_1 X_{i1} + \beta_2 Z_i + u_{i1} \]
  and
  \[ Y_{i2} = \beta_0 + \beta_1 X_{i2} + \beta_2 Z_i + u_{i2} \]

- Subtracting period 1 from period 2 gives
  \[ Y_{i2} - Y_{i1} = (\beta_0 + \beta_1 X_{i2} + \beta_2 Z_i + u_{i2}) - (\beta_0 + \beta_1 X_{i1} + \beta_2 Z_i + u_{i1}) \]

- Applying OLS to:
  \[ Y_{i2} - Y_{i1} = \beta_1 (X_{i2} - X_{i1}) + (u_{i2} - u_{i1}) \]
  will produce an unbiased and consistent estimator of \( \beta_1 \)

- Advantage of this regression is that we do not need data on \( Z \)

- By analyzing changes in dependent variable we automatically control for time-invariant unobserved factors
Data from 1982 and 1988

Traffic deaths and alcohol taxes: before–after

\[
\text{Fatality}_{i,1988} - \text{Fatality}_{i,1982} = -0.07 - 1.04 (\text{BeerTax}_{i,1988} - \text{BeerTax}_{i,1982})
\]

(0.06) (0.42)
Panel data with more than 2 time periods

<table>
<thead>
<tr>
<th>state</th>
<th>year</th>
<th>fatalityrate</th>
<th>beertax</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL</td>
<td>1982</td>
<td>2.12836</td>
<td>1.539379</td>
</tr>
<tr>
<td>AL</td>
<td>1983</td>
<td>2.34848</td>
<td>1.788991</td>
</tr>
<tr>
<td>AL</td>
<td>1984</td>
<td>2.33643</td>
<td>1.714286</td>
</tr>
<tr>
<td>AL</td>
<td>1985</td>
<td>2.19348</td>
<td>1.652542</td>
</tr>
<tr>
<td>AL</td>
<td>1986</td>
<td>2.66914</td>
<td>1.609907</td>
</tr>
<tr>
<td>AL</td>
<td>1987</td>
<td>2.71859</td>
<td>1.56</td>
</tr>
<tr>
<td>AL</td>
<td>1988</td>
<td>2.49391</td>
<td>1.501444</td>
</tr>
<tr>
<td>AZ</td>
<td>1982</td>
<td>2.49914</td>
<td>2147971</td>
</tr>
<tr>
<td>AZ</td>
<td>1983</td>
<td>2.26738</td>
<td>206422</td>
</tr>
<tr>
<td>AZ</td>
<td>1984</td>
<td>2.82878</td>
<td>2967033</td>
</tr>
<tr>
<td>AZ</td>
<td>1985</td>
<td>2.80201</td>
<td>3813559</td>
</tr>
<tr>
<td>AZ</td>
<td>1986</td>
<td>3.07106</td>
<td>371517</td>
</tr>
<tr>
<td>AZ</td>
<td>1987</td>
<td>2.76728</td>
<td>36</td>
</tr>
<tr>
<td>AZ</td>
<td>1988</td>
<td>2.70565</td>
<td>346487</td>
</tr>
<tr>
<td>AR</td>
<td>1982</td>
<td>2.38405</td>
<td>650358</td>
</tr>
<tr>
<td>AR</td>
<td>1983</td>
<td>2.3957</td>
<td>6754587</td>
</tr>
<tr>
<td>AR</td>
<td>1984</td>
<td>2.23785</td>
<td>5989011</td>
</tr>
<tr>
<td>AR</td>
<td>1985</td>
<td>2.26367</td>
<td>5773305</td>
</tr>
<tr>
<td>AR</td>
<td>1986</td>
<td>2.54323</td>
<td>5624355</td>
</tr>
<tr>
<td>AR</td>
<td>1987</td>
<td>2.67588</td>
<td>545</td>
</tr>
<tr>
<td>AR</td>
<td>1988</td>
<td>2.54697</td>
<td>5245429</td>
</tr>
</tbody>
</table>
Panel data with more than 2 time periods

- Panel data with $T \geq 2$

\[ Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}, \quad i = 1, \ldots, n; \quad t = 1, \ldots, T \]

- $Y_{it}$ is dependent variable; $X_{it}$ is explanatory variable; $Z_i$ are state specific, time invariant variables

- Equation can be interpreted as model with $n$ specific intercepts (one for each state)

\[ Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}, \quad \text{with} \quad \alpha_i = \beta_0 + \beta_2 Z_i \]

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

- $\alpha_i$ models impact of omitted time-invariant variables on $Y_{it}$
State specific intercepts

Predicted fatality rate

beer tax

Alabama
Arizona
Arkansa
California
Having data on $Y_{it}$ and $X_{it}$ how to determine $\beta_1$?

