Guidelines for Final exam HMET4220 – Applied Micro Econometrics, Spring 2024

Instructions

This open book home-exam is to be solved independently. You can discuss the exam with your fellow students, but you should submit an individual assignment in your own words. Copying text from others is not allowed. You can use the statistical software (e.g. Stata, R Python, etc.) and text editor of your choice (e.g. Rmarkdown, Word, \LaTeX, Google docs, etc.). Your code should be included as a readable appendix, or, if you use Rmarkdown, the code can be included in the text using codechunks (set warnings and messages to FALSE such that only the code and output are visible in the document). No matter which programs you use, you should convert the document into a single pdf that you upload to Inspera.

You don't need to add a reference list if you only use references from the curriculum. Just use inline references with the author names and year of publication, and you can also refer directly to lecture slides. If you use references that are not included in the curriculum, you need to add a reference list that includes the author names, year of publication, and name of journal of the references that are not in the curriculum. Use only references from peer-reviewed journals. In general, do not use direct quotations. If you find it essential to use a direct quotation, remember to use quotation marks, and refer directly to page and paragraph number. This also applies to content from course materials such as the lecture slides and seminar notes.

The first exercise count 35 percent and the second exercise counts 65 percent. Plan your time thereafter.

The exam is anonymous. Do not add your name anywhere in the document. Use your candidate number.

Exercise 1: Provider quality (35%)

In this exercise, students can earn up to 35 points in total. The points allocated to each exercise are highlighted in bold.

Countries with publicly funded healthcare debate whether for-profit or not-for-profit providers should deliver health services. Consider the context where an influential commentator from a leading newspaper has argued that for-profit hospitals, motivated by profit incentives and higher co-payments, deliver superior care at lower costs than public hospitals. The commentator support his claim by referencing a government report indicating that the unadjusted 30-day mortality rate for Acute Myocardial Infarction (AMI) patients is 5% at for-profit hospitals, compared to 8% at public hospitals.

Assume you are provided with data from the government report, which contains information on a sample of AMI patients treated at for-profit and public hospitals in 2018. The data set includes whether the patient died within 30-days, general demographic and socio-economic information (age, gender, county and municipality of residence, income, education), information on health status (the most common comorbidites of AMI) and post-admission complications (re-admitted to hospital because of infections, irregular heartbeats or other conditions).

1.0. Formulate and interpret a linear regression model to explore the relationship between 30-day mortality and the type of hospital (for-profit vs. public). Write the regression equation clearly, and discuss the likely size and signs of the coefficients in the model.

Guide to answer:

4 points

The students should present a linear regression like this:

$$y_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

where y_i is an indicator equal to 1 if patient i died within 30-days of hospital treatment, and D_i is an indicator equal to 1 if patient i is treated at a for-profit hosptial and zero if the patient is treated at a public hospital.

- The students should interpret the coefficients.
- If no control variables are included, then you can also quantify the coefficients using the information in the exercise: $\beta_0 = 0.08$ and $\beta_1 = -0.03$
- The students could also include a vector of control variables into this regression from the list of potential controls listed in the exercise.
- If controls variables are included the β_1 is interpreted as the association between for-profit status and 30-day mortality when the correlation with for-profit status and the included control variables are taken out.

1.1 Define the specific causal effect you are investigating.

Guide to answer:

4 points

The individual causal effect of D (for-profit hospital treatment) on y (30-day mortality risk) in words is: the contrast in 30-day mortality risk for someone treated at a for-profit hospital to the 30-day mortality risk if that someone had been treated at a public hospital instead.

Preferable the students should also define the causal effect using the potential outcomes framework:

- Y_i^1 is the probability of death within 30-days if AMI patient i is treated at a for-profit hospital
- Y_i^0 is the probability of death within 30-days if AMI patient i is treated at a public hospital
- D_i is a treatment indicator equal to 1 if individual i is treated in a for-profit hospital
- D_i is a treatment indicator equal to 0 if individual i is treated in a public hospital

The students can also discuss at least one of the following treatment effects:

- The individual treatment effect is $\delta_i = Y_i^1 - Y_i^0$
- The average treatment effect (ATE) is $E[\delta_i]$
- The average treatment effect on the treated (ATT): $E[Y_i^1|D_i = 1] - E[Y_i^1|D_i = 0]$

1.2 Evaluate whether your linear regression model from Question **1.0.** adequately identifies the causal effect of interest. Discuss the limitations or assumptions of this model.

Guide to answer:

4 points

- The simple bivariate regression says nothing about causality, only that OLS is unbiased for mean differences in mortality by exercise types.
- Including control variables the regression will identifiy the causal effect if the model is correctly specified and that potential outcomes are linear in parameters.

1.3 Construct a causal diagram for this research question:

a) Identify and justify which observable variables should be included in the causal diagram. Consider simplifying the variable space to avoid clutter.

b) Propose any relevant latent or unobserved variables that should be included in the diagram.

