Guidelines for Final exam HMET4220 – Applied Micro Econometrics, Spring 2023

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 25 percent and the second exercise counts 75 percent. Plan your time thereafter.

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

Exercise 1: Physical exercise and health (25%)

A research article suggests that moderate exercise can significantly improve health. The claim is supported by an analysis of individuals, where the group that did the least exercise (4 minutes per day of movement) had the highest mortality rate. The group that did slightly more exercise (6 minutes per day of moderate exercise) had half the mortality rate of the least exercise group.

1.0. Define the causal effect of increasing moderate exercise from 4 to 6 minutes on mortality using the potential outcomes framework.

Guide to answer:

The students should define:

- Y_i^1 is the probability of death if individual i exercise 6 minutes per day
- Y_i^0 is the probability of death if individual i exercise 4 minutes per day
- D_i is a treatment indicator equal to 1 if individual i exercise 6 minutes per day
- D_i is a treatment indicator equal to 0 if individual i exercise 4 minutes per day
- The students should define the individual treatment effect: $\delta_i = Y_i^1 - Y_i^0$

And discuss at least one of the following average treatment effects:

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

1.1. Suppose you have a sample of people who exercised for either 4 or 6 minutes per day for a year, and you have tracked their mortality for one year. Write out the linear regression equation with mortality as the dependent variable (=1 if the individual died) and an indicator variable that takes a value of 1 for the group that exercised for 6 minutes (treatment group) and 0 for the group that exercised for 4 minutes (control group). Interpret the regression equation.

Guide to answer:

- The regression: $y_i = \beta_0 + \beta_1 D_i + \epsilon_i$
- The students should explain each parameter.
- Good students will recognize that without a causal interpreation the ϵ_i is the CEF residual, which is mean independent of D_i in the population.

1.2. Explain in a few sentences whether we can obtain an estimate of the mean difference in mortality between the two exercise groups that is both unbiased and consistent using the regression model you described in the previous exercise?

Guide to answer:

- Here we are simply asking whether the linear regression can be used to estimate the difference in means between the two groups.
- Students should recognize that β_1 will be the difference in means between the treatment and control group, but the parameter should not be given a causal interpretation unless the treatment is randomly assigned.
- The regression will be an unbiased and consistent estimator of the mean difference.
- The students should also explain what is meant by unbiasdness and consistency.

1.3. Use the potential outcomes framework to show the bias of the difference in means estimator of the ATE. What is the likely sign of the bias? Explain your answer.

Guide to answer:

- In the above regression model, β_1 is the difference in means of y_i between the exercise group, or, as it was named in the textbook "the simple difference in outcomes (SDO)".
- The students should then be able to show that the bias is given by the sum of selection bias (SB) and heterogenous treatment effect bias (HTEB), which, in terms of potential outcomes is given by:
- $SB = E[Y_i^0|D_i = 1] - E[Y_i^0|D_i = 0]$.
- $HTEB = (1 - \pi)(ATT - ATU)$, where π is the share that receives treatment.
- $ATT = E[Y_i^1|D_i = 1] - E[Y_i^0|D_i = 1]$
- $ATU = E[Y_i^1|D_i = 0] - E[Y_i^0|D_i = 0]$
- $SDO = ATE + SB + (1 - \pi)(ATT - ATU)$
- Students should describe selection bias in detail, and explain why it is likely to be negative (i.e. the mortality rate of people choosing to exercise more is lower than people who choose to exercise less).
- Students should also discuss whether HTEB is likely to be positive or negative.
- HTEB is negative if people who choose to exercise more also have a higher return to exercise.

1.4. Suggest and justify a research design to identify the causal effect of physical exercise, and discuss potential pitfalls of your proposed research design.

Guide to answer:

- Randomized controlled trial (RCT). The RCT would involve randomly assigning participants to two groups: an exercise group and a control group.
    - Students should explain why a RCT solves the problem of selection bias.
    - Potential pitfalls is compliance and attrition i.e. whether people follow the treatment and risk of people dropping out.
    - Another pitfall is that people know whether they are in the control or treatment group and might experience a placebo effect.
- Students can also suggest any other natural experiment. If they do that they need to discuss the identifying assumption in their proposed strategy.

Exercise 2: The effects of Medicaid expansions on insurance coverage (75%)

In this exercise, you are asked to discuss and replicate some of the analysis in Carey, Miller and Wherry (2020): The Impact of Insurance Expansions on the Already Insured: The Affordable Care Act and Medicare, American Economic Journal: Applied Economics 2020, 12(4): 288–318.