- Population regression model: $Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}$
- In order to estimate the model we have to quantify $\alpha_i$
- Solution: create $n$ dummy variables $D1_i, ..., Dn_i$
  - with $D1_i = 1$ if $i = 1$ and 0 otherwise,
  - with $D2_i = 1$ if $i = 2$ and 0 otherwise, ....
- Population regression model can be written as:
  $$Y_{it} = \beta_1 X_{it} + \alpha_1 D1_i + \alpha_2 D2_i + ... + \alpha_n Dn_i + u_{it}$$
Alternatively, population regression model can be written as:

\[ Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \ldots + \gamma_n Dn_i + u_{it} \]

with \( \beta_0 = \alpha_1 \) and \( \gamma_i = \alpha_i - \beta_0 \) for \( i > 1 \)

- Interpretation of \( \beta_1 \) identical for both representations

- Ordinary Least Squares (OLS): choose \( \hat{\beta}_0, \hat{\beta}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_n \) to minimize squared prediction mistakes (SSR):

\[
\sum_{i=1}^{n} \sum_{t=1}^{T} \left( Y_{it} - \hat{\beta}_0 - \hat{\beta}_1 X_{it} - \hat{\gamma}_2 D2_i - \ldots - \hat{\gamma}_n Dn_i \right)^2
\]

- SSR is function of \( \hat{\beta}_0, \hat{\beta}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_n \)
Fixed effect regression model
Least squares with dummy variables

\[
\sum_{i=1}^{n} \sum_{t=1}^{T} \left( Y_{it} - \hat{\beta}_0 - \hat{\beta}_1 X_{it} - \hat{\gamma}_2 D_{2i} - \ldots - \hat{\gamma}_n D_{ni} \right)^2
\]

OLS procedure:

- Take partial derivatives of SSR w.r.t. \(\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma}_2\ldots, \hat{\gamma}_n\)
- Equal partial derivatives to zero resulting in \(n + 1\) equations with \(n + 1\) unknown coefficients
- Solutions are the OLS estimators \(\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma}_2\ldots, \hat{\gamma}_n\)
Fixed effect regression model
Least squares with dummy variables

- Analytical formulas require matrix algebra
- Algebraic properties OLS estimators (normal equations, linearity) same as for simple regression model
- Extension to multiple $X$’s straightforward: $n + k$ normal equations
- OLS procedure is also labeled Least Squares Dummy Variables (LSDV) method
- Dummy variable trap: Never include all $n$ dummy variables and the constant term!
Fixed effect regression model
Within estimation

- Typically \( n \) is large in panel data applications
- With large \( n \) computer will face numerical problem when solving system of \( n + 1 \) equations
- OLS estimator can be calculated in two steps
- First step: demean \( Y_{it} \) and \( X_{it} \)
- Second step: use OLS on demeaned variables
Fixed effect regression model
Within estimation

- We have

\[ Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it} \]

\[ \bar{Y}_i = \beta_1 \bar{X}_i + \alpha_i + \bar{u}_i \]

- \( \bar{Y}_i = \frac{1}{T} \sum_{t=1}^{T} Y_{it} \), etc. is entity mean

- Subtracting both expressions leads to

\[ Y_{it} - \bar{Y}_i = (\beta_1 X_{it} + \alpha_i + u_{it}) - (\beta_1 \bar{X}_i + \alpha_i + \bar{u}_i) \]

\[ \tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it} \]

- \( \tilde{Y}_{it} = Y_{it} - \bar{Y}_i \), etc. is entity demeaned variable

- \( \alpha_i \) has disappeared; OLS on demeaned variables involves solving one normal equation only!
Fixed effect regression model
Within estimation