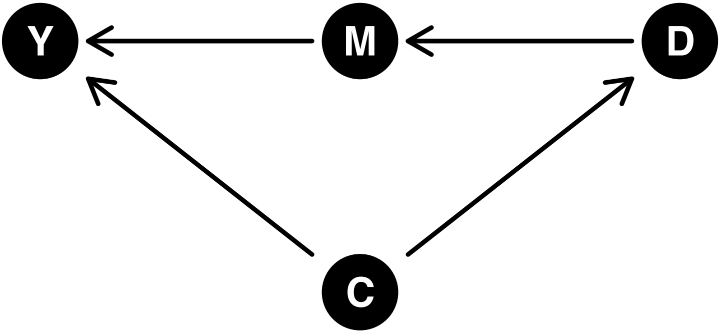
c) Draw the causal diagram incorporating the variables from parts a) and b).

Guide to answer:

6 points

- A critical aspect of this exercise is for students to identify potential **confounders**—variables that influence both the treatment and the outcome. For example, consider the scenario where for-profit hospitals with co-payments prefer treating healthier individuals. This preference may lead to a patient demographic with higher socio-economic status, which correlates with better health outcomes. Therefore, socio-economic status should be considered a confounder (denoted as C in the diagram) and can be visually represented in a causal diagram as shown.

- Additionally, it is important to correctly categorize post-admission complications as **mediators**, not confounders, as illustrated in the figure (where they are denoted as M).



1.4 Based on your diagram suggest a more sophisticated regression model than the one used in Question **1.0.** to better identify the causal effect of interest. Discuss whether this new model can accurately identify the causal effect.

Guide to answer:

4 points

- Here the students should include the variables that they identified as confounders in the previous exercise into the regression model.
- Post-admission complications should not be included as controls.
- They should problemaitize that this regression model is able to identify the causal effect of interest, such that the identifying assumption holds: All confounders are included in a correctly specified model.

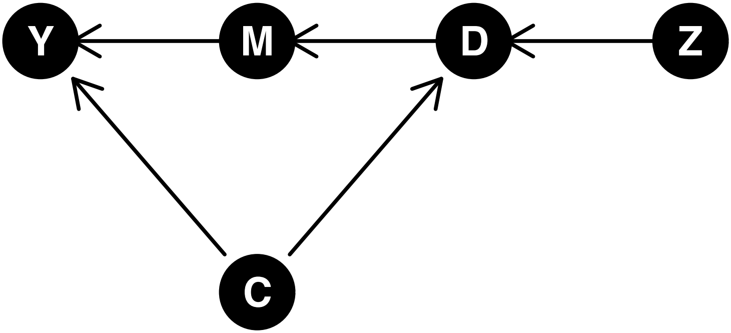
1.5 Consider a research article that uses proximity to for-profit hospitals as an instrumental variable to measure the impact of for-profit treatment on 30-day mortality. They employ a dummy variable indicating the presence of a for-profit hospital in the municipality:

Include this instrument in your causal diagram from Question **1.4** Assess whether the proximity to for-profit hospitals is a valid instrument for determining the effect of treatment type on patient outcomes.

Guide to answer:

6 points

- A valid instrument (Z) should be as-good-as randomly assigned, and therefore independently distributed of confounders ("independence assumptions"), and should only influence the outcome through the treatment variable ("excludability assumption").
- If the instrument Z satisifies these assumptions it can be drawn into the figure like the picture shows.
- Howver, whether you live close to a for-profit hospital is not randomly assigned, and therefore highly likely to be correlated with confounding variables. For example, that for-profit hospital are placed where people can afford the higher out-of-pocket payments (or the insurance to cover them). Therefore you would expect there to be an arrow from C to Z in the figure. Therefore, proximity to for-profit hospital is not a valid instrument.



1.6 In response to the high demand at a specialized for-profit hospital for treating less severe (non-STEMI) AMI cases, a county governor initiates a lottery system to allocate patients between for-profit and public hospitals.

After the lottery was initiated, every diagnosed AMI patient is assigned a random lottery number. The county has a set capacity at the for-profit hospital; patients with lottery numbers below this capacity threshold can choose to receive treatment at the for-profit hospital.

a) Propose and justify a research design using this lottery to explore the causal relationship between the type of hospital treatment and 30-day mortality.

b) Describe the specific effect identified by your design. Discuss whether it is representative of the general population.

c) Identify potential pitfalls of your proposed research design.

d) Discuss the potential ethical concerns related to using the "hospital lottery" as part of a research design. How might these concerns influence the interpretation and application of the research findings?

Guide to answer:

7 points

There are several potential methods to analyze the impact of hospital type on patient outcomes using the lottery system as a basis for assignment. These methods include analyzing mean differences, employing instrumental variables (IV), conducting regression discontinuity (RD), and performing a difference-in-differences (DD) analysis. Students can choose to explain one of these methods, discussing how it can be applied to the hospital lottery context and why it effectively addresses selection bias.

Methods

1. Intention-to-Treat Analysis

The simplest approach is to perform an intention-to-treat analysis, comparing the mean outcomes between the treatment group (those with lottery numbers below the capacity threshold) and the control group (those with lottery numbers above the threshold). This method is considered intention-to-treat because not all individuals assigned to the treatment group will opt for treatment at the for-profit hospital, mirroring real-world scenarios where not all patients follow prescribed treatments.

