

# EEP/IAS 118 - Introductory Applied Econometrics

## Problem Set 5, Spring 2025, Villas-Boas

Due in Gradescope - Midnight, May 4

Submit materials (all typed answers and R outputs) as one combined pdf on [Gradescope](#).

### To receive full credit:

- **Demonstrate all steps** (code, calculations, and output) to obtain the answer.
- Correctly **assign pages** on Gradescope.
- **Do not use canned functions** for confidence intervals or test functions (i.e. `linearHypothesis()`). Compute these statistics "by hand". You can use functions like `sd()` or `qt()` to get values to plug into the formulas.

This problem set uses two data sets: `pset5_2025.dta` & `exercise4_2025.dta`.

## Preamble

When writing R code, it is a good habit to start your notebooks with a preamble, a section where you load all necessary packages, set paths, or declare other options. Use the below code cell to load in packages you will use throughout the problem set ( `haven` , `tidyverse` , `lfe` , `lmtest` , `sandwich` , `stargazer` ).

You will also need to install a new package `mfx` and load it.

Set scientific display off using `options(scipen=999)`

```
In [20]: # Insert your code here
```

## Exercise 1: Treatment Probability

### Data description

Data description: In this exercise, you will use data on housing prices for two years, for a sample of houses, and also information on the announcement of the construction of a garbage and recycling drop-off facility. Characteristics of

houses in the sample are also available in the dataset. For this exercise, you will use the Stata file `pset5_2025.dta` provided on Datahub and bCourses.

Note that several problems require you to produce custom summary statistics and regression tables using `stargazer`. For more information on producing these types of tables, see Coding Bootcamp 5 posted on Datahub.

The first dependent variable of interest is whether a recycling and garbage drop-off facility will be near a certain house in the sample. Let the variable `treatment` be equal to 1 or 0, where 1 means the house ever had a recycling facility installed near it and zero otherwise. You will use a linear probability model to explain the probability of a region having a facility as a function of observables of the area at a time before any facility was installed. You will also estimate a logit specification, interpret marginal effects, and perform hypothesis testing.

*Please perform all the calculations for this exercise using real 1978 prices (`rprice`).*

#### Readme for data variables

Variable name	Definition
year	1978 or 1981
age	Age of the house, in years
nbh	Neighborhood identification number, from 1 to 6
price	Selling price of the house
rooms	Number of rooms in the house
area	Square footage of house
land	Square footage of lot
baths	Number of bathrooms
dist	Distance from house to garbage and recycling drop off facility, in feet
rprice	(Real) Price, in 1978 dollars

## Ungraded Question: Data Setup and Explore

(i) First, load the data. (ii) Then, define a new variable called `treated`, which is equal to one if a house is located 22000 ft or less to an upcoming recycling and garbage facility (to be installed in 1981) and equal to zero otherwise. (iii) Finally, filter only the observations for 1978 (before any facilities were constructed) and compare how many observations were lost after filtering.

```
In [21]: # Insert your code here
```

## Question 1.1: Linear Probability Model Regressions

We want to explore what factors **in 1978** are correlated with the probability of a house being treated ( `treated` = 1 or 0; where 1 means that houses will be treated in 1981 when the facility is constructed).

Estimate **3 linear probability model regressions**, as specified below, and present the estimates in a three-column table called Table 1. Make sure you **use robust standard errors** in all regressions. Make sure to use only data from 1978. In your table, denote with a star \* the coefficients that are significant at the 10% level, two stars \*\* those significant at the 5%, and three stars \*\*\* those significant at the 1% level. The models are as follows:

- Column 1: Specify a constant and `age` as regressors.
- Column 2: Add the `rooms` variable to the model specified in Column 1.
- Column 3: Add the `land` variable to the model specified in Column 2.

*Hint: See Coding Bootcamp Part 5 for help producing these tables with stargazer. For reference, we include some example code using stargazer below. Adapt the sample code to answer the question.*

```
reg6a <- lm(treated ~ age, data78)

robust_6a <- sqrt(diag(vcovHC(reg6a, type = "HC1")))

...

stargazer(reg6a, reg6b, reg6c,
          type = "text",
          se = list(robust_6a, robust_6b, robust_6c),
          title = "Table 3: Probability factors of being
treated in 1978")
```

```
In [22]: # Insert your code here
```

## Question 1.2: Interpreting Linear Probability Model Regressions

(a) Which coefficient measures the estimated correlated change in treatment probability when the age of the house changes by one year, controlling for no

other covariates? Looking at the significance stars, is it significantly different from 0 at the 5 percent level? **Please answer in a maximum of 2 sentences.**

→ Type your answer here

(b) Which coefficient measures the estimated correlated change in treatment probability if the house land square footage increases by one square foot holding rooms and age constant? Is it statistically significant at the 5 percent level?