The file `ehec.csv` is available on Inspera, and it contains a panel dataset at the state level about health insurance coverage and Medicaid expansion. You can perform analyses that are similar to those of Carey et al. (2020), but you cannot replicate their analysis since they used confidential data. The variable `dins` indicates the share of low-income childless adults with health insurance. The variable `yexp2` shows the year when a state expanded Medicaid coverage under the Affordable Care Act; it is missing if the state never expanded. The variable `year` identifies the observation year, and `stfips` is a state identifier. 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 "The Impact of Insurance Expansions on the Already Insured: The Affordable Care Act and Medicare" by Carey, Miller and Wherry (2020). 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 and its identifying assumptions. Focus on the following questions: What are the coefficient(s) of interest(s)? What is (are) the key identifying assumption(s)? Are the identifying assumptions likely to hold? 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 3 pages (12pt, double spaced).

Guide to answer:

- 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 identifying assumptions, coefficients of interest, threats to the identifying assumption, data limitations, suggestions for alternative analyses.

2.1. Load and describe the panel data. How many states are included in the panel data set? What is the time period covered by the panel data set? What is the frequency (annual, quarterly or weekly) of observations in the panel data set? What are the starting and ending years? How many observations are there in total for each state in the panel data set?

Guide to answer:

```r
library(data.table)
d ← fread("ehec.csv")
print("number of obs.")
```

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

```r
nrow(d)
```

```
#> [1] 552
```

```r
print("number of states:")
```

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

```r
length(unique(d$stfips))
```

```
#> [1] 46
```

```r
print("number of obs. per state:")
```

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

```r
table(d[, .(n = .N), by = stfips]$n)
```

```
#> 
#> 12 
#> 46 
```

```r
print("years of observations and how many states are observed each year")
```

```
#> [1] "years of observations and how many states are observed each year"
```

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

```
#> 
#> 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 
#>   46   46   46   46   46   46   46   46   46   46   46   46 
```

In a good answer, the candidate should provide accurate and complete information about the number of states, time period covered, frequency of observations, starting and ending years, and total observations per state in the panel data set. Answering the last question the candidate should be able recognize that the panel is balanced. The information should be presented in a clear and organized manner.

2.2. Describe and comment on the variation in the dependent variable "share of low-income childless adults with health insurance" across states and over time using plots and/or descriptive statistics tables. How has the share of low-income childless adults with health insurance changed over time in different states? Are there any states that consistently show higher or lower levels of health insurance coverage for low-income childless adults over time? What is the overall trend in health insurance coverage for low-income childless adults across all states over time?

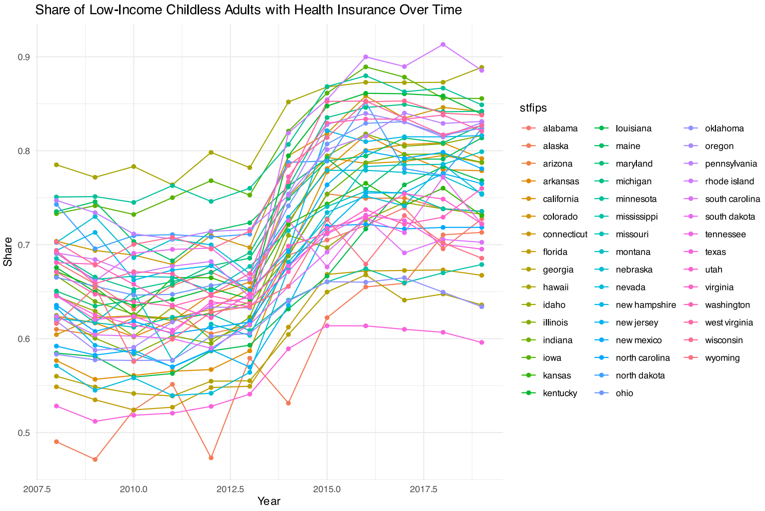
Guide to answer:

- Here the students should make some graphs and tables that answers all the questions above.
- The figures should be self explanatory, and should include a title, y- and x-labels.
- The students should also explain the figures in their answers. In their answers, they should
- Examples of tables and figures that can be included:

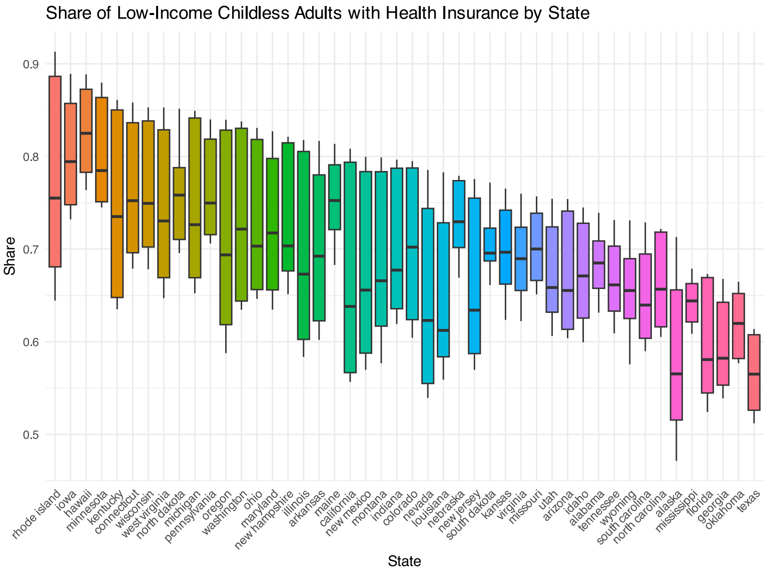
```
# Calculate descriptive statistics by state
summary_table ← d %>%
  group_by(stfips) %>%
  summarise(mean_share = mean(dins),
            median_share = median(dins),
            min_share = min(dins),
            max_share = max(dins),
            sd_share = sd(dins))

# Print the summary table
print(summary_table)
```

```
#> # A tibble: 46 × 6
#>    stfips      mean_share median_share min_share max_share sd_share
#>    <chr>            <dbl>        <dbl>     <dbl>     <dbl>    <dbl>
#>  1 alabama          0.686        0.685     0.631     0.739   0.0333
#>  2 alaska           0.582        0.565     0.471     0.713   0.0884
#>  3 arizona          0.674        0.655     0.604     0.754   0.0665
#>  4 arkansas         0.704        0.692     0.602     0.817   0.0830
#>  5 california       0.674        0.638     0.557     0.808   0.114
#>  6 colorado         0.705        0.702     0.604     0.795   0.0836
#>  7 connecticut      0.764        0.752     0.679     0.858   0.0735
#>  8 florida          0.600        0.581     0.524     0.673   0.0663
#>  9 georgia          0.595        0.582     0.539     0.668   0.0500
#> 10 hawaii           0.826        0.825     0.764     0.889   0.0486
#> # ℹ 36 more rows
```

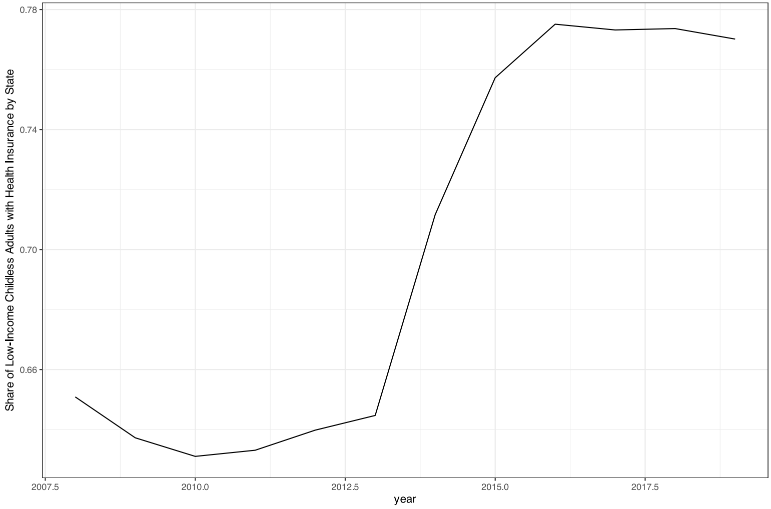
```r
# Load necessary packages
library(ggplot2)

# Create a line chart of share over time
line_chart <- ggplot(d, aes(x = year, y = dins, color = stfips)) +
  geom_line() +
  geom_point() +
  labs(title = "Share of Low-Income Childless Adults with Health Insurance Over Time",
       x = "Year",
       y = "Share") +
  theme_minimal()

# Print the line chart
print(line_chart)
```



```r
# Create a boxplot of share by state
boxplot <- ggplot(d, aes(x = group, y = dins, fill = group)) +
  geom_boxplot() +
  labs(title = "Share of Low-Income Childless Adults with Health Insurance by State",
       x = "State",
       y = "Share") +
  theme_minimal() +
  theme(legend.position = "none") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Print the boxplot
print(boxplot)
```



```
d %>%
  group_by(year) %>%
  summarise(dins = mean(dins)) %>%
  ggplot(aes(x = year, y = dins)) +
  geom_line() +
  ylab("Share of Low-Income Childless Adults with Health Insurance by State") +
  theme_bw()
```



