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 [1]:
```
# Install new package
install.packages("mfx", quietly = TRUE)

# Load libraries
suppressMessages(suppressWarnings({
library(haven)
library(tidyverse)
library(lfe)
library(lmtest)
library(sandwich)
library(stargazer)
library(mfx)
```

```
}))

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

```
Installing package into ‘/srv/r’
(as ‘lib’ is unspecified)

also installing the dependencies ‘flexmix’, ‘modeltools’, ‘betareg’
```

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

Variable name	Definition
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 and compare how many observations were lost after filtering.

In [2]:
```
# Read in data
data <- read_dta("pset5_2025.dta")

# Number of observations
obs <- nrow(data)

# Create treatment
data <- data %>%
        mutate(treated = ifelse(dist <= 22000,1,0))

# Filter data by year
data78 <- data %>% filter(year == 1978)
data81 <- data %>% filter(year == 1981)

# Number of observations for 1978
obs78 <- nrow(data78)

cat('Number of observations:', obs, '\n')
cat('Number of observations in 1978:', obs78)
```

```
Number of observations: 319
Number of observations in 1978: 179
```

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 [3]:

```
# Insert your code here
reg6a <- lm(treated ~ age, data78)
reg6b <- lm(treated ~ age + rooms, data78)
reg6c <- lm(treated ~ age + rooms + land, data78)

robust_6a <- sqrt(diag(vcovHC(reg6a, type = "HC1")))
robust_6b <- sqrt(diag(vcovHC(reg6b, type = "HC1")))
robust_6c <- sqrt(diag(vcovHC(reg6c, type = "HC1")))

stargazer(reg6a, reg6b, reg6c,
          type = "text",
          se = list(robust_6a, robust_6b, robust_6c),
          title = "Table 1: Probability factors of being treated in 1978")
```

Table 1: Probability factors of being treated in 1978

	Dependent variable: treated		
	(1)	(2)	(3)
age	0.004***	0.004***	0.004***
	(0.001)	(0.001)	(0.001)
rooms		-0.088**	-0.075**
		(0.036)	(0.038)
land			-0.00000
			(0.00000)
Constant	0.472***	1.048***	1.036***
	(0.043)	(0.234)	(0.225)
Observations	179	179	179
R2	0.084	0.112	0.137
Adjusted R2	0.079	0.102	0.122
Residual Std. Error	0.479 (df = 177)	0.472 (df = 176)	0.467 (df = 175)
F Statistic	16.234*** (df = 1; 177)	11.154*** (df = 2; 176)	9.238*** (df = 3; 175)
Note:			*p<0.1; **p<0.05; ***p<0.01

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

The coefficient on age in model (1) in column 1 from Table 1 measures the estimated correlated change in treatment probability when the age of the house changes by one year, controlling for no other covariates. It is statistically significant at the 1% level, thus also at the 5% level.

(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

The coefficient on land in model (3) in column 3 measures the estimated correlated change in treatment probability if lot size increases by one square foot holding rooms and age constant. It is not statistically significant at the 10%, 5%, 1% level.

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 [4]:

```
# Insert your code here
## Option one: Use fitted.values formula (award full credit)
data78$predicted_prob = reg6c$fitted.values
data78 %>% filter(predicted_prob < 0 | predicted_prob > 1) %>% nrow()

## Option two: Manually calculate fitted values from estimated regression co
data78$predicted_prob_v2 <- 1.036 + 0.004*data78$age - 0.075*data78$rooms
data78 %>% filter(predicted_prob_v2 < 0 | predicted_prob_v2 > 1) %>% nrow()
```

4
5

➡️ Type your answer here

We obtain 4 observations with a predicted probability of less than 0 or greater than 1. This highlights the issue with linear probability models in that this model does not necessarily predict results between 0 and 1 as a probability would realistically behave.

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 [5]:

```
# Insert your code here
# Give full points if student either does atmean = FALSE or atmean = TRUE

# Answer if atmean = TRUE (award full points)
logit1 <- glm(treated ~ age + rooms + land, data78,
               family = binomial(link = "logit"))
summary(logit1)

logitmfx(treated ~ age + rooms + land, data78, atmean = TRUE)