2. Instrumental Variable (IV) Approach

An IV approach focuses on identifying the local average treatment effect (LATE) for those patients who opt for for-profit hospital treatment when their lottery number allows them to choose this option. This method is particularly useful for isolating the causal impact of hospital type from other confounding factors.

3. Difference-in-Differences (DD) Analysis

If students have access to panel data, they can conduct a DD analysis. This approach would compare changes in outcomes over time between groups that are exposed to different treatments, thereby controlling for time-invariant unobserved heterogeneity that might affect the results.

4. Regression Discontinuity (RD)

Although a RD approach could be considered, it is less relevant in this scenario since the lottery effectively randomizes patients, ensuring that characteristics are balanced across all lottery numbers. Therefore, RD does not add significant value over simpler randomization methods in this specific case.

By choosing and elaborating on one of these methods, students can delve into how each approach helps overcome selection bias, ensuring the robustness of the causal inferences drawn from the analysis.

b) Specific Effect Identified and Representativeness

Specific Effect Identified:

The primary effect identified by this design is the causal impact of receiving treatment at a for-profit hospital on the 30-day mortality rate of non-STEMI AMI patients compared to receiving treatment at a public hospital. This effect is directly attributable to the type of hospital due to the random assignment.

Representativeness:

The effect identified may not necessarily be representative of the general population. This limitation arises because the effect measured is the local average treatment effect (LATE), which specifically applies to the subset of patients whose hospital assignment is influenced by the lottery (i.e., those who are indifferent between hospitals). Therefore, results may not generalize to patients with strong preferences or needs for one type of hospital over another.

c) Potential Pitfalls of the Research Design

Non-compliance

Patients might refuse to adhere to their assigned treatment due to preferences for a specific hospital, which could lead to issues with treatment fidelity.

Spillover Effects

Interaction or information sharing among patients from different hospitals might dilute or confound the effect of the treatment.

Generalizability

As noted, the results are representative only of those affected by the lottery system, which might limit the applicability of findings to all non-STEMI AMI patients.

Limited Control Over Hospital Practices

Differences in hospital practices that are not related to being a for-profit or public entity could influence outcomes. That is, there could be other characetristics than the for-profit status that create the treatment effect, which is not possible to control for in the analysis.

d) Ethical Concerns and Influence on Research Interpretation

Ethical Concerns

Equity and Fairness

Using a lottery system to allocate health resources could be seen as inequitable or unfair, especially if perceived differences in the quality of care between hospital types exist. However, given that there are capacity constraints, a lottery might be the most equitable allocation mechanism since all have the same probability of being selected.

Patient Autonomy

The lottery system might override patient choice, where patients might have preferences for hospitals based on location, past experiences, or perceived quality.

Risk of Harm

If one type of hospital provides significantly inferior care, randomly assigning patients could expose them to increased risk.

Influence on Research Interpretation and Application

These ethical concerns can affect how the research findings are interpreted and applied. For instance, if stakeholders view the allocation method as unethical, they may question the legitimacy or moral standing of the research, potentially leading to resistance against implementing policy changes based on the findings. Transparency about the benefits and risks, along with rigorous ethical oversight, is crucial to maintain trust and validity in the research process.

Exercise 2: Cross-country differences in infant mortality (65%)

In this exercise, students can earn up to 65 points in total. The points allocated to each exercise are highlighted in bold.

In this exercise, you are asked to discuss and run some of the analysis in Chen, Oster and Williams (2016): "Why Is Infant Mortality Higher in the United States than in Europe?" published in the American Economic Journal: Economic policy 2016, 8(2): 89–124.

The dataset titled `birth_data.csv`, available on Inspera, includes simulated samples: 1% of the U.S. data, and 10% each for Austria and Finland. This dataset allows you to perform similar types of analyses to those conducted by Chen et al. (2016). However, since this is simulated data, replicating their exact analysis is not possible. On the last page, there is a table that provides descriptions and names of the variables.

2.0. Write a short report of the attached paper "Why Is Infant Mortality Higher in the United States than in Europe?" by Chen, Oster and Williams (2016). The report should include a summary of the paper, and a critical discussion of the empirical approach. The summary should identify the research questions that the paper tries to answer, how the paper answers the questions, and the results (about 1 page). The discussion of the empirical approach should give a description and critical assessment of the applied methods. Focus on the following questions: Which regression models are used, and what are the coefficient(s) of interest(s)? Are the variables and models employed relevant for the research question? Are there data limitations, and do you have any suggestions for alternative analyses and sensitivity checks? The report can be less, but should be no longer than 800 words ~ 3 pages.

Guide to answer:

10 points

- This should be cohesive text written in the students' own words and arguments i.e. the text should not be a copy of the text and arguments from the article.
- In the first page, it is important that the student is able to effectively identify the papers' core argument: What is the research question and statement?
- It is important that the students are able to critically discuss the empirical approach. They should identify the key variables and coefficients of interest, data limitations, suggestions for alternative analyses.
- The paper is descriptive, and the students should interpret the models as descriptive and not interpret them causally.

2.1. Load and describe the data. Which countries are included in the data set? What is the time period? How many observations are there in total for each country?

