# Problem Set 4

*Prof. Conlon*

*Due: 3/29/19*

## Packages to Install

**The packages used this week are**

- estimatr (Tidyverse version of lm function)
- plm (Panel Data package)

## Problem 1 (Analytical Exercise)

Assume we have the following dynamic panel model:

$$y_{i,t} = \gamma y_{i,t-1} + \alpha_i + \epsilon_{i,t},$$

We have the following assumptions:

$$|\gamma| < 1$$
$$y_{i,0} \text{ is known (i.e. non-random)}$$
$$\mathbb{E}[T]\epsilon_{i,t}|y_{i,t-1}, ..., y_{i,0}, \alpha_i = 0 \text{ (Sequential Exogeneity)}$$

Our goal is to consistently estimate $\gamma$.

1. Show that the first-difference estimator of $\gamma$ is unbiased if $\alpha_i = 0$.
   (Hint: To show unbiased, show that the $\mathbb{E}[\hat{\alpha})i] = \alpha_i$.)

2. Is the first difference estimator of $\gamma$ still unbiased if $\alpha_i \neq 0$?
   (Hint: Try showing that $cov(\epsilon_{i,t}, y_{i,t+1}) \neq 0$. Why is this condition useful?)

3. Let $\Delta y_{i,t} = y_{i,t} - y_{i,t-1}$, what is $cov(\Delta \epsilon_{i,t}, y_{i,t-2})$? What is $cov(\Delta y_{i,t}, y_{i,t-2})$?

4. Using your answer from part (4), propose a strategy to consistently estimate $\gamma$?
   (Hint: The estimator will be a 2SLS (or IV regression).)

## Problem 2 (Coding Exercise)

For this exercise we will be using data on cigarette sales available in the **plm** package. To load this dataset, simply use the following command:

The data has the following columns:

- **sales** – cigarette sales in packs per capita

- **pimin** – minimum price in adjoining states per pack of cigarettes

- **ndi** – per capita disposable income

- **cpi** – consumer price index (1983=100)

- **pop16** – population above the age of 16

- **pop** – population

- **price** – price per page of cigarettes
- **year** – the year
- **state** – number for the state

We will be estimating a series of panel models trying to estimate the price elasticity of demand.

1. The price elasticity of cigarette demand, $E_d$ is the percentage change of sales of cigarettes divided by the percentage change of price of cigarettes. Let $S_t$ be sales of cigarettes at time $t$ and $p_t$ be price of cigarettes at time $t$, then the price elasticity of demand is:

$$E_d = \frac{\Delta S_t / S_t}{\Delta p_t / p_t}$$

   To estimate price elasticities in a linear regression, turn sales and price into logs. Why do we do this?

2. Assume that we are estimating the following cross-sectional regression:

$$log(S_{i,t}) = \gamma + \beta log(p_{i,t}) + \alpha_1 ndi_{i,t} + \alpha_2 cpi_{i,t} + \alpha_3 pop16_{i,t} + \epsilon_{i,t}$$

   Estimate this regression, reporting the price elasticity of demand? What is potentially wrong with this regression?

3. One important issue could be heterogeneity in sales across states, $\gamma$ varies across states. However, we may not know whether we want to model this as a fixed effect or a random effect. One test often used is the Hausman-Wu test, available as phtest in the plm package. Estimate a within estimator, a random effect estimator and then use a Hausman-Wu test to report which specification is the correct one. Be sure to understand what the null and alternative hypothesis of this test are in order to report the correct interpretation of the test.

4. We begin to believe that not only heterogeneity in mean sales, but dynamics are important, so we decide to estimate the following model:

$$log(S_{i,t}) = \gamma_i + \tau log(S_{i,t-1}) + \beta log(p_{i,t}) + \alpha_1 ndi_{i,t} + \alpha_2 cpi_{i,t} + \alpha_3 pop16_{i,t} + \epsilon_{i,t}$$

   We are first going to have you implement the Anderson-Hsiao estimator:

5. Create a vector of first differences for each term in the regression above.

6. Let $\Delta S_{i,t} = log(S_{i,t}) - log(S_{i,t-1})$, we will be running the following two-stage least squares regression. In your first stage, use $log(S_{i,t-1})$ as an instrument for $\Delta S_{i,t}$, then run a 2SLS to get a new estimate for the price elasticity. Report any differences.

   An issue with the Anderson-Hsiao estimator is that it is not efficient because it does not use all available information. The reason is quite simple, if $log(S_{i,t-1})$ is a valid instrument for $\Delta S_{i,t}$ and $log(S_{i,t-1})$ depends on $log(S_{i,t-2})$ then $log(S_{i,t-2})$ is also a valid instrument for $\Delta S_{i,t}$. This idea is the logic underpinning the Arellano-Bond estimator. This estimator is quite hard to code up, but luckily plm includes this estimator as the pgmm function.

7. Using a generalized method of moments estimator (Arellano-Bond), re-estimate the model. Report any differences.