2.3. You are interested in how the expansion of Medicaid coverage affected the share of low-income childless adults with health insurance. You consider running the following regression model:

$$dins_{jt} = \alpha_j + \gamma_t + \beta D_{jt} + \epsilon_{jt} \quad (1)$$

where $dins_{jt}$ is the share of low-income childless adults with health insurance in state s at year t, α_j is a state-specific effect, γ_t is a year-specific effect, D_{jt} is an indicator variable equal 1 for the treatment group in the after-treatment period, and ϵ_{jt} is an error term.

Describe the model in detail. What is the coefficient(s) of interest? Which variation are you exploiting to estimate the coefficient(s) of interest?

Guide to answer:

- Students should recognize that this is a two-way fixed effects model (controls for both state- and year-specific effects).
- The unit- and year-specific effects will control for unobserved, time-invariant differences between states and for any changes that affect all units equally over time.
- The parameter of interest is β, and we use unit-specific variation over time to estimate this paratmeter when we control for unit- and year-specific effects
- Good answers will include that β will be an unbiased estimator of the average treatment effect on the treated (ATT) under strict assumptions regarding parallel trends, and homogenous treatment effects across states and over time.

2.4. How many states expanded Medicaid during each year of the study period? Discuss whether the fact that states are treated in different years can cause a bias when estimating the causal effect using regression Equation 1.

Guide to answer:

```
d[!is.na(yexp2), .(n = length(unique(stfips))), by = .(yexp2)][order(yexp2)]
```

```
#>    yexp2  n
#> 1:  2014 22
#> 2:  2015  3
#> 3:  2016  2
#> 4:  2017  1
#> 5:  2019  2
```

- States expanded Medicaid in 2014, 2015, 2016, 2017 and 2019. The number of states that expanded in each of these year are shown in the table above.
- In class, students were told to only use the TWFE model when units are treated at the same time, and they were referred to section 18.2 in the textbook "The Effect".
- Good answers should recognize, but do not go into detail, that we have rollout design, and recognize that rollout design cause problems because already-treated groups are used as a control group, and this causes problems when effects varies across groups and/or over time.
- Really good answers mention this in exercise **2.0.** when critically assessing the given paper.

2.5. In the rest of the exercise we are only going to analyze the treatment effect of states that expanded Medicaid in 2014. Drop states that expanded Medicaid later than 2014, such that only never-treated can be used as control. How many states are in the treatment group, and how many states are in the (never-treated) control group?

Guide to answer:

```
d14 <- d[yexp2 == 2014 | is.na(yexp2)]
# Make treatment dummy
d14[, D := case_when(!is.na(yexp2) ~ 1, T ~ 0)]
d14[, .(n = length(unique(stfips))), by = "D"]
```

```
#>    D  n
#> 1: 0 16
#> 2: 1 22
```

- There are 22 states treated in 2014 ("the treatment group") and 16 states in the control group.

2.6. Does the pre-treatment level of the dependent variable predict whether states choose to expand Medicaid in 2014? Explain in a few sentences whether it is likely that the treatment is randomly assigned.

Guide to answer:

- There are many ways to do this. One way to do it is to collapse the data across in the pre-treatment period for the treatment and control states, and use a regression model to test whether the average level of the dependent variable predicts treatment:

```
d.agg <- d14[year < 2014, .(dins = mean(dins)), by = .(stfips, D)]
broom::tidy(lm(dins ~ D, data = d.agg))
```

```
#> # A tibble: 2 × 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)   0.619     0.0140     44.2  5.56e-33
#> 2 D             0.0355    0.0184      1.93 6.15e- 2
```

- We see that treatment sates that choose to expand Medicaid have on average higher share of low-income childless adults with health insurance compared to control states.
- In expectation, there should be no level difference between treatment and control sates if the treatment was randomly assigned.
- Bonus points for students that point out the sample size is so small (there aren't that many states) so by chance there could be level differences between treatment and control states even under randomized assignment.
- Irrespective of the regression that is run the students should interpret the size, significance, and sign of estimated coefficients.