Guide to answer:

3 points

```
library(data.table)
d <- fread("birth_data.csv")
print("number of obs.")
```

```
#> [1] "number of obs."
```

```
nrow(d)
```

```
#> [1] 322804
```

```
print("number of Countries:")
```

```
#> [1] "number of Countries:"
```

```
length(unique(d$country))
```

```
#> [1] 3
```

```
print("years of observations and how many observations per year")
```

```
#> [1] "years of observations and how many observations per year"
```

```
table(d$year)
```

```
#> 
#>  2000  2001  2002  2003  2004  2005 
#> 53950 53495 53233 53930 53805 54391
```

```
print("number of obs. per country:")
```

```
#> [1] "number of obs. per country:"
```

```
d[, .(n = .N), by = country]
```

```
#>    country      n
#>     <char>  <int>
#> 1:      AT  47196
#> 2:      FI  34226
#> 3:      US 241382
```

In a good answer, the candidate should provide accurate and complete information about the number of countries, time period covered, frequency of observations, starting and ending years, and total observations per country.

2.2. Limit the analysis to the group referred to as the 'Demographic sample' in Chen et al. (2016). Create a table of summary statistics for this sample across different countries, similar to what is presented in Panel B, "Demographic Sample" of Table 1 in Chen et al. (2016). Analyze and discuss the differences observed among these countries based on the summary statistics.

Guide to answer:

10 points

- Here the students should make a nice Table similar in Panel B, "Demographic Sample" of Table 1 in Chen et al. (2016).
- Part of the exercise is to choose the correct sample. The 'Demographic sample' is defined in the paper and is "limited to singleton births at ≥ 22 weeks of gestation and ≥ 500 grams with birth weight and gestational age observed". In addition observations with missing demographic covariates should be dropped. The students should define the sample in their answer, and only present summary statistics for this sample.

Here I just show the numbers, but the students should organize this information into a self-explanatory table for clearer understanding:

```r
# Restrict to demographic sample

d.demo <- d %>%
  mutate(across(where(is.character), ~na_if(., ""))) %>%
  filter(gestation >= 22 & birwt >= 500 & singleton == 1) %>%
  filter(!(is.na(male) | is.na(mage) | is.na(married))) %>%
  filter(!(country == "US" & (is.na(black_us) | is.na(education_cat_us)))) %>%
  filter(!(country == "AT" & (is.na(immigrant_at) | is.na(education_cat_at)))) %>%
  filter(!(country == "FI" & is.na(occupation_cat_fi)))


## Make variables for summary statistics

d.demo <- d.demo %>%
  mutate(black_or_immigrant = case_when(black_us == 1 | immigrant_at == 1 ~ 1,
                                        black_us == 0 | immigrant_at == 0 ~ 0, T ~ NA),
         at_least_college = case_when(country == "US" & education_cat_us == "college degree or
                                      country == "AT" & education_cat_at == "university or teac
                                      country == "FI" ~ NA, T ~ 0),
         upper_white_collar = case_when(country == "FI" & occupation_cat_fi == "upper white col
                                        country != "FI" ~ NA, T ~ 0))

# Calculate descriptive statistics by state

d.demo %>%
  group_by(country) %>%
  summarise(
    die        = mean(die*1000),
    gestation = mean(gestation),
    birwt      = mean(birwt),
    male       = mean(male),
    mage       = mean(mage),
    married    = mean(married),
    black_or_immigrant = mean(black_or_immigrant),
    at_least_college = mean(at_least_college),
    upper_white_collar = mean(upper_white_collar),
    number_of_observations = n()
    ) %>%
  pivot_longer(cols = -country, names_to = "variable", values_to = "value") %>%
  pivot_wider(names_from = country, values_from = value) %>%
  select(variable, US, AT, FI)
```

```
#> # A tibble: 10 × 4
#>    variable                       US        AT        FI
#>    <chr>                       <dbl>     <dbl>     <dbl>
#>  1 die                          4.55      2.94      2.65
#>  2 gestation                   39.0      38.0      36.2
#>  3 birwt                     3335.     3346.     3552.
#>  4 male                         0.512     0.512     0.513
#>  5 mage                        27.0      28.3      29.0
#>  6 married                      0.653     0.653     0.599
#>  7 black_or_immigrant           0.149     0.239    NA
#>  8 at_least_college             0.257     0.119    NA
#>  9 upper_white_collar         NA        NA          0.218
#> 10 number_of_observations 231132     45192     32773
```

2.3. Write down the regressions equation(s) used in Panel A "United States versus Finland" and Panel B "United States versus Austria" of Table 3 in Chen et al. (2016). Explain and provide interpretation for the equation(s).

Guide to answer:

6 points

The regression equation should look something like this:

$$y_{ij} = \beta_0^{C_i} + \beta_1^{C_i} US_{j\{i\}} + \sum_{k=1}^{K} \gamma_k^{C_i} BWC_i + \epsilon_{ij}^{C_i}$$

where $C_i \in \{AT_i, FI_i\}$. This implies separate models are run for each country comparison, with each country having its own intercepts and coefficients.