<table>
<thead>
<tr>
<th>state</th>
<th>year</th>
<th>fatalityrate</th>
<th>Meanfatality</th>
<th>DmeanFatal</th>
<th>beertax</th>
<th>Meanbeertax</th>
<th>DmeanBeertax</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AL</td>
<td>1982</td>
<td>2.12836</td>
<td>2.41267</td>
<td>-.2842672</td>
<td>1.539379</td>
<td>1.623793</td>
</tr>
<tr>
<td>2</td>
<td>AL</td>
<td>1983</td>
<td>2.34848</td>
<td>2.41267</td>
<td>-.0641472</td>
<td>1.788991</td>
<td>1.623793</td>
</tr>
<tr>
<td>3</td>
<td>AL</td>
<td>1984</td>
<td>2.33643</td>
<td>2.41267</td>
<td>-.0761971</td>
<td>1.714286</td>
<td>1.623793</td>
</tr>
<tr>
<td>4</td>
<td>AL</td>
<td>1985</td>
<td>2.19348</td>
<td>2.41267</td>
<td>-.2191472</td>
<td>1.652542</td>
<td>1.623793</td>
</tr>
<tr>
<td>5</td>
<td>AL</td>
<td>1986</td>
<td>2.66914</td>
<td>2.41267</td>
<td>.2565126</td>
<td>1.609907</td>
<td>1.623793</td>
</tr>
<tr>
<td>6</td>
<td>AL</td>
<td>1987</td>
<td>2.71859</td>
<td>2.41267</td>
<td>.3059628</td>
<td>1.56</td>
<td>1.623793</td>
</tr>
<tr>
<td>7</td>
<td>AL</td>
<td>1988</td>
<td>2.49391</td>
<td>2.41267</td>
<td>.081829</td>
<td>1.501444</td>
<td>1.623793</td>
</tr>
<tr>
<td>8</td>
<td>AZ</td>
<td>1982</td>
<td>2.49914</td>
<td>2.7059</td>
<td>-.2067599</td>
<td>2.147971</td>
<td>.3110403</td>
</tr>
<tr>
<td>9</td>
<td>AZ</td>
<td>1983</td>
<td>2.26738</td>
<td>2.7059</td>
<td>.43852</td>
<td>2.06422</td>
<td>.3110403</td>
</tr>
<tr>
<td>10</td>
<td>AZ</td>
<td>1984</td>
<td>2.82878</td>
<td>2.7059</td>
<td>.12288</td>
<td>2.967033</td>
<td>.3110403</td>
</tr>
<tr>
<td>11</td>
<td>AZ</td>
<td>1985</td>
<td>2.80201</td>
<td>2.7059</td>
<td>.0961101</td>
<td>3.813559</td>
<td>.3110403</td>
</tr>
<tr>
<td>12</td>
<td>AZ</td>
<td>1986</td>
<td>3.07106</td>
<td>2.7059</td>
<td>.36516</td>
<td>3.71517</td>
<td>.3110403</td>
</tr>
<tr>
<td>13</td>
<td>AZ</td>
<td>1987</td>
<td>2.76728</td>
<td>2.7059</td>
<td>.0613801</td>
<td>.36</td>
<td>.3110403</td>
</tr>
<tr>
<td>14</td>
<td>AZ</td>
<td>1988</td>
<td>2.70565</td>
<td>2.7059</td>
<td>-.0002499</td>
<td>.346487</td>
<td>.3110403</td>
</tr>
<tr>
<td>15</td>
<td>AR</td>
<td>1982</td>
<td>2.38405</td>
<td>2.435336</td>
<td>-.0512855</td>
<td>.650358</td>
<td>.5905753</td>
</tr>
<tr>
<td>16</td>
<td>AR</td>
<td>1983</td>
<td>2.3957</td>
<td>2.435336</td>
<td>-.0396357</td>
<td>.675458</td>
<td>.5905753</td>
</tr>
<tr>
<td>17</td>
<td>AR</td>
<td>1984</td>
<td>2.23785</td>
<td>2.435336</td>
<td>-.1974857</td>
<td>.5989011</td>
<td>.5905753</td>
</tr>
<tr>
<td>18</td>
<td>AR</td>
<td>1985</td>
<td>2.26367</td>
<td>2.435336</td>
<td>-.1716657</td>
<td>.5773305</td>
<td>.5905753</td>
</tr>
<tr>
<td>19</td>
<td>AR</td>
<td>1986</td>
<td>2.54323</td>
<td>2.435336</td>
<td>.1078944</td>
<td>.5624355</td>
<td>.5905753</td>
</tr>
<tr>
<td>20</td>
<td>AR</td>
<td>1987</td>
<td>2.67588</td>
<td>2.435336</td>
<td>.2405446</td>
<td>.545</td>
<td>.5905753</td>
</tr>
<tr>
<td>21</td>
<td>AR</td>
<td>1988</td>
<td>2.54697</td>
<td>2.435336</td>
<td>.1116343</td>
<td>.5245429</td>
<td>.5905753</td>
</tr>
</tbody>
</table>
Fixed effect regression model
Within estimation