# Answer if atmean = FALSE (award full points)

logit1_v2 <- glm(treated ~ age + rooms + land, data78,
               family = binomial(link = "logit"))
summary(logit1_v2)

logitmfx(treated ~ age + rooms + land, data78, atmean = FALSE)
```

```
Call:
glm(formula = treated ~ age + rooms + land, family = binomial(link = "logi
t"),
    data = data78)

Coefficients:
                Estimate   Std. Error z value Pr(>|z|)
(Intercept)  2.216420719  1.275498180   1.738  0.08227 .
age          0.021684245  0.007047764   3.077  0.00209 **
rooms       -0.293261428  0.194762267  -1.506  0.13213
land        -0.000010627  0.000006387  -1.664  0.09616 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 246.13  on 178  degrees of freedom
Residual deviance: 217.22  on 175  degrees of freedom
AIC: 225.22

Number of Fisher Scoring iterations: 5
Call:
logitmfx(formula = treated ~ age + rooms + land, data = data78,
    atmean = TRUE)

Marginal Effects:
              dF/dx     Std. Err.       z    P>|z|
age    0.0053087820  0.0016825343  3.1552 0.001604 **
rooms -0.0717968727  0.0476385955 -1.5071 0.131781
land  -0.0000026016  0.0000015759 -1.6508 0.098770 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Call:
glm(formula = treated ~ age + rooms + land, family = binomial(link = "logi
t"),
    data = data78)

Coefficients:
                Estimate   Std. Error z value Pr(>|z|)
(Intercept)  2.216420719  1.275498180   1.738  0.08227 .
age          0.021684245  0.007047764   3.077  0.00209 **
rooms       -0.293261428  0.194762267  -1.506  0.13213
land        -0.000010627  0.000006387  -1.664  0.09616 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 246.13  on 178  degrees of freedom
Residual deviance: 217.22  on 175  degrees of freedom
AIC: 225.22

Number of Fisher Scoring iterations: 5
```

```
Call:
logitmfx(formula = treated ~ age + rooms + land, data = data78,
    atmean = FALSE)

Marginal Effects:
              dF/dx     Std. Err.        z   P>|z|
age    0.0045883829  0.0016449060  2.7894 0.00528 **
rooms -0.0620540731  0.0425154756 -1.4596 0.14441
land  -0.0000022486  0.0000013958 -1.6110 0.10719
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

➡️ Type your answer here

Correct interpretation for atmean = TRUE: Holding all other variables constant, a 1 year increase in the age of the house is correlated with an increase in the probability of receiving the treatment by approximately 0.53 percentage points. This result is statistically significant at the 1% significance level.

Correct interpretation for atmean = FALSE: Holding all other variables constant, a 1 year increase in the age of the house is correlated with an increase in the probability of receiving the treatment by approximately 0.46 percentage points. This result is statistically significant at the 1% significance level.

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

This is a joint hypothesis test for a Logit Model.

$$H_0 : \beta_2 = \beta_3 = 0$$

$$H_A : \beta_2 \neq 0 \text{ or } \beta_3 \neq 0 \text{ or } both$$

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

In [6]:
```
# Insert your code here

logit2 <- glm(treated ~ age, data78, family = binomial(link = "logit"))
summary(logit2)

LL_R <- logLik(logit2)

LL_UR <- logLik(logit1)

LL_R
LL_UR

LikeRatio <- 2 * (LL_UR - LL_R)

print(paste("The Log-Likelihood function value for the unrestricted model is
            LL_UR))
print(paste("The Log-Likelihood function value for the restricted model is",
            LL_R))
print(paste("The Likelihood Ratio test statistic is",
            LikeRatio))
```

```
Call:
glm(formula = treated ~ age, family = binomial(link = "logit"),
    data = data78)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.245653   0.188181  -1.305 0.191754
age          0.028632   0.008629   3.318 0.000906 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 246.13  on 178  degrees of freedom
Residual deviance: 225.74  on 177  degrees of freedom
AIC: 229.74

Number of Fisher Scoring iterations: 5
'log Lik.' -112.8678 (df=2)
'log Lik.' -108.6122 (df=4)
[1] "The Log-Likelihood function value for the unrestricted model is -108.61
223454433"
[1] "The Log-Likelihood function value for the restricted model is -112.8678
41064789"
[1] "The Likelihood Ratio test statistic is 8.51121304091802"
```

For the Likelihood Ratio Test, the test statistic is

$$LR = 2(LL_{UR} - LL_R) = 8.5112$$

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

In [7]:
```
# Insert your code here

q.05 <- qchisq(0.95, 2)
print(paste("The critical value for alpha = 0.05 is ",
            q.05))
```

```
[1] "The critical value for alpha = 0.05 is  5.99146454710798"
```

Our test statistic follows a chi-squared distribution with q=2 degrees of freedom. In this case, our critical value for alpha = .05 is 5.9915

Step 4: Define your rejection rule

➡️ Type your answer here

We reject the null hypothesis if the test statistic is greater than the critical value.