- y_{ij} denotes four different outcomes for birth i in country j: one-year mortality (Column 1 & 2 in Table 6 of the replication paper); deaths up to one week (Column 3); deaths from one week to one month, conditional on surviving to one week (Column 4); and deaths from one to 12 months, conditional on surviving to one month (Column 5). All outcomes are measured as deaths per 1,000 births.
- $\beta_0^{C_i}$ represents the country-specific intercept for country C_i (either Austria or Finland).
- $\beta_1^{C_i}$ is the country-specific coefficient for the US dummy variable $US_{j\{i\}}$, which takes a value of 1 for births occurring in the US, and 0 otherwise.
- $\sum_{k=1}^{K} \gamma_k^{C_i} BWC_i$ captures the indicator variables of K birth weight categories specific to each country C_i, adjusting the neonatal mortality based on these categories. In Column these controls are not included. **In Column 1, these controls are not included**
- $\epsilon_{ij}^{C_i}$ is the country-specific error term for each observation.

2.4. Use the regression models from **2.3.** to conduct the analysis from Panel A 'United States versus Finland' and Panel B 'United States versus Austria' in Table 3 of Chen et al. (2016). Document your process, present your results, and discuss the implications of the results.

Note that because the data is simulated, your results will differ from those in the paper.

Guide to answer:

10 points

The students should document their process and results (preferably, in a self-explanatory table), and discuss the implications of the results. Here I just show the results:

- Students should preferably use robust standard errors, as is used in the replication article
- I use the same sample as I used in 2.2, but this was not stated explicitly so the student can use the sample with missing demographic characteristics.
- The students should control flexibly for age, but will likely not use the exact same specification as in this guide, and therefore the results might be a little different.

```r
## Use only the Demographic sample with only non-missing observations

# use d which is shorter d.demo

d <- copy(d.demo)

setDT(d)

## Making dependent variables:
d[, lt_week_cond := case_when(death_day > 7 ~ 0, T ~ die)]
d[, week_to_month_cond := case_when(death_day > 30 ~ 0,
                                    death_day <= 7 ~ NA, T ~ die)]
d[, post_neonatal_cond := case_when(death_day <= 30 ~ NA, T ~ die)]
## Making independent variables
d[, birth_weight_cat := cut(birwt, breaks = c(0, seq(500, 6000, by = 500)), labels = FALSE, inc
d[, us := case_when(country == "US" ~ 1, T ~ 0)]
```

```r
m1 <- feols(die*1000 ~ us, data = d[country %in% c("US", "FI")], vcov = "hetero")
m2 <- feols(die*1000 ~ us | birth_weight_cat, data = d[country %in% c("US", "FI")], vcov = "het
m3 <- feols(lt_week_cond*1000 ~ us | birth_weight_cat, data = d[country %in% c("US", "FI")], vc
m4 <- feols(week_to_month_cond*1000 ~ us | birth_weight_cat, data = d[country %in% c("US", "FI"
m5 <- feols(post_neonatal_cond*1000 ~ us | birth_weight_cat, data = d[country %in% c("US", "FI"
etable(m1, m2, m3, m4, m5, signif.code = c("***"=0.01, "**"=0.05, "*"=0.10))
```

```
#>                                 m1                m2                m3
#> Dependent Var.:           die*1000          die*1000 lt_week_cond*1000
#>
#> Constant         2.655*** (0.2842)
#> us               1.897*** (0.3168) 0.8011** (0.3261) 0.3161** (0.1373)
#> Fixed-Effects:   ----------------- ----------------- -----------------
#> birth_weight_cat                No               Yes               Yes
#> ________________ _________________ _________________ _________________
#> S.E. type        Heteroskeda.-rob. Heteroskeda.-rob. Heteroskeda.-rob.
#> Observations               263,905           263,905           263,905
#> R2                         9.11e-5           0.01902           0.01317
#> Within R2                       --           1.58e-5           1.22e-5
#>
#>                                       m4                     m5
#> Dependent Var.:  week_to_month_cond*1000 post_neonatal_cond*1000
#>
#> Constant
#> us                     -0.0385 (0.0600)        0.5316* (0.2911)
#> Fixed-Effects:   ----------------------- -----------------------
#> birth_weight_cat                     Yes                     Yes
#> ________________ _______________________ _______________________
#> S.E. type        Heteroskedasticity-rob. Heteroskedasticity-rob.
#> Observations                     263,678                 263,662
#> R2                               0.00403                 0.00976
#> Within R2                        2.54e-6                 8.75e-6
#> ---
#> Signif. codes: 0 '***' 0.01 '**' 0.05 '*' 0.1 ' ' 1
```

```r
m1 <- feols(die*1000 ~ us, data = d[country %in% c("US", "AT")], vcov = "hetero")
m2 <- feols(die*1000 ~ us | birth_weight_cat, data = d[country %in% c("US", "AT")], vcov = "het
m3 <- feols(lt_week_cond*1000 ~ us | birth_weight_cat, data = d[country %in% c("US", "AT")], vc
m4 <- feols(week_to_month_cond*1000 ~ us | birth_weight_cat, data = d[country %in% c("US", "AT"
m5 <- feols(post_neonatal_cond*1000 ~ us | birth_weight_cat, data = d[country %in% c("US", "AT"
etable(m1, m2, m3, m4, m5, signif.code = c("***"=0.01, "**"=0.05, "*"=0.10))
```

```
#>                                 m1                m2                 m3
#> Dependent Var.:           die*1000          die*1000  lt_week_cond*1000
#>
#> Constant         2.943*** (0.2548)
#> us               1.609*** (0.2907) 1.136*** (0.2868) 0.3502*** (0.1125)
#> Fixed-Effects:   ----------------- ----------------- ------------------
#> birth_weight_cat                No               Yes                Yes
#> ________________ _________________ _________________ __________________
#> S.E. type        Heteroskeda.-rob. Heteroskeda.-rob. Heteroskedas.-rob.
#> Observations               276,324           276,324            276,324
#> R2                         8.29e-5           0.01892            0.01412
#> Within R2                       --           4.21e-5            2.04e-5
#>
#>                                       m4                      m5
#> Dependent Var.:  week_to_month_cond*1000 post_neonatal_cond*1000
#>
#> Constant
#> us                    -0.1613** (0.0687)      0.9500*** (0.2562)
#> Fixed-Effects:   ----------------------- -----------------------
#> birth_weight_cat                     Yes                     Yes
#> ________________ _______________________ _______________________
#> S.E. type        Heteroskedasticity-rob. Heteroskedasticity-rob.
#> Observations                     276,093                 276,071
#> R2                               0.00337                 0.00943
#> Within R2                        4.48e-5                  3.7e-5
#> ---
#> Signif. codes: 0 '***' 0.01 '**' 0.05 '*' 0.1 ' ' 1
```

