Empirical Example: Panel Data Model

We are interested in how beer tax affects fatality rates on roads. The data file fatality.csv contains (balanced) panel data for annual observations of 48 US states from year 1982 to 1988. The description of variables is below

- statename: state (a string variable)
- year: year
- unrate: state unemployment rate
- beertax: beer tax
- mlda: minimum legal driving age
- pop: total population
- pop1715: population with age of 15-17
- unus: US unemployment rate
- y: fatality per 10000 people

First we use the 1982 data only. That means the data are cross sectional. Regressing y onto beertax using OLS gives an insignificant coefficient estimate of .148. The positive sign is unexpected, implying that rising beertax will cause more fatality. By using the 1988 data, we get still positive but significant coefficient of .438. If we ignore the special structure of the panel data and treat it as pooled cross sections (1982 cross section plus 1983 cross section plus...), we get positive estimate of .364, between the 1982 and 1988 estimates. We know omitted variable may generate this misleading positive effect.

Panel data are special because we have repeated observations for the same state (entity) over time. Therefore we can consider differenced data. Hopefully the process of time differencing can remove the time-invariant unobserved factor, $a_i$. More explicitly, suppose the true models for 1982 and 1988 are

\begin{align}
y_{i,t=1982} &= \beta_0 + \beta_1 x_{i,t=1982} + a_i + u_{i,t=1982} \\
y_{i,t=1988} &= \beta_0 + \beta_1 x_{i,t=1988} + a_i + u_{i,t=1988}
\end{align}

(1)
where \( a_i + u_{it} \) is the composite error term. Note that \( a_i \) has no time subscript, so denotes time-invariant unobserved factors such as people’s attitude toward drunk driving. \( u_{it} \) is time-varying and state-varying, and is called idiosyncratic error. In general we believe

\[
E(x_i a_i) \neq 0,
\]

so the key regressor \( x \) is correlated with the time-invariant unobserved factors, and there is endogeneity issue. As a result, the pooled OLS estimator or OLS estimator applied to a cross section in a specific year is biased. To remove \( a_i \), let

\[
\Delta y_{i6} \equiv y_{i,t=1988} - y_{i,t=1982} \tag{3}
\]

\[
\Delta x_{i6} \equiv x_{i,t=1988} - x_{i,t=1982} \tag{4}
\]
denote the six-year-difference. Then subtracting (1) from (2) leads to

\[
\Delta y_{i6} = \beta_1 \Delta x_{i6} + \Delta u_{i6} \tag{5}
\]
The (six-year) difference estimator is the OLS estimator \( \hat{\beta}_1 \) in the differenced regression (5). Because \( a_i \) is absent in (5), the difference estimator is consistent provided that

\[
E(\Delta x_{i6} \Delta u_{i6}) = 0 \tag{6}
\]

If assumption (6) fails, we need to find IV for \( \Delta x_{i6} \). The STATA commands to obtain the differenced data and run differenced regression (5) are

```
encode statename, gen(id);
sort id year;
by id: gen dy6 = y[_n]-y[_n-6];
by id: gen dx6 = beertax[_n]-beertax[_n-6];
reg dy6 dx6 if year==1988, r;
```

where we convert the string statename into a numeric id, and we use 1988 data only because other years have missing values. In this case the six-year difference estimate is -1.04 with p-value of 0.005. So rising beer tax reduces fatality significantly. This result is unbiased no matter whether \( x \) and \( a \) are correlated, but we do require assumption (6) holds.

Because we have 7 years of observations for each state, another estimator that is better than pooled OLS is the fixed effect estimator. The FE estimator is also called within
estimator, because it is based on the within regression

\[
\ddot{y}_{it} = \beta_1 \ddot{x}_{it} + \ddot{u}_{it}
\]  

(7)

where the double dots denote the time-demeaned variable:

\[
\ddot{y}_{it} = y_{it} - \bar{y}_i, \bar{y}_i = T^{-1} \sum_{t=1}^{T} y_{it}
\]  

(8)

\[
\ddot{x}_{it} = x_{it} - \bar{x}_i, \bar{x}_i = T^{-1} \sum_{t=1}^{T} x_{it}
\]  

(9)

\[
\ddot{u}_{it} = u_{it} - \bar{u}_i, \bar{u}_i = T^{-1} \sum_{t=1}^{T} u_{it}
\]  

(10)

Basically the time-demeaned variable measures the variation within each panel (the 7 years observations for each state is a panel). Because \( a_i \) is time-invariant, it follows that

\[
\ddot{a}_i = a_i - \bar{a}_i = a_i - a_i = 0.
\]

Hence \( a_i \) is absent in the within regression (7). The FE (within) estimator is the OLS estimator \( \hat{\beta}_1 \) in the within regression (7), and is consistent provided that

\[
E(\ddot{x}_{it} \ddot{u}_{it}) = 0
\]  

(11)

Instruments are needed if assumption (11) fails. The STATA commands to fit the within regression (7) are

\begin{verbatim}
xset id year;
xtrreg y beertax, r fe;
areg y beertax, absorb(id) r;
\end{verbatim}

So we declare the panel set. Next we can use either xtreg (with option fe) or areg (with option absorb) to obtain the FE estimate. In this case, the FE estimate is -.655, smaller (in absolute value) than the six-year difference estimate, with p-value of .029, so significant at 5% level. The FE estimate is unbiased if assumption (11) holds, regardless of whether \( x \) and \( a \) are correlated. The nice thing about xtreg and areg is, they automatically carry out the within transformation and generate the time-demeaned variable.
Because in general $u_{it}$ is serially correlated, it is important to add option r in xtreg or areg so that the cluster-robust standard error is used.

