KAN-CEAPV2505U

Econometric Analysis of Firm Data

July 30, 2025

Copenhagen Business School
Department of Economics
Ralf A. Wilke
Mette Franck
2025/2026, Autumn semester

Problem Set 5: Panel Data Models

1. Marketing firms are interested in knowing how potential customers respond to the design of logos. In neuromarketing, the effect of a stimulus is measured in various ways, including facial coding to categorize the physical expression of emotion. This exercise uses facial expression data on browfurrow, joy and facial asymmetry that was recorded in an experimental setting. 27 individuals have been presented a sequence of four new draft logos of a business association and various distractors. We have data for each so called stimulus for the first 10 seconds they are shown to participants. While the distractors are not supposed to result in a physical expression, the goal of the exercise is to test whether the draft logos lead to any response. The data (motion.dta) is kindly being provided

by Imotions A/S. For an explanation of the measures of physical expression see readme.rtf. The responses for each individual are plotted for the 4 logos in Figures browfurrow_logoX.pdf, joy_logoX.pdf and frontalasymmetry_logoX.pdf.

(a) On the grounds of the data available, how would you estimate the base effect of showing any distractor and the effect of showing the 4 logos? Write down the panel data model.

(b) Estimate the model by POLS.

(c) Estimate the model by FD for all individuals and by stratifying by gender. Why is FD and not FE the preferred panel model in this application? Compare your results with your POLS results.

2. Use the data in GPA3.dta for this exercise. The data set is for 366 student-athletes from a large university for autumn and spring semesters. Because you have two terms of data for each student, an unobserved effects model is appropriate. The primary question of interest is this: Do athletes perform more poorly in school during the semester their sport is in season?

   (a) Use pooled OLS to estimate a model with term GPA ($trmgpa$) as the dependent variable. The explanatory variables are $spring$, sat, $hsperc$, $female$, $black$, $white$, $frstsem$, $tothrs$, $crsgpa$, $season$. Interpret the coefficient on $season$. Is it statistically significant?

   (b) Most of the athletes who play their sport only in the fall are football players. Suppose the ability levels of football players differ systematically from those of other athletes. If ability is not adequately captured by SAT score and high school percentage, explain why pooled OLS estimators will be biased.

   (c) Now, use the data differenced across the two terms. Which variables drop out? Now, test for an in-season effect.

   (d) Can you think of one or more potentially important, time varying variables that have been omitted from the analysis?

3. Use the data in airfare.dta for this exercise. Load the data and explore its content (variables, structure). We are interested in estimating the determinants of the price for airfares. Our model is

$$log(fare_{it}) = \theta_t + \beta_1 concen_{it} + \beta_2 log(dist_i) + \beta_3 [log(dist_i)]^2 + a_i + u_{it}, \quad t = 1, ..., 4,$$

where θ_t means that we allow for different year intercepts.

(a) Estimate the above equation by pooled OLS, being sure to include year dummies. If $\Delta concen = 0.10$, what is the estimated percentage increase in $fare$?

(b) What is the usual OLS 95% confidence interval for β_1? Why is it probably not reliable? Find the heteroscedasticity robust and "fully" (heteroscedasticity and serial correlation) robust 95% CI for β_1. Compare it to the usual CI and comment. Hint: for heteroscedasticity robust standard errors use the option ", robust", for fully robust standard errors use the option ", robust cluster(id)"

(c) Describe what is happening with the quadratic in $log(dist)$. In particular, for what value of $dist$ does the relationship between $log(fare)$ and $dist$ become positive? Is the turning point outside the range of the data?

(d) Now estimate the equation using fixed effects. What is the FE estimate of β_1?

(e) Name two characteristics of a route (other than distance between stops) that are captured by a_i. Might these be correlated with $concen_{it}$?

These problems are partly taken from the Wooldridge (2020) textbook.