2.5. Describe and comment on cross-country differences in the association between socio-economic status and infant mortality. Use plots and/or tables of descriptive statistics to support your discussion.

Guide to answer:

6 points

The students should explore the association between infant variable and variables describing socio-economic status, such as education/occupation, marriage rates, and immigrant status.

Examples of figures the students can make are in section "IV Demographics of Postneonatal Disadvantage" in the original paper.

2.6. Write down and interpret the regression equation(s) used in Panel A "Postneonatal mortality" of Table 6 in Chen at al. (2016) with and without birth weight controls.

Guide to answer:

10 points

The regression equation is the following, which is run separately for the US versus Austria and US versus Finland:

$$y_{ij} = \beta_0 + \beta_1 US_j + \beta_2 \times advantaged_i + \beta_3 \times advantaged_i \times US_j + e_{ij}$$

Without control variables, this is a staturated model, and therefore:

- $E[y_{ij}|US_j = 0, advantaged = 0] = \beta_0$ is the mean number of post-neonatal deaths per 1,000 births for the non-advantaged group in Austria or Finland.
- $E[y_{ij}|US_j = 1, advantaged = 0] = \beta_0 + \beta_1$ is mean the number of post-neonatal deaths per 1,000 births for the non-advantaged group in the US.
- $E[y_{ij}|US_j = 0, advantaged = 1] = \beta_0 + \beta_2$ is the mean number of post-neonatal deaths per 1,000 births for the advantaged group in Austria or Finland.
- $E[y_{ij}|US_j = 1, advantaged = 1] = \beta_0 + \beta_1 + \beta_2 + \beta_3$ is the mean number of post-neonatal deaths per 1,000 births for the advantaged group in the US.

Therefore, we have these interpretations:

- β_1 is the difference in the mean the number of post-neonatal deaths per 1,000 births in the non-advantaged group between the US and Austria or Finland. If $\beta_1 > 0$ it means that the US have higher post-neonatal death rate in the non-advantaged group compared to the non-advantaged group in Austria or Finland.
- β_2 is the difference in the mean the number of post-neonatal deaths per 1,000 births between the non-advantaged and advantaged group in Austria or Finland. If $\beta_2 < 0$, the advantaged group has a lower post-neonatal death rate compared to the non-advantaged group.
- β_3 is the difference in the mean the number of post-neonatal deaths per 1,000 births in the advantaged group compared to the non-advantaged group in the US:

$$E[y_{ij}|US_j = 1, advantaged = 1] - E[y_{ij}|US_j = 1, advantaged = 0] = (\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_1) = \beta_2 + \beta_3$$

and the difference in the mean the number of post-neonatal deaths per 1,000 births in the advantaged group compared to the non-advantaged group in the Finland or Austria:

$$E[y_{ij}|US_j = 0, advantaged = 1] - E[y_{ij}|US_j = 0, advantaged = 0] = (\beta_0 + \beta_2) - \beta_0 = \beta_2$$