According to the Frisch-Waugh theorem, the FE estimator can be equivalently obtained in the dummy variable regression (DVR)

\[ y_{it} = \beta_0 + \beta_1 x_{it} + \sum_{j=1}^{n-1} c_j D_{ji} + \text{error} \]  

where $D_{ji}$ is the state (entity) dummy defined as

\[ D_{ji} = \begin{cases} 
1, & \text{state } j; \\
0, & \text{o/w.} 
\end{cases} \]  

You can think of the state dummy as the perfect proxy variable for time-invariant unobserved factor $a_i$. Notice that when the intercept term is included we need to drop one state dummy in order to avoid multicollinearity. The STATA commands to generate state dummies and fit DVR (12) are

```stata
qui tab state, gen(sd);
reg y beertax sd1-sd47, vce(cluster id);
```

The coefficient estimate of $\hat{\beta}_1$ in DVR is (and should be) numerically equivalent to the FE estimate, though the standard error differs. In DVR $a_i$ is not treated as an unobserved random factor, instead, it is treated as the state-specific intercept term. That is why we include state dummies in (12).

The stata commands to obtain time-demeaned variables and run within regression manually are

```stata
bysort id: egen ybar = mean(y);
bysort id: gen dmy = y - ybar;
bysort id: egen xbar = mean(beertax);
bysort id: gen dmx = beertax - xbar;
reg dmy dmx, noc;
```

where we use command egen because we want to get the time mean $\bar{y}_i = T^{-1} \sum_{t=1}^{T} y_{it}$ for each $i$, not the overall mean $\bar{y} = (nT)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} y_{it}$. The command reg dmy dmx, noc produces correct coefficient estimate, but incorrect standard error, t-value and p-value.
The above commands may help if we need some intermediate results. For instance, xtreg reports \(\sigma_u\), which is the standard error of state fixed effect \(a_i\). It can be obtained by running stata commands of

```stata
gen ai = ybar - (-0.6558746)*xbar;
list ai;
sum ai if year==1988;
```

The ai should be the same as the state-specific intercept terms reported by the below DVR

```stata
reg y beertax sd*, noc vce(cluster id);
```

xtreg also reports \(\sigma_e\), which is the standard error of \(u_{it}\). It can be obtained by running commands of

```stata
gen uhat = y - ai - (-0.6558746)*beertax;
gen uhat2 = uhat^2;
qui sum uhat2;
dis "sigme is " sqrt(r(sum)/(336-1-48));
```

Note that the degree of freedom is \(nT - k - n\). The rho reported by xtreg is \(\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}\).

We may also consider a model with both time-invariant and entity-invariant unobserved factors

\[
y_{it} = \beta_0 + \beta_1 x_{it} + a_i + \theta_t + u_{it}, \quad E(x_{it}a_i) \neq 0, E(x_{it}\theta_t) \neq 0
\]  

(14)

where the state fixed effect \(a_i\) denotes unobserved factors that vary across entities but are invariant to time; time fixed effect \(\theta_t\) denotes unobserved factors that vary over time but are invariant to entities (such as national trend). In the one-level model, \(a_i\) can be approximated by entity dummies. In a similar fashion, in the two-level model (14), \(\theta_t\) can be approximated by time dummies. The stata commands to generate time dummies and obtain the two level FE estimator are

```stata
qui tab year, gen (yd);
areg y beertax yd*, absorb(id) r;
```

and for this problem, the two level FE estimate is -.639, slightly smaller (in absolute value) than the one-level FE estimate -.655. This fact indicates that for this problem the unobserved time-varying factors may be unimportant, consistent with \(\rho = .934\) reported by xtreg.

To summarize, we find evidence that rising beer tax significantly lowers the fatality rate. The six-year difference estimate is the biggest because it considers the long run effect of tax.
Summary of Panel Data Models

The idea of various panel data models can be summarized as follows. We assume the error term has three components:

\[ e_{it} = a_i + \theta_t + u_{it} \]  

(15)

where \( a_i \) is time-invariant, \( \theta_t \) is entity-invariant, and idiosyncratic error \( u_{it} \) is time-varying and state-varying. The key is to remove \( a_i \) and \( \theta_t \) since they are potentially correlated with \( x \). By using the one-level FE estimator or difference estimator we can remove \( a_i \); by using the two-level FE estimator (with both entity and time fixed effects) we can remove \( a_i \) and \( \theta_t \). However, the consistency of both FE estimator and difference estimator requires that the regressor be uncorrelated with idiosyncratic error \( u \). If \( x \) and \( u \) are correlated, endogeneity issue still exists (but with smaller impact than pooled OLS) and instrumental estimation is needed.

It is much easier to find valid IV if we have panel data. After obtaining the time-demeanded or differenced data, the instrument \( z \) should satisfy

\[ E(z\bar{x}) \neq 0, E(z\bar{u}) = 0 \]  

(16)

\[ E(z\Delta x) \neq 0, E(z\Delta u) = 0 \]  

(17)

In other words, in the setting of panel data model, instrument \( z \) should be uncorrelated with idiosyncratic error \( u \) only. \( z \) can be correlated with \( a_i \) or \( \theta_t \).

There is another (popular but ad hoc) way to handle the time-varying and state-varying unobserved factors. Consider an entity-specific trend, which is the interaction term of entity-dummy and a linear trend

\[
\text{entity-specific trend}_j = Dj \ast t
\]  

(18)

where \( Dj \) is specified in (13). The entity-specific trend works well if \( u_{it} \) is indeed trending linearly. After first-differencing, the entity-specific trend becomes entity-dummy:

\[
Dj \ast t - Dj \ast (t - 1) = Dj
\]

This fact suggests that we need to estimate the differenced equation by fixed effect. See computer exercise C14.4 in the textbook for an example.