Step 5: Decide and interpret

➡️ Type your answer here

Since our test statistic 8.5112 > 5.9915, our critical value, we reject the null hypothesis at the 5% significance level. We reject the null hypothesis that `rooms` and `land` are jointly not significant in predicting Treatment status in favor of the alternative hypothesis that `rooms` and `land` are jointly significant in predicting Treatment Status at the 5% significance level.

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) $price_{it} = \beta_0 + \beta_1 treated_i + \beta_2 after_t + \beta_3 treated_i * after_t + u_{it}$

(2.2)
$price_{it} = \beta_0 + \beta_1 treated_i + \beta_2 after_t + \beta_3 treated_i * after_t + \beta_4 rooms_{it} + \beta_5 l$

Question 2.1:

Suppose you have the following summary statistics table. Use this table to manually calculate what β_0, β_1, β_2, and β_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 (**After** = 0)	$60,000$	$70,000$
After (**After** = 1)	$90,000$	$80,000$

➡️ Type your answer here

- $\beta_0 = E[price|Before, Control] = 60,000$
- $\beta_1 = E[price|Before, Treated] - E[price|Before, Control] = 70,000 - 6($
- $\beta_2 = E[price|After, Control] - E[price|Before, Control] = 90,000 - 60,$
- $\beta_3 = \triangle\overline{Price}_T - \triangle\overline{Price}_C = (80,000 - 90,000) - (70,000 - 60,000) = -$

Question 2.2:

What are the conditions needed so that we can interpret the coefficient β_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

β_3 can be interpreted as a difference-in-differences causal impact under the assumption that the difference between 1978 and 1981 for the untreated houses is a good counterfactual for the treated houses, i.e. treated houses and untreated houses have **parallel trends** in home prices prior to 1981.

We can test the validity of the identification assumption by looking at the slopes in home prices in the pre-period (before 1981) leading up to the treatment, to show that parallel trends hold. In order to do this, we would need to collect more years of pre-1981 data and could run a difference-in-differences specification of equation (2.1), only on years prior to the intervention (pre-1981).

Specifically, if we had multiple years of data before the treatment we could estimate the following equation using only the data before the treatment (prior to 1981)

$$price_{it} = \beta_0 + \beta_1 treated_i + \beta_2 pre_t + \beta_3 treated_i * pre_t + u_{it}$$

Then I would conduct a two-sided hypothesis test of β_3:

$$H_0 : \beta_3 = 0$$

$$H_A : \beta_3 \neq 0$$

If I failed to reject the null, then this would be evidence in support of the parallel trends assumption.

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 β_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 [8]:

```
# Insert your code here

data$after <- ifelse(data$year == 1981, 1, 0)
data$treatedxafter<-data$treated*data$after

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

reg2_1 <- lm(price ~ treated+after+treatedxafter, data = data)
#summary(reg2_1)

reg2_2 <- lm(price ~ treated+after+treatedxafter+rooms+land+age,
             data = data)
#summary(reg2_2)

stargazer(reg2_1, reg2_2, type = "text",
title = "Table 2: Effects of recycling and garbage facility treatment on hom
```

Table 2: Effects of recycling and garbage facility treatment on home prices

	Dependent variable: price	
	(1)	(2)
treated	-7,034.448	9,924.282**
	(5,561.163)	(4,685.297)
after	50,475.540***	48,252.500***
	(6,445.136)	(5,158.593)
treatedxafter	-10,189.210	-10,161.480
	(8,464.739)	(6,764.484)
rooms		19,553.210***
		(1,889.959)
land		0.185***
		(0.044)
age		-343.391***
		(53.417)
Constant	80,518.600***	-58,209.910***
	(4,135.773)	(13,065.700)
Observations	319	319
R2	0.276	0.544
Adjusted R2	0.269	0.535
Residual Std. Error	36,991.480 (df = 315)	29,503.670 (df = 312)
F Statistic	40.064*** (df = 3; 315)	62.020*** (df = 6; 312)
Note:		*p<0.1; **p<0.05; ***p<0.01

➡️ Type your answer here

In both models, the estimated coefficient on $treatedxafter$ is not statistically significantly different from zero. I would therefore conclude that the facility treatment had no significant impact on housing prices.