- Entity demeaning is often called the Within transformation.
- Within transformation is generalization of "before-after" analysis to more than \( T = 2 \) periods.
- Before-after:
  \[
  Y_{i2} - Y_{i1} = \beta_1 (X_{i2} - X_{i1}) + (u_{i2} - u_{i1})
  \]
- Within:
  \[
  Y_{it} - \bar{Y}_i = \beta_1 (X_{it} - \bar{X}_i) + (u_{it} - \bar{u}_i)
  \]
- LSDV and Within estimators are identical:
  \[
  \hat{FatalityRate}_{it} = -0.66 \times BeerTax_{it} + State \text{ dummies} \\
  (0.19)
  \]
  \[
  (\hat{FatalityRate}_{it} - \hat{FatalityRate}) = -0.66 \times (BeerTax_{it} - \bar{BeerTax}) \\
  (0.19)
  \]
In addition to entity effects we can also include time effects in the model.

Time effects control for omitted variables that are common to all entities but vary over time.

Typical example of time effects: macroeconomic conditions or federal policy measures are common to all entities (e.g. states) but vary over time.

Panel data model with entity and time effects:

\[ Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it} \]
Fixed effects regression model

- OLS estimation straightforward extension of LSDV/Within estimators of model with only entity fixed effects

- LSDV: create $T$ dummy variables $B_1 t .... B_T t$

$$
Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D_{2i} + ... + \gamma_n D_{ni} \\
+ \delta_2 B_{2t} + \delta_3 B_{3t} + ... + \delta_T B_{Tt} + u_{it}
$$

- Within estimation: Deviating $Y_{it}$ and $X_{it}$ from their entity and time-period means

- The effect of the tax on beer on the traffic fatality rate:

$$
\text{FatalityRate}_{it} = -0.64 \text{ BeerTax}_{it} + \text{ State dummies} + \text{ Time dummies} \\
(0.20)
$$
$Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it}$

Statistical assumptions are:

ASS #1: $\mathbb{E}(u_{it}|X_{i1}, \ldots, X_{iT}, \alpha_i, \lambda_t) = 0$

ASS #2: $(X_{i1}, \ldots, X_{iT}, Y_{i1}, \ldots, Y_{iT})$ are i.i.d. over the cross-section

ASS #3: Large outliers are unlikely

ASS #4: No perfect multicollinearity

ASS #5: $\text{cov}(u_{it}, u_{is}|X_{i1}, \ldots, X_{iT}, \alpha_i, \lambda_t) = 0$ for $t \neq s$
ASS #1 to ASS #5 imply that:

- OLS estimator $\hat{\beta}_1$ is *unbiased* and *consistent* estimator of $\beta_1$
- OLS estimators approximately have a normal distribution

**Remarks:**

- ASS #1 is most important
- Extension to multiple $X$’s straightforward

\[ Y_{it} = \beta_1 X_{1it} + \beta_2 X_{2it} + \ldots + \beta_k X_{kit} + \alpha_i + \lambda_t + u_{it} \]

- Additional assumption ASS #5 implies that error terms are uncorrelated over time (no autocorrelation)
• Violation of assumption #5: error terms are correlated over time: $\text{Cov}(u_{it}, u_{is}) \neq 0$

• $u_{it}$ contains time-varying factors that affect the traffic fatality rate (but that are uncorrelated with the beer tax)

• These omitted factors might for a given entity be correlated over time

• Examples: downturn in local economy, road improvement project

• Not correcting for autocorrelation leads to standard errors which are often too low
Fixed effects regression model
Clustered standard errors

• Solution: compute HAC-standard errors (clustered se’s)
  • robust to arbitrary correlation within clusters (entities)
  • robust to heteroskedasticity
  • assume no correlation across entities

• Clustered standard errors valid whether or not there is heteroskedasticity and/or autocorrelation

• Use of clustered standard errors problematic when number of entities is below 50 (or 42)

