# Homework Assignment 4

Max H. Farrell – Chicago Booth
BUS41100 Applied Regression Analysis

*Due at the beginning of class of week 6*

## 1 Grunfeld Data

The data set `Grunfeld.csv` contains data from 11 firms over 20 years (1935–1954), and for each firm/year we observe:[1]

| | |
|---|---|
| $I_{i,t}$ = `invest` | gross investment |
| $V_{i,t}$ = `value` | market value of the firm at the end of the previous year |
| $C_{i,t}$ = `capital` | value of the stock of plant and equipment at the end of the previous year |

Often, we use the notation $n$ for the total sample size, $N$ for the per-time sample size, and $T$ for the time periods. So here we have $N = 11$, $T = 20$, and $n = NT = 220$ total observations.[2] The key is that we do not have 220 total independent observations. We have less *information* than that (remember we think of observations as noisy signals).

Our goal is to study investment as a function of the other two. We think of `value` for firm $i$ and time $t$ as anticipated profit, and then the investment for that firm that year is the amount of replacement investment required.

A first pass at the model is therefor:

$$I_{i,t} = \beta_0 + \beta_1 V_{i,t} + \beta_2 C_{i,t} + \varepsilon_{i,t}. \tag{1}$$

**(a)** Estimate the regression model (1). Interpret the coefficient estimates.

**(b)** In this context, what are the assumptions that make this approach valid? Which of these may be violated and why?

But we believe that different firms have substantively/qualitatively different decision-making styles. Different firms are fundamentally different, and so they are not part of the same phenomenon, and it does not make sense to entirely pool the data together. This is exactly what motivates the *fixed effects* model:

$$I_{i,t} = \beta_{0,i} + \beta_1 V_{i,t} + \beta_2 C_{i,t} + \varepsilon_{i,t}, \tag{2}$$

which lets each firm have it's own intercept. We *can't* care about the fixed effects specifically: there are $N$ dummy variables in the model (2), but because each firm's behavior is dependent over time, we only have $N$ truly independent pieces of information. So we don't even have enough information to precisely nail down these estimates! In other words, in the estimated version of (2),

$$\hat{I}_{i,t} = b_{0,i} + b_1 V_{i,t} + b_2 C_{i,t},$$

[1] This is quite a famous data set in economics, and is regularly used a textbook example. It is also quite old, and quite small.

[2] This is called a *balanced* panel because all units are observed at all time points. In real life, unbalanced panels are more common, where some units are not observed at certain times, either they drop out entirely, drop out then come back, or are added to the sample.

the estimates $b_1$ and $b_2$ are good estimates for $\beta_1$ and $\beta_2$, just like always, but $b_{0,i}$ is *not* a good estimate for $\beta_{0,i}$. For example, the sample distribution result from week 2 will *not* be true: $b_{0,i} \not\sim \mathcal{N}(\beta_{0,i}, \sigma^2_{b_{0,i}})$. Nonetheless, because $b_1$ and $b_2$ are good estimates for $\beta_1$ and $\beta_2$, we can estimate the model anyway.

**(c)** Estimate the regression model (2). Interpret the coefficient estimates on `value` and `capital`. Compare your findings, both the coefficient estimates and the standard errors, to part **(a)**. Are you assuming $\text{cor}(\beta_{0,i}, V_{i,t} + C_{i,t})$ is zero or not? Justify your choice statistically.

We might instead believe that although all firms behave similarly, different years are different (perhaps due to government regulation or tax incentives). This motivates a very similar model, but with time fixed effects instead:

$$I_{i,t} = \beta_{0,t} + \beta_1 V_{i,t} + \beta_2 C_{i,t} + \varepsilon_{i,t}, \tag{3}$$

**(d)** Estimate the regression model (3). Interpret the coefficient estimates on `value` and `capital`. Compare your findings, both the coefficient estimates and the standard errors, to parts **(a)** and **(c)**. Are you assuming $\text{cor}(\beta_{0,t}, V_{i,t} + C_{i,t})$ is zero or not? Justify your choice statistically.

Of course, we can combine the two ideas easily enough, and have both time and firm fixed effects:

$$I_{i,t} = \beta_{0,i} + \beta_{0,t} + \beta_1 V_{i,t} + \beta_2 C_{i,t} + \varepsilon_{i,t}, \tag{4}$$

Now each year and each firm have a specific effect in the model: there are 30 dummy variables!

**(e)** Estimate the regression model (4). Interpret the coefficient estimates on `value` and `capital`. Compare your findings, both the coefficient estimates and the standard errors, to parts **(a)**, **(c)**, and **(d)**.