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 etimate 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

We can estimate this causal effect using a regression discontinuity design:

$$BMI_{j,2023} = \beta_0 + \beta_1 T_j + \beta_2(BMI_{j,2022} - X) + \beta_3 T_j * (BMI_{j,2022} - X) + u_{j,t}$$

Here, both the outcome and the running variable are BMI - but at different time periods. The outcome is average BMI in 2023, and the running variable is average BMI in the year prior (2022). The treatment variable is SSB tax regulation in the current year and the cutoff is given by X. So, $T_i = 1$ if $BMI_{j,2022} > X$, and is =0 otherwise. The treatment effect of the SSB tax is given by β_1.

For a regression discontinuity model, the key assumption to interpret the β_1 coefficient as causal is that the relationship between average BMI levels in 2023 and average BMI levels in 2022 would be continuous at the threshold X if it were not for the SSB tax intervention. This assumption might be violated if, for example, schools could manipulate the reported BMI of their students in order to avoid the tax.

Note: Diff-in-Diff and other research methods can be used, if appropriate assumptions hold. However, RDD is the specification we are looking for, because the set-up clearly indicates the treatment status (whether the school gets the SSB tax) was determined based on a threshold in the running variable. So, we know RDD assumptions are satisfied for us to leverage.

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	-0.8684	-0.1110	0.0071	-0.6429
	(1.2332)	(1.0901)	(1.1720)	(0.4283)
union		3.3242***	3.2880**	0.0000
		(0.9826)	(0.9891)	(.)
employ		0.0210*	0.0202	0.0015
		(0.0103)	(0.0105)	(0.0127)
sales		-0.0000*	-0.0000*	0.0000
		(0.0000)	(0.0000)	(0.0000)
1987.year			0.0000	0.0000
			(.)	(.)
1988.year			-0.2878	-0.2129
			(1.1157)	(0.3622)
1989.year			-0.7409	-0.9107*
			(1.0752)	(0.3559)
_cons	3.9991***	2.4933***	2.8077**	3.7909***
	(0.5218)	(0.6325)	(0.8587)	(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

A) We cannot estimate coefficients on all 3 year indicators without running into perfect collinearity issues.

B) There are N-1 individual fixed effects estimated in column (4) when there is a constant.

C) The union variable does not get estimated in column (4), because it is fully absorbed by the firm fixed effects in that specification. Remember that firm-level fixed effects will capture any effect from factors which are constant over time for firms, such as union presence.

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-formated table.

In [9]:
```
# read in the new dataset
dataEx4 <- read_dta("exercise4_2025.dta")
#head(dataquestion5)
```

In [10]:
```
# Method using factor(), which more closely replicates the existing table, w

reg5_3 <- lm(scrap ~ grant + union +  employ +
             sales + factor(year), dataEx4)

reg5_4 <- lm(scrap ~ grant + employ +
             sales + factor(year) + factor(fcode), dataEx4)

stargazer(reg5_3, reg5_4, type = "text", title = "Scrap and grants (Exercise

# Method using felm, which automatically drops the constant

reg5_3_alt <- felm(scrap ~ grant + union +  employ +
             sales | year, dataEx4)

reg5_4_alt <- felm(scrap ~ grant + union +  employ +
             sales| year + fcode, dataEx4)

stargazer(reg5_3_alt, reg5_4_alt, type = "text", title = "Table 3: Scrap and
```

Scrap and grants (Exercise 4.2)