**Please answer in a maximum of 2 sentences.**

→ Type your answer here

### Question 1.3: Linear Probability Model Predictions

Based on the linear probability model of column (3) in Table 1, create a variable equal to the predicted probabilities. Calculate how many predicted probabilities are less than zero and greater than one. **In no more than 2 sentences**, comment on what problem does this highlight, if any, of using a linear probability model.

In [23]: *# Insert your code here*

→ Type your answer here

### Question 1.4: Logit Model

Estimate the same right-hand side specification as in column 3 of Table 1 above but now use a Logit model. After you estimate the model, type the marginal effects command in R to obtain the estimated marginal effects. Please print both results.

What do you conclude in terms of the marginal effect of age on the probability of receiving the treatment at the 5% significance level? **Please answer in a maximum of 2 sentences.**

*Hint: See Section 13. For reference, we include some sample code below. Adapt the sample code to answer the question.*

```
logit6c <- glm(treated ~ ..., family = binomial(link =  
"logit"))  
logitmfx(treated ~ ..., atmean = TRUE)
```

In [24]: *# Insert your code here*

→ Type your answer here

## Question 1.5: Likelihood Ratio Test with Logit Models

$$Treatment_i = \beta_0 + \beta_1 age_i + \beta_2 rooms_i + \beta_3 land_i$$

Conduct a joint hypothesis test that `rooms` and `land` are jointly significant in predicting treatment status in your **logit model**. What do you conclude at the 5 percent significance level? Use the five steps of hypothesis testing and interpret Step 5 in a maximum of 2 sentences.

*Hint: You will need to estimate an additional logit model as well as use the results from the logit model you estimated in Question 1.4 above. The Likelihood Ratio Test (or Chi-Squared test) is an F-test for the logit model. See section 13 notes. Do not use canned hypothesis testing functions.*

Step 1: Define your hypotheses

 Type your answer here

Step 2: Compute your test statistic by using your unrestricted (previous question model) and restricted model.

In [25]: `# Insert your code here`

 Type your answer here

Step 3: Find your critical value using the appropriate distribution.

In [26]: `# Insert your code here`

 Type your answer here

Step 4: Define your rejection rule

 Type your answer here

Step 5: Decide and interpret

 Type your answer here

## Exercise 2: Differences-in-Differences

This question focuses on the effects of the facility construction treatment on housing price ( `price` ). We now use all available data from years 1978 (before any facilities were constructed) and 1981 (after facilities were constructed). Let

`treated*after` be the interaction of **treated** and an indicator variable called **after** (**after** is equal to one if the year is 1981, 0 otherwise).

Let the two models (2.1) and (2.2) be given by the following regressions:

$$(2.1) \text{ price}_{it} = \beta_0 + \beta_1 \text{treated}_i + \beta_2 \text{after}_t + \beta_3 \text{treated}_i * \text{after}_t + u_{it}$$

(2.2)

$$\text{price}_{it} = \beta_0 + \beta_1 \text{treated}_i + \beta_2 \text{after}_t + \beta_3 \text{treated}_i * \text{after}_t + \beta_4 \text{rooms}_{it} + \beta_5 \text{baths}_{it}$$

## Question 2.1:

Suppose you have the following summary statistics table. Use this table to manually calculate what  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  would be in Equation 2.1. Make sure to show your work and all calculations that you conducted. To answer this question, only use the information in the table provided below, do not run any additional regressions at this point with the dataset.

	Average Price	Control (Treated = 0)	Treated (Treated = 1)
Before ( <b>After</b> = 0)		60,000	70,000
After ( <b>After</b> = 1)		90,000	80,000

In [27]: `# Insert your code here`

→ Type your answer here

## Question 2.2:

What are the conditions needed so that we can interpret the coefficient  $\beta_3$  from equation 2.1 as the causal impact of receiving a recycling and garbage facility near a house (the treatment) on house prices? What would be a simple set of tests you could run to support this? Do not run these tests but explain what data you would use and collect and what tests you would run.

→ Type your answer here

## Question 2.3:

Now estimate both models using the `pset5_2025.dta` dataset. Display the estimates in a table with two columns, and label this table `Table 2`. Place the estimates from model (2.1) in column 1 and the results from model (2.2) with additional covariates in column 2. Assume you meet the conditions from question 2.2 to interpret  $\beta_3$  as the causal impact of the treatment on housing prices. Would you conclude that the treatment had a significant impact on

housing prices? Why or why not? Justify your response in a maximum of 2 sentences.

