**Lab Assignment 10 - Week 12**

**I. Objectives**: The objectives are to explore methods of analyzing panel data and time-series cross-section data. For the latter, you will run traditional fixed-effects and random-effects models, but can also examine models with lagged dependent variables, and models that attempt to control for serial correlation using feasible generalized least squares.

**II. Datasets**: *fatality.dta*

**III. Packages**: *foreign*, *lattice*, *lmtest*, *pcse*, *plm*, *tseries*

**IV. Preparation**

1) Open RStudio by double-clicking the icon or selecting RStudio from the Windows Start menu.

2) Open the POLS6481-Spring2021-UH-lab Project and perform a Git pull.

3) If not using Projects, Git, and *here* download files, place in working directory, and make changes as needed.

4) Open the R script by typing *Ctrl*+*O* or by clicking on File in the upper-left corner, using the dropdown menu, and navigating to the script in your working directory.

5) Open the R script by typing *Ctrl*+*O* or by clicking on File in the upper-left corner, using the dropdown menu, and navigating to the script in your working directory.

6) If you are missing any packages from the script, you should see a message similar to the one in the screenshot below (this screenshot is also available full size in the lab folder). Click install or install any needed packages manually:

Graphical user interface

Description automatically generated

6) Run lines 1–7 in the R script to load five packages that you will need. Commands to install packages *plm*  and *pcse* are included in the comments for lines 5 and 7.

**V. Instructions for Lab 12**

The dataset that you will be using is a balanced panel, with traffic fatalities in every year from 1982 through 1988, for 48 states. Because there are so many options for estimating regression models with panel data, it usually is best to start with a focused research question, and then choose the model that will best assist you to answer that question.

Your question will be: Can states reduce drunk driving by raising taxes on beer? This approach to policy-making is called a “Pigouvian” tax, akin to taxing pollution in order to deter polluters.

Load the dataset by running line 9 or typing the following code, changing the directory if needed:

> data <- read.dta("C:/Users/Scott/fatality.dta")

The data frame that you will work with is simply named data.

Run line 10 of the R script to set the data as panel data, telling R that the state variable differentiates the cross-sectional entities and that the year variable differentiates the time points. Run line 11 to examine the panel data structure. This is a *balanced panel* because it has observations for every state in every year.

Run line 13 to create the dependent variable, multiplying values of the variable *mrall* by 10,000 to create the variable *vfrall*.

Run lines 16–19 to generate a figure that plots fatality rates against beer taxes in 1982, including a regression line. Next, run lines 21–24 to generate a figure that plots fatality rates against beer taxes in 1988. Note that additional code has been included in line 17 and line 22 to ensure that the two plots have the same axis dimensions and thus are more directly comparable.

Describe the overall relationship between the level of beer taxes and the number of vehicular fatalities (per 10,000 population). Is the direction of the relationship surprising to you? (*Hopefully you are surprised and provoked to think about why cross-sectional policy data presents challenges that panel data analysis can address.*)

**A. Pooled Cross-Sectional Model**

The first stage of the analysis will examine all seven years in the data set (1982 through 1988) and add a few control variables. Run line 27 to generate the natural log of real income per capita in the state, and run line 28 to re-scale average vehicle miles driven per driver.

Run line 30 to estimate the regression of vehicular fatalities on beer taxes, income, and miles driven for all seven years. When you run line 31 to display the results you should observe that higher levels of the beer tax are associated with higher levels of fatalities (although not significantly), which contradicts logic and the “differences” results above. Fill in the coefficients and standard errors in the table on page 6 of this worksheet.

Run line 33 to generate the lagged value of fatalities, and run line 34 to add the lagged dependent variable to the model estimated previously. While this addition dramatically improves the model’s R2 (from .39 to .89), what happens to the effect of beer taxes? You should comment on both the coefficient and the standard error…

Obviously the cross-sectional analysis (with or without the lagged dependent variable) does not get us the “right” answer to our question. It’s important to note that beer taxes do not change much over time. To verify this, run line 36 to create the lagged value of beer taxes, and run line 37 to examine the correlation between beer taxes at *t* and at *t*–1.

One possibility that might explain the counter-intuitive results above is that beer taxes are raised in reaction to higher vehicle fatalities. You can examine this possibility by running line 39. What do you observe?

**B. Fixed Effects Model**

The ‘fixed effects’ model can be conceived of as allowing the intercept to shift for each cross-sectional unit. In this example, this means estimating a separate intercept for each state, which is constant across the seven years in the model. This is one way of handling unobserved heterogeneity, i.e., an unexplained yet consistent difference in vehicular fatalities across states.

We are going to run the fixed effects model two ways:

1. Run lines 42−43 to estimate the fixed-effects regression using the factor() command. This model includes dummy variables for every state, which are shown in the model summary, which is why this is called the least squares dummy variables (LSDV) model. R automatically dropped the first state in the list (AK, or Alaska) in order to ensure we don’t fall into the “dummy variable trap.”