	Dependent variable:	
	scrap	
	(1)	(2)
grant	0.007	-0.643
	(1.172)	(0.428)
union	3.288***	
	(0.989)	
employ	0.020*	0.002
	(0.010)	(0.013)
sales	-0.00000**	0.000
	(0.00000)	(0.00000)
factor(year)1988	-0.288	-0.213
	(1.116)	(0.362)
factor(year)1989	-0.741	-0.911**
	(1.075)	(0.356)
factor(fcode)410538		3.063
		(1.885)
factor(fcode)410563		5.962***
		(1.334)
factor(fcode)410565		5.773***
		(1.529)
factor(fcode)410566		5.559***
		(1.429)
factor(fcode)410567		0.603
		(1.346)
factor(fcode)410577		1.582
		(1.356)
factor(fcode)410593		1.072
		(1.492)
factor(fcode)410596		6.725***
		(1.421)
factor(fcode)410606		0.346
		(1.413)
factor(fcode)410626		0.836
		(1.396)
factor(fcode)410629		0.791

	(1.407)
factor(fcode)410653	1.137
	(1.431)
factor(fcode)410665	0.303
	(1.579)
factor(fcode)410685	0.225
	(1.401)
factor(fcode)418011	5.218***
	(1.355)
factor(fcode)418021	1.276
	(1.456)
factor(fcode)418035	4.915***
	(1.391)
factor(fcode)418045	2.057
	(1.429)
factor(fcode)418054	1.452
	(1.351)
factor(fcode)418065	1.082
	(2.395)
factor(fcode)418076	0.885
	(1.347)
factor(fcode)418083	2.126
	(1.533)
factor(fcode)418091	3.157*
	(1.642)
factor(fcode)418097	1.534
	(1.465)
factor(fcode)418107	0.907
	(1.412)
factor(fcode)418118	0.804
	(1.437)
factor(fcode)418125	0.653
	(1.349)
factor(fcode)418163	7.685***
	(1.328)
factor(fcode)418168	0.174
	(1.520)

factor(fcode)418177	18.440***
	(1.393)
factor(fcode)418237	0.473
	(1.415)
factor(fcode)419198	5.609***
	(1.445)
factor(fcode)419201	2.054
	(1.431)
factor(fcode)419242	4.709***
	(1.345)
factor(fcode)419268	1.436
	(1.474)
factor(fcode)419272	20.733***
	(1.375)
factor(fcode)419289	1.560
	(1.444)
factor(fcode)419297	0.073
	(1.452)
factor(fcode)419307	0.776
	(2.424)
factor(fcode)419339	0.519
	(1.365)
factor(fcode)419343	4.021***
	(1.475)
factor(fcode)419357	1.357
	(1.331)
factor(fcode)419378	2.110
	(1.410)
factor(fcode)419381	0.803
	(1.453)
factor(fcode)419388	8.572***
	(1.378)
factor(fcode)419409	2.428*
	(1.325)
factor(fcode)419432	4.614*
	(2.463)
factor(fcode)419459	1.600
	(1.428)

```
factor(fcode)419482                                  2.938*
                                                     (1.567)

factor(fcode)419483                                 24.887***
                                                     (1.400)

Constant                  2.808***                   0.293
                           (0.859)                  (1.148)

-----------------------------------------------------------------
Observations                148                        148
R2                         0.092                      0.941
Adjusted R2                0.053                      0.906
Residual Std. Error   5.116 (df = 141)        1.608 (df = 92)
F Statistic         2.384** (df = 6; 141) 26.908*** (df = 55; 92)
=================================================================
Note:                                 *p<0.1; **p<0.05; ***p<0.01
```

```
Warning message in chol.default(mat, pivot = TRUE, tol = tol):
“the matrix is either rank-deficient or not positive definite”
```

```
Table 3: Scrap and grants (Exercise 4.2)
====================================================
                          Dependent variable:
                    --------------------------------
                                 scrap
                          (1)              (2)
----------------------------------------------------
grant                    0.007           -0.643
                        (1.172)          (0.428)

union                   3.288***
                        (0.989)

employ                   0.020*           0.002
                        (0.010)          (0.013)

sales                  -0.00000**         0.000
                       (0.00000)        (0.00000)

----------------------------------------------------
Observations              148              148
R2                       0.092            0.941
Adjusted R2              0.053            0.906
Residual Std. Error 5.116 (df = 141) 1.608 (df = 92)
====================================================
Note:                    *p<0.1; **p<0.05; ***p<0.01
```

Question 4.3:

Your results in Column 4 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 sentneces.

➡️ Type your answer here

In order for a two-way fixed effects model to be causally identified, we need to assume that there are no omitted variables that vary by firm across time that are correlated with the likelihood that a firm receives a grant and a firm's scrap production.

THE END

In [ ]:

```

```