That is, β_3 is a difference-in-differences:

$$\begin{aligned} &\{E[y_{ij}|US_j = 1, advantaged = 1] - E[y_{ij}|US_j = 1, advantaged = 0]\} \\ -&\{E[y_{ij}|US_j = 0, advantaged = 1] - E[y_{ij}|US_j = 0, advantaged = 0]\} \\ &= (\beta_2 + \beta_3) - \beta_2 = \beta_3 \end{aligned}$$

and tells you whether the difference in post-neonatal death rates for the advantaged group versus non-advantaged is different in the US compared to Austria or Finland. If $\beta_3 < 0$, it means that the difference between the advantaged and non-advantaged group is more pronounced in the US compared to Austria or Finland. For example, consider the scenario where both coefficients β_2 and $\beta_2 + \beta_3$ are negative, indicating lower post-neonatal death rates in the advantaged group compared to the non-advantaged group in both countries. If $\beta_2 + \beta_3 < \beta_2 < 0$, this implies that the difference in post-neonatal death rates between the high and non-high SES groups is more pronounced in the US than in Austria or Finland. Specifically, the advantaged group in the US experiences even lower death rates relative to their disadvantaged counterparts than the corresponding SES group difference in Austria. This highlights a stronger relative advantage for the high SES group in the US context compared to Austria.

2.7. Apply the regression models from your response to question **2.6.** to perform the analysis described in Panel A 'Postneonatal Mortality' of Table 6 in Chen et al. (2016). You should implement and analyze a total of four models. Document your process, present your results, and discuss the implications of the results. Tip: You can use the `linearHypothesis()` function from the `car` package to test the equality between the advantaged group in the US and the advantaged group in the comparison country.

Please note that because the data is simulated, your results will vary from those reported in the paper.

Guide to answer:

10 points

The students should document their process and results (preferably, in a self-explanatory table), and discuss the implications of the results. They should also explain and coduct the test. Here I just show the results for the regressions, below the regression results I give some more details on the test:

```
#** Needed variables

d[, `:=` (high_ses = case_when(
  country == "FI" & occupation_cat_fi == "upper white collar or entrepreneur" ~ 1,
  country == "US" & education_cat_us == "college degree or more" ~ 1,
  country == "AT" & education_cat_at == "university or teaching college" ~ 1, T ~ 0),
          white = case_when(
            country == "FI" ~ 1,
            country == "AT" & immigrant_at == 0 ~ 1,
            country == "US" & black_us == 0 ~ 1, T ~ 0))][
                            , high_ses_married_white := as.integer((high_ses == 1 & married =
                          ][
                            , us_high_married_white := high_ses_married_white*us
                          ]
```

```
AT_US <- feols(post_neonatal_cond*1000 ~ us + us_high_married_white+high_ses_married_white+us,
AT_US_brwt <- feols(post_neonatal_cond*1000 ~ us + us_high_married_white+high_ses_married_white

FI_US <- feols(post_neonatal_cond*1000 ~ us + us_high_married_white+high_ses_married_white+us,
FI_US_brwt <- feols(post_neonatal_cond*1000 ~ us + us_high_married_white+high_ses_married_white


etable(AT_US, AT_US_brwt, FI_US, FI_US_brwt)
```

```
#>                                           AT_US              AT_US_brwt
#> Dependent Var.:         post_neonatal_cond*1000 post_neonatal_cond*1000
#>
#> Constant                      2.370*** (0.2379)
#> us                            1.780*** (0.2795)       1.388*** (0.2762)
#> us_high_married_white        -2.759*** (0.7881)      -2.792*** (0.7845)
#> high_ses_married_white         -0.5971 (0.7611)         0.3787 (0.7579)
#> Fixed-Effects:          ----------------------- -----------------------
#> birth_weight_cat                             No                     Yes
#> _______________________ _______________________ _______________________
#> S.E. type               Heteroskedasticity-rob. Heteroskedasticity-rob.
#> Observations                            276,071                 276,071
#> R2                                      0.00046                 0.00964
#> Within R2                                    --                 0.00024
#>
#>                                           FI_US              FI_US_brwt
#> Dependent Var.:         post_neonatal_cond*1000 post_neonatal_cond*1000
#>
#> Constant                      2.199*** (0.2813)
#> us                            1.950*** (0.3173)        1.021** (0.3235)
#> us_high_married_white        -2.752*** (0.6620)      -2.753*** (0.6638)
#> high_ses_married_white         -0.6046 (0.6296)         0.3556 (0.6296)
#> Fixed-Effects:          ----------------------- -----------------------
#> birth_weight_cat                             No                     Yes
#> _______________________ _______________________ _______________________
#> S.E. type               Heteroskedasticity-rob. Heteroskedasticity-rob.
#> Observations                            263,662                 263,662
#> R2                                      0.00048                 0.00997
#> Within R2                                    --                 0.00022
#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Really good assignments can also show that the model without control variables is saturated, and therfore the coeffcients can simply be calculated by simple extractions of means among the groups:

```
summary(feols(post_neonatal_cond*1000 ~ us + us_high_married_white+high_ses_married_white+us, c
```

```
#> OLS estimation, Dep. Var.: post_neonatal_cond * 1000
#> Observations: 276,071
#> Standard-errors: Heteroskedasticity-robust
#>                         Estimate Std. Error   t value   Pr(>|t|)
#> (Intercept)             2.369612   0.237874  9.961612  < 2.2e-16 ***
#> us                      1.779874   0.279510  6.367844 1.9200e-10 ***
#> us_high_married_white  -2.759074   0.788120 -3.500831 4.6388e-04 ***
#> high_ses_married_white -0.597086   0.761121 -0.784482 4.3276e-01
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> RMSE: 58.0   Adj. R2: 4.482e-4
```