In [28]: `# Insert your code here`

→ Type your answer here

## Exercise 3: Does a sugar sweetened beverage (SSB) tax decrease obesity among middle school students?

In a school district, a superintendent announced that if a student population had an average Body Mass index (BMI) greater than  $X$  in 2022 (where  $X$  is the 85th percentile for that age in 2022), the vending machines for soda would be subject to a SSB tax in 2023, whereas schools that had average student BMI less than  $X$  in 2022 would not be subjected to the SSB tax in 2023. Suppose that you have data for two time periods, 2022 and 2023, for a random sample of schools  $j$  on the average BMI of its students.

How would you estimate the causal effect of the SSB tax on the outcome  $Y_{j,2023}$ , where  $Y_{j,2023}$  is the average BMI in school  $j$  in year 2023 (after the new SSB tax is implemented)? What impact evaluation method (research design) would be most likely to give you a causal estimate?

1. Write down the exact regression you would run and define each variable.  
(Hint: Pay attention to your subscripts!)
2. Specify which coefficient in your regression would be interpreted as the causal effect of the SSB tax on students BMI.
3. What assumption is key for you to interpret the coefficient as a causal effect of the SSB tax?

→ Type your answer here

## Exercise 4: Fixed Effects Panel Regression

Open the dataset for exercise 4 (**exercise4\_2025.dta**), which has three years of data (1987-1989) for firms on how much scrap they produce and other firm characteristics, such as whether they have a union( `union` ), annual sales ( `sales` ), and number of employees ( `employ` ). Some firms received a grant in 1988 to reduce scrap production, represented by the dummy variable `grant` , which = 1 if firm  $j$  received the grant in year  $t$  and is = 0 otherwise. The table below shows the results from estimating models of scrap by firm  $j$  in year  $t$  on

the variables specified in the rows, for 4 different specifications, with each specification represented by a separate column:

- Column 1: Includes an intercept and controls for **grant**
- Column 2: Includes an intercept, the controls in Column 1, and also controls for **union**, **sales** and **employ**
- Column 3: Includes an intercept, the controls in Column 2, and also includes year fixed effects (an indicator variable for each year)
- Column 4: Includes an intercept, the controls in Column 3, and also includes firm fixed effects (an indicator variable for each firm)

Effect of Grant Treatment on scrap Generated by Firms in Michigan				
	(1) REG31	(2) REG32	(3) REG33	(4) REG34
grant	<b>-0.8684</b> (1.2332)	<b>-0.1110</b> (1.0901)	<b>0.0071</b> (1.1720)	<b>-0.6429</b> (0.4283)
union		<b>3.3242***</b> (0.9826)	<b>3.2880**</b> (0.9891)	<b>0.0000</b> (.)
employ		<b>0.0210*</b> (0.0103)	<b>0.0202</b> (0.0105)	<b>0.0015</b> (0.0127)
sales		<b>-0.0000*</b> (0.0000)	<b>-0.0000*</b> (0.0000)	<b>0.0000</b> (0.0000)
1987.year			<b>0.0000</b> (.)	<b>0.0000</b> (.)
1988.year			<b>-0.2878</b> (1.1157)	<b>-0.2129</b> (0.3622)
1989.year			<b>-0.7409</b> (1.0752)	<b>-0.9107*</b> (0.3559)
_cons	<b>3.9991***</b> (0.5218)	<b>2.4933***</b> (0.6325)	<b>2.8077**</b> (0.8587)	<b>3.7909***</b> (0.6160)
N	162	148	148	148
r2	0.0031	0.0889	0.0921	0.9415
aic	1043.1810	906.5456	910.0254	502.2632
F	0.4959	3.4887	2.3841	2.4235
Standard errors in parentheses				
* p<0.05, ** p<0.01, *** p<0.001				

Question 4.1:



A) We have three years of data, 1987, 1988, and 1989. Why does the year 1987 indicator not get estimated in column (3)?

B) If we have  $N$  firms, how many firm fixed effects are estimated in column (4) when there is a constant?

C) Why does the union variable not get estimated in column (4)?

 Type your answer here


## Question 4.2:

Using the `exercise4_2025.dta` dataset, estimate in R the specification in 3.3 and 3.4 (replicate them both) in a table called `Table 3`. Make sure to show all relevant code and use `stargazer` to produce a well-formatted table.

In [29]: `# Insert your code here`

## Question 4.3:

Your results in Column 4 of Table 3 include two-way fixed effects (year and firm fixed effects). What is the key identifying assumption that is needed for your results to identify the causal relationship of the grant on firms' scrap production? Explain your answer in a maximum of 2 sentences.

 Type your answer here

# THE END