• In stata: **command, cluster(entity)**
The effect of a tax on beer on traffic fatalities

| Dependent variable: traffic fatality rate (number of deaths per 10 000) |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Beer tax                        | 0.36***         | -0.66***        | -0.64***        | -0.59***        | -0.59*          |
|                                 | (0.06)          | (0.19)          | (0.20)          | (0.18)          | (0.33)          |
| State fixed effects             | -               | yes             | yes             | yes             | yes             |
| Time fixed effects              | -               | -               | yes             | yes             | yes             |
| Additional control variables    | -               | -               | -               | yes             | yes             |
| Clustered standard errors       | -               | -               | -               | -               | yes             |
| N                               | 336             | 336             | 336             | 336             | 336             |

Note: * significant at 10% level, ** significant at 5% level, *** significant at 1% level. Control variables: Unemployment rate, per capita income, minimum legal drinking age.
Panel data: an example
returns to schooling

\[ Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it} \]

- \( Y_{it} \) is logarithm of individual earnings; \( X_{it} \) is years of completed education
- \( \alpha_i \) unobserved ability
- Likely to be cross-sectional correlation between \( X_{it} \) and \( \alpha_i \), hence standard cross-sectional analysis with OLS fails
- However, in this case panel data does not solve the problem because \( X_{it} \) typically lacks time series variation (\( X_{it} = X_i \))
- We have to resort to cross-sectional methods (instrumental variables) to identify returns to schooling
Is there an effect of cigarette taxes on smoking behavior?

\[ Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it} \]

- \( Y_{it} \) number of packages per capita in state \( i \) in year \( t \), \( X_{it} \) is real tax on cigarettes in state \( i \) in year \( t \)
- \( \alpha_i \) is a state specific effect which includes state characteristics which are constant over time
- Data for 48 U.S. states in 2 time periods: 1985 and 1995
Panel data: Cigarette taxes and smoking

\[ \text{Lpackpc} = \log \text{number of packages per capita in state } i \text{ in year } t \]
\[ \text{rtax} = \text{real avr cigarette specific tax during fiscal year in state } i \]
\[ \text{Lperinc} = \log \text{per capita real income} \]

```
. regress lpackpc rtax lperinc
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 96</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1.76908655</td>
<td>2</td>
<td>.884543277</td>
<td>F( 2, 93) = 21.25</td>
</tr>
<tr>
<td>Residual</td>
<td>3.87049389</td>
<td>93</td>
<td>.041618214</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>5.63958045</td>
<td>95</td>
<td>.059364005</td>
<td>R-squared = 0.3137</td>
</tr>
</tbody>
</table>

| Variable | Coef. | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|----------|-------|-----------|------|------|-----------------------|
| rtax     | -.0156393 | .0027975 | -5.59 | 0.000 | -.0211946 to -.0100839 |
| lperinc  | -.0139092 | .158696 | -0.09 | 0.930 | -.3290481 to .3012296 |
| _cons    | 5.206614 | .3781071 | 13.77 | 0.000 | 4.455769 to 5.95746 |
Panel data: Cigarette taxes and smoking

Before-After estimation

. gen diff_rtax= rtax1995- rtax1985
. gen diff_lpackpc= lpackpc1995- lpackpc1985
. gen diff_lperinc= lperinc1995- lperinc1985

. regress  diff_lpackpc diff_rtax diff_lperinc, nocons

| Source |       SS       df       MS              Number of obs =      48 |
|--------+-----------------------------------------------|
| Model  |  3.33475011     2  1.66737506           F(  2,    46) =  145.66 |
| Residual |  .526571782    46  .011447213           Prob > F      =  0.0000 |
| Total  |  3.86132189    48  .080444206           R-squared     =  0.8636 |
|        |]<=-----------------------------------------------|
|        | Number of obs =      48 |
|        | Adj R-squared =  0.8577 |
|        | Root MSE      =  .10699 |

| diff_lpackpc | Coef.   Std. Err.   t     P>|t|      [95% Conf. Interval] |
|-------------+-------------------|-------------------|-------------------|-------------------|-------------------|
| diff_rtax   | -.0169369   .0020119   -8.42   0.000   -.0209865   -.0128872 |
| diff_lperinc| -1.011625   .1325691    -7.63   0.000  -1.278473   -.7447771 |
Least squares with dummy variables (no constant term)

```
. regress lpackpc rtax lperinc stateB*, nocons

          Source |        SS       df       MS                      Number of obs =   96
-------------+---------------------------------------------------------------
           Model |  2094.15728    50  41.883147                      F(  50,    46) =  7317.61
        Residual |   .263285891    46   .005723606                  Prob > F    =  0.0000
-------------+---------------------------------------------------------------
            Total |  2094.42057    96  21.8168809                     Adj R-squared =   0.9997