```
mean(d[country == "AT" & high_ses_married_white == 1]$post_neonatal_cond, na.rm = T)
```

```
#> [1] 0.001772526
```

```
mean(d[country == "AT" & high_ses_married_white == 0]$post_neonatal_cond, na.rm = T)
```

```
#> [1] 0.002369612
```

```
mean(d[country == "AT" & high_ses_married_white == 1]$post_neonatal_cond, na.rm = T)-mean(d[cou
```

```
#> [1] -0.0005970857
```

```
mean(d[country == "US" & high_ses_married_white == 1]$post_neonatal_cond, na.rm = T)
```

```
#> [1] 0.0007933258
```

```
mean(d[country == "US" & high_ses_married_white == 0]$post_neonatal_cond, na.rm = T)
```

```
#> [1] 0.004149486
```

```
mean(d[country == "US" & high_ses_married_white == 1]$post_neonatal_cond, na.rm = T)-mean(d[cou
```

```
#> [1] -0.00335616
```

```
## diff-in-diff
((mean(d[country == "US" & high_ses_married_white == 1]$post_neonatal_cond, na.rm = T)-mean(d[c
```

```
#> [1] -2.759074
```

Running an F-test, for whether the death rates in the advantaged group is the same in both countries:

- $E[y_{ij}|US_j = 0, advantaged = 1] = \beta_0 + \beta_2$ is the mean number of post-neonatal deaths per 1,000 births for the advantaged group in Austria or Finland.
- $E[y_{ij}|US_j = 1, advantaged = 1] = \beta_0 + \beta_1 + \beta_2 + \beta_3$ is the mean number of post-neonatal deaths per 1,000 births for the advantaged group in the US.

For these to be equal we need $\beta_1 + \beta_3 = 0$, this can be tested using a F-test.

We test the hypothesis that the sum of the coefficients for \texttt{US} and \texttt{advantaged_i\times US_j} in a regression is zero. The hypotheses are formulated as follows:

- Null Hypothesis (H_0) :: The combined effect of being in the U.S. and the additional effect of being in the U.S. while also being high SES, married, and white on the dependent variable is zero. Mathematically, this is expressed as:

$$\beta_1 + \beta_3 = 0$$

- Alternative Hypothesis (H_a) : The combined effect is not zero, implying that these characteristics together have a statistically significant effect on the dependent variable. Mathematically, this is:

$$\beta_1 + \beta_3 \neq 0$$

To perform this test, an F-statistic is computed based on the sum of the squares of the deviations of the predicted values from the actual values, adjusted for the degrees of freedom. If the p-value associated with this F-statistic is below a predetermined threshold (commonly 0.05), the null hypothesis is rejected, suggesting a significant combined effect of the predictors on the dependent variable.

```r
library(car)
linearHypothesis(AT_US, "us + us_high_married_white = 0")
```

```
#> Linear hypothesis test
#> 
#> Hypothesis:
#> us  + us_high_married_white = 0
#> 
#> Model 1: restricted model
#> Model 2: post_neonatal_cond * 1000 ~ us + us_high_married_white + high_ses_married_white +
#>     us
#> 
#>   Res.Df Df  Chisq Pr(>Chisq)
#> 1 276068
#> 2 276067  1 1.7658     0.1839
```

```r
linearHypothesis(FI_US, "us + us_high_married_white = 0")
```

```
#> Linear hypothesis test
#> 
#> Hypothesis:
#> us  + us_high_married_white = 0
#> 
#> Model 1: restricted model
#> Model 2: post_neonatal_cond * 1000 ~ us + us_high_married_white + high_ses_married_white + 
#>     us
#> 
#>   Res.Df Df  Chisq Pr(>Chisq)
#> 1 263659                     
#> 2 263658  1 1.9016     0.1679
```

Description of variables and names

The dataset `birth_data.csv` includes the data for conducting exercise 2.

Variable	Description
`country`	Country identifier (US = The United States; AT = Austria; FI = Finland)
`year`	Year identifier
`gestation`	Gestational age in weeks
`die`	Died within 1 year indicator (1 = Died; 0 = Survived)
`death_day`	Day of death, ranges from 1 to 365
`singleton`	Singleton birth indicator (1 = yes; 0 = no)
`birwt`	Birth weight in grams
`male`	Male infant (1 = Male; 0 = Female)
`mage`	Mother's age in years
`married`	Mother is married (1 = yes; 0 = no)
`black_us`	Mother is black in the US (1 = yes; 0 = no)
`education_cat_us`	Education categories in the US (1 = College degree or more; 2 = High school degree; 3 = Less than high school degree; 4 = Some college)
`immigrant_at`	Mother is an immigrant in Austria (1 = yes; 0 = no)
`education_cat_at`	Education categories in Austria (1 = Compulsory school; 2 = High school with A-levels; 3 = University or teaching college; 4 = Vocational school)
`occupation_cat_fi`	Occupation categories in Finland (Blue collar, Lower white collar, Upper white collar or Entrepreneur)