2.7. Calculate the average of the dependent variable in the before and after treatment period for the treatment and control group, and calculate and interpret the 2×2 difference-in-differences estimate.

Guide to answer:

```
d14[, before.after := case_when(year < 2014 ~ "before", T ~ "after")]
t.did <- dcast(d14[, .(dins = mean(dins)), by = .(before.after, D)], formula = D ~ before.after
t.did[, post_pre_diff := after - before]
t.did
```

```
#>    D     after    before post_pre_diff
#> 1: 0 0.6979453 0.6189702    0.07897515
#> 2: 1 0.8083730 0.6544622    0.15391084
```

```
did.est <- t.did[D == 1,]$post_pre_diff - t.did[D == 0,]$post_pre_diff
did.est
```

```
#> [1] 0.07493569
```

- The students should interpret the size of the effect, which is a roughly 7.5 percentage point increase in the outcome in the treament group compared to the control group.

2.8. Estimate regression Equation 1. Remember to cluster your standard errors at the state level. Interpret the coefficient in front of `dins`, and comment on the size and significance of its estimate.

Guide to answer:

```
d14 <- d14 %>% mutate(Dpost = case_when(yexp2 == 2014 & year >= 2014 ~ 1, T ~ 0))
library(fixest)
twfe_dd <- feols(dins ~ Dpost | stfips + year,
                      cluster = "stfips",
                      data = d14)

broom::tidy(twfe_dd)
```

```
#> # A tibble: 1 × 5
#>   term  estimate std.error statistic       p.value
#>   <chr>    <dbl>     <dbl>     <dbl>         <dbl>
#> 1 Dpost   0.0749   0.00954      7.86 0.00000000211
```

- Students should interpret the coefficient its unit of measurement and size, and whether it is significant.

2.9. What is the key identifying assumption that allows us to interpret the estimate from the previous exercise as an estimate of a causal parameter? Does the potential level difference found in exercise **2.6.** pose a threat for the identification of the causal effect?

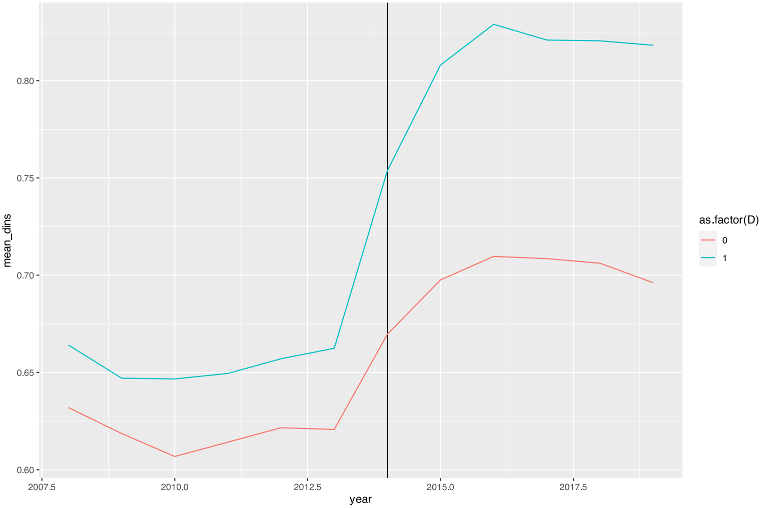
Guide to answer:

- The identifying assumption is the parallel trends assumption, which states that the outcome would have followed prarllel trends in treatment and control in the absence of treatment.
- The level difference itself does not need to be a problem for the identifying assumption, but you have to assume that states at different levels follow the same trend over time in the absence of treatment. Good students discuss whether this is realistic.

2.10. Create a graph that displays the average of `dins` in both the treatment and control groups throughout the entire study period. Compare and discuss the trends in the treatment and control group before and after the treatment.

Guide to answer:

```
d14 %>%
  group_by(D, year) %>%
  summarise(mean_dins = mean(dins)) %>%
  ggplot(aes(x = year, y = mean_dins, color = as.factor(D))) + geom_vline(xintercept = 2014) +
```



- The students should explain and describe the plot.
- The figure should be self explanatory: title, y- and x-labels, and preferable a vertical line at the treatment year.
- It is good if the control and treated lines are separated in shape or color.

2.11. Use a linear regression model to estimate linear trends in the pre-treatment period, and then perform a statistical test at a 95% level to examine whether there are any differences in the trends between the treated and control group.

Guide to answer:

```
tidy(lm(dins ~ year*D, data = d14[year < 2014]))
```

```
#> # A tibble: 4 × 5
#>   term        estimate std.error statistic p.value
#>   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
#> 1 (Intercept)  2.91      6.81        0.427   0.669
#> 2 year        -0.00114   0.00339    -0.337   0.737
#> 3 D           -3.66      8.96       -0.409   0.683
#> 4 year:D       0.00184   0.00445     0.413   0.680
```

- Students should interpret the coefficient (the coefficient of interest is the interaction between year and the treatment dummy), and comment on whether we can reject that there are differenecs in (linear) trends before the intervention.

2.12. Compute a relative year variable which is centered at the treatment year, which counts the number of years before the intervention (negative numbers) and after the intervention (positive numbers).

Guide to answer:

```
d <- d %>%
     mutate(relative_year = year - 2014)
```

2.13. Estimate an event-study model: A difference-in-difference model where you allow the effect to differ by year (using `year = -1` as the reference year), with standard errors clustered at the state level. Show and discuss the results.

Guide to answer:

```
library(fixest)
twfe_event <- feols(dins ~ i(year, D, ref = 2013) | stfips + year,
                      cluster = "stfips",
                      data = d14)
twfe_event <- summary(twfe_event)
twfe_event
```

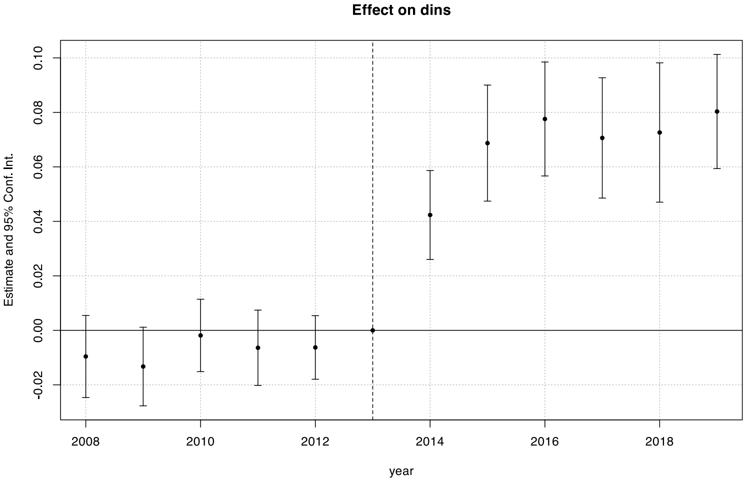
```
#> OLS estimation, Dep. Var.: dins
#> Observations: 456
#> Fixed-effects: stfips: 38,  year: 12
#> Standard-errors: Clustered (stfips)
#>               Estimate Std. Error   t value   Pr(>|t|)
#> year::2008:D -0.009596   0.007686 -1.248526 2.1968e-01
#> year::2009:D -0.013277   0.007359 -1.804314 7.9327e-02 .
#> year::2010:D -0.001871   0.006777 -0.276090 7.8402e-01
#> year::2011:D -0.006401   0.007050 -0.907921 3.6980e-01
#> year::2012:D -0.006286   0.005951 -1.056413 2.9763e-01
#> year::2014:D  0.042340   0.008322  5.087889 1.0748e-05 ***
#> year::2015:D  0.068713   0.010863  6.325215 2.2809e-07 ***
#> year::2016:D  0.077578   0.010665  7.273756 1.2276e-08 ***
#> year::2017:D  0.070620   0.011267  6.267763 2.7270e-07 ***
#> year::2018:D  0.072612   0.013049  5.564663 2.4409e-06 ***
#> year::2019:D  0.080320   0.010693  7.511587 5.9652e-09 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> RMSE: 0.019887     Adj. R2: 0.944044
#>                  Within R2: 0.478908
```

- The students should discuss the size of the effect estimates, and whether there are signs of significant pre-treatment differences.

2.14. Create a plot of the coefficients you estimated in **2.13.** with the relative years on the x-axis and the coefficient values on the y-axis. Include the 95% confidence intervals around each point estimate by using cluster-robust standard errors. Describe the impact of