These are all very common models for panel data. Depending on the particular context/example, one of (2), (3), or (4) might make more sense. If your project ends up involving panel data, consider these models carefully in your example.

## 2 Pricing Experiment

Return to the pricing experiment we studied in class. There, we showed that an increase in price *caused* lower purchasing but also higher profits. Therefore, if the firm wanted to increase profits, they should raise their price to \$249. Now we will explore this further by discussing uncertainty and targeted treatments.

**(a)** How confident are you that raising the price to \$249 will yield an increase in profits?

The data includes a variable `customerSize` that gives the size of the customer firm (remember, the customers of this business are themselves firms). The sizes are ranked 0, 1, 2, for small, medium, and large firms.

**(b)** Using a *single* regression (i.e. one `lm()` command), compute the profit effect for each firm size individually. Compare these to the overall effect we found in class. Does this pattern make sense to you?

**(c)** Using these results, decide on the optimal pricing strategy to maximize profits when the service can charge different prices to different customers based on their size. We are imagining that when a firm goes to the service, they first fill out several questions, including their firm size, and then are shown a price based on these answers. (This is known as *third-degree* price discrimination.)

**(d)** Can the recruiting service improve its profit? By how much? *(Careful computing the profit from your strategy. When using the data, think about which observations were exposed to which price, and how many of each type of firm you have.)*

## 3 Price Elasticity and Cheese

This question considers sales volume as well as price and display activity for packages of Borden Sliced Cheese. The data, available as `cheese.csv` on the course site, are taken from Rossi, Allenby, and McCulloch's *Bayesian Statistics and Marketing*. For each of 88 stores (`store`) in different US cities, we have repeated observations of the sales volume (`vol`, in terms of packages sold), unit price (`price`), and whether the product was advertised with an in-store display (`disp` = 1 for display).

Answer the following questions in a clear and concise manner. Include the appropriate plots and hypothesis tests to illustrate and support your conclusions. Present your solutions as though you are a consultant seeking to inform and convince a (very statistics savvy) client of your results.

**(a)** Ignoring `price`, do the in-store `disp`lays have an affect on log `sales`? Is there reason to suspect that your result is confounded by pricing strategies?

**(b)** A better question: is `price` elasticity for Borden cheese affected by the presence of in-store advertisement?

**(i)** Test this by running two separate regressions. Note that testing if one value is equal to another is the same as testing if the difference is equal to zero. Also, if $b$ and $b^\star$ are least squares coefficients from *independent* regression fits, then $\text{sd}(b - b^\star) = \sqrt{s_b^2 + s_{b^\star}^2}$ (and $n$ is big enough that the $b$'s are all normal).

**(ii)** How can you test this with only one regression? Is the result the same?

**(c)** Do you have a possible economic explanation for your results in **(b)**?

**(d)** Parts **(a)** and **(b)** both say "affect". Write a justification that your regression models truly capture a causal effect. Then, discuss how these results can be used even if causality fails.

## 4 Can observational studies replicate experiments?

In class we studied data from the National Supported Work (NSW) experiment. Men were randomized to either receive job training (the treatment) or not (control). We found the average treatment effect was \$1,794.34: treated men could expect to earn this much more than controls. What if we didn't have a randomized experiment? Could we still estimate this *causal* effect from *observational* data? That is, without an experiment, we want to assess the impact of job training on earnings. Answering these questions is the goal of this problem.

Suppose that the same 185 men received job training, and we want to know what effect this training had on income, but *we do not have access to the NSW control sample*. That is, no experiment was done. We need a control sample to compare too, and for this we have 2490 men drawn from the Panel Study of Income Dynamics (PSID). This is now observational data. The full data consists of the 185 treated men and these 2490 men to act as controls (file `nsw_psid.csv`).[3]

[3]The combination of the NSW and PSID data is one of the most-studied data sets in labor economics; many studies have done (versions of) what you are about to do.

**(a)** Before beginning the analysis, summarize the main concern when it comes to using observational data for the analysis. Why might the PSID *comparison* group not be a good *control* group?

**(b)** Using the PSID control sample as though it were the control group for a randomized trial, estimate the average treatment effect.

**(c)** Does the PSID sample appear to be a good control group for this purpose? What characteristics of the men help answer this question? Provide data-based evidence and discussion for your conclusion, either way you decide. How does this shed light on your finding in **(a)**?

**(d)** Using the above analysis to guide you, build a regression that attempts to control for any sources of nonrandomization. Does using the partial $F$ test help you further select/remove variables? Does your regression-based treatment effect estimate recover the experimental benchmark treatment effect estimate? Discuss the uncertainty of your regression-based estimate.