-------------------------------------------------------------------------------
                  lpackpc |     Coef.    Std. Err.     t    P>|t|     [95% Conf. Interval]
-------------------------------------------------------------------------------
          rtax |  -.0169369    .0020119  -8.42   0.000  -.0209865   -.0128872
         lperinc |  -1.011625    .1325691  -7.63   0.000  -1.278473   -.7447771
           stateB1 |   7.663688    .3037711   25.23   0.000    7.052229    8.275148
           stateB2 |   7.834448    .2926539   26.77   0.000    7.245367    8.423535
           stateB3 |   7.678433    .3121525   24.60   0.000    7.050103    8.306763
           stateB4 |   7.666270    .3392221   22.60   0.000    6.983451    8.349088
           stateB5 |   7.834448    .3193189   24.57   0.000    7.201603    8.487114
           stateB6 |   7.926666    .3154175   25.13   0.000    7.291758    8.561563
           stateB7 |   7.644741    .2936826   26.03   0.000    7.053589    8.235894
           stateB8 |   7.825943    .3275694   23.89   0.000    7.166580    8.485306
```

Panel data: Cigarette taxes and smoking
Least squares with dummy variables with constant term

. regress lpackpc rtax lperinc stateB*

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 96</th>
</tr>
</thead>
<tbody>
<tr>
<td>F( 49, 46) = 19.17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>5.37629455</td>
<td>49</td>
<td>.109720297</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>.263285891</td>
<td>46</td>
<td>.005723606</td>
<td>R-squared = 0.9533</td>
</tr>
<tr>
<td>Total</td>
<td>5.63958045</td>
<td>95</td>
<td>.059364005</td>
<td>Adj R-squared = 0.9036</td>
</tr>
</tbody>
</table>

| lpackpc | Coef.   | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|---------|---------|-----------|-------|------|---------------------|
| rtax    | -0.169369 | 0.020119  | -8.42 | 0.000 | -0.209865 - 0.128872 |
| lperinc | -1.011625 | 0.1325691 | -7.63 | 0.000 | -1.278473 - 0.7447771 |
| stateB1 | -0.3530275 | 0.0900694 | -1.70 | 0.096 | -0.5392832 - 0.1758463 |
| stateB2 | 0.0177322 | 0.1005272 | 0.18  | 0.861 | -0.1846185 - 0.220083 |
| stateB42| -0.771239 | 0.0918679 | -8.40 | 0.000 | -0.9561594 - 0.5863186 |
| stateB43| (dropped) |
| stateB44| 0.1757536 | 0.0854144 | 2.06  | 0.045 | 0.0038233 - 0.347684 |
| stateB45| 0.0276429 | 0.0948094 | 0.29  | 0.772 | -0.1631985 - 0.2184843 |
| stateB46| 0.1099444 | 0.0918156 | 1.20  | 0.237 | -0.0748708 - 0.2947597 |
| stateB47| -0.1719747 | 0.0959042 | -1.79 | 0.080 | -0.3650198 - 0.0210705 |
| stateB48| 0.0092272 | 0.0787188 | 0.12  | 0.907 | -0.1492255 - 0.16768 |
| _cons   | 7.816716  | 0.3458507 | 22.60 | 0.000 | 7.120554 - 8.512877 |
Within estimation

```
. xtreg lpackpc rtax lperinc, fe i(STATE)

Fixed-effects (within) regression
Group variable: STATE

Number of obs = 96
Number of groups = 48

R-sq: within = 0.8636
       between = 0.0896
       overall = 0.2354

F(2,46) = 145.66
corr(u_i, Xb) = -0.5687
Prob > F = 0.0000

------------------------------------------------------------------------------
 lpackpc  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      rtax  |  -.0169369   .0020119    -8.42   0.000    -.0209865   -.0128872
     lperinc |  -1.011625   .1325691    -7.63   0.000    -1.278473   -.7447771
        _cons |   7.856714   .3150362    24.94   0.000     7.222579    8.490849
-------------+----------------------------------------------------------------
     sigma_u |  .25232518
     sigma_e |  .07565452
        rho  |   .91751731 (fraction of variance due to u_i)
-------------+----------------------------------------------------------------
F test that all u_i=0:   F(47, 46) = 13.41
Prob > F = 0.0000
```