The model does not provide an *F* test of their joint significance, although you could compute one; the best way would be to compare the *R*2 (which is sky-high) to the *R*2 of line 58. The *F* test will have 47 numerator degrees of freedom (one for each cross-sectional unit) and 285 denominator degrees of freedom (336 observations, minus 47 dummies, minus 4 coefficients estimated - including the intercept). You can perform such a test using the anova() command as shown in line 44. The F statistic is 35.77, which has a highly significant *p* value, indicating that the 47 cross-sectional unit dummy variables are jointly significant.

1. Run lines 46–47 to use the *plm* package to estimate the “within” model. The “within” estimator regresses deviations from the average value of the dependent variable (within each state) on deviations from the average values of each independent variable (within each state). You should notice that the coefficients and standard errors for *beertax*, *lincperc*, and *vmilespd* are all identical to the result of a).

Run line 48 to carry out a Breusch-Pagan test, which is a Lagrange multiplier test (yielding a chi-square statistic) of heteroskedasticity. When we were using cross-section data, this test was carried out with the bptest() command in the *lmtest* package.

Run lines 50–51 to estimate the “pooling” model, which is simply the static model applied to panel data. (You can type summary(base) to verify this.) Run line 52 to display the result of a test of the joint significance of the 47 cross-sectional unit dummy variables. This is simply another way to get to the same result that you had after running line 44.

Aside: We have not used the between-groups estimator, which regresses the average value of the dependent variable in each state on the average values of each independent variable in each group / state. If you replaced "pooling" with "between" in line 50, then you would run the between estimator. If you replaced "pooling" with "fd" in line 50, then you would run the first-difference estimator, but this is rarely justified in time-series cross-sectional analysis.

**C. Random Effects Model**

The key random effects assumption is that the correlation between the fixed effects and the included regressors (*beertax*, *lincperc*, and *vmilespd*) is equal to zero; the fixed effects model should provide you with information about whether such an assumption would be appropriate – however I’m not able to find this information in R! In this example, the assumption is not appropriate – I recall from working these problems in Stata – however we will estimate the random effects model for practice.

Run lines 55–56 to estimate the random effects model. Note that you can include variables that do not vary over time in a state with random effects models. For instance, if you thought Southern and Midwestern states were distinct from coastal states, then a random effects model would allow you to add dummy variables for regions. Variables such as these are dropped automatically from fixed effects models, because they are perfectly multicollinear with a subset of the dummies / intercept shifts.

Compare the coefficients on *beertax* in the fixed-effects and random-effects models, and make sure to fill in both sets of coefficients (and standard errors) in the table on page 6.

In the next-to-last line of the table on page 6, also write down the value of *theta* (θ) which we will discuss in class. When *theta* is close to 0, the random effects model is close to a pooled cross-section model; when *theta* is close to 1, the random effects model is close to a fixed effects model. What value does R report for this model?

Run line 57 to conduct the Hausman test, which helps us to choose between fixed and random effects. Under the null hypothesis, the random effects estimator is efficient, and both the fixed- and random-effects are consistent (so the estimates should not differ systematically). Under the alternative hypothesis, the random effects estimator is inconsistent; fixed effects are preferred if we reject the null hypothesis. A low *p* value indicates rejecting the null hypothesis in favor of the alternative hypothesis. What is the value of *p* here? What do we conclude from this test: fixed- or random- effects?

**D. Other Models**

In a very famous article, “What to do and not to do with time-series cross-section data,” Beck and Katz criticize Feasible GLS estimators, and introduce “panel-corrected standard errors.” Lines 60 and 61 show you versions of the model with panel-corrected standard errors using the *pcse* package; note that the input must be an *lm* object, so we could use the model *fixed1* (which was LSDV) or *base* (which was simple OLS). I recommend running the latter line of code because with fewer variables, the results will be much easier to inspect. The *pcse* command gives you the variance-covariance matrix of the estimators, the panel-corrected standard errors, the estimated coefficients (which are identical to what the model summary reported) and the t-statistics, recomputed using the panel corrected s.e.

Lines 63−65 show you a version of the GLS estimator using the *pggls* package. This was popular a few decades ago, but now it seems that scholars prefer other methods.

Before you exit, line 68 adds another package, *lattice*, which allows you to generate some interesting plots. (The installation command is in the comments for Line 68). In line 69, you can plot the dependent variable over time, for each cross-sectional unit (state) independently. This can allow you to see whether the cross-sectional units tend to move together over time, in which case you might wish to include fixed- or random-effects for time units (which is not difficult, but is unnecessary here) instead of or in addition to cross-sectional units.

To clear the **Environment**, type rm(list=ls()) or click on the broom icon. To clear the **Console** window, type Ctrl-*l*

Regression coefficients and standard error

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Static model  line 59 / 80 | Lagged y  line 62 | Fixed Effects  line 72 / 76 | Random Effects  line 85 |
| Intercept | ( ) | ( ) | ( ) | ( ) |
| Beer tax | ( ) | ( ) | ( ) | ( ) |
| Income | ( ) | ( ) | ( ) | ( ) |
| Miles | ( ) | ( ) | ( ) | ( ) |
| yt-1 |  | ( ) |  |  |
| R2 |  |  |  |  |
|  |  |  |  | *theta* = |
|  |  |  | *Hausman test*: χ2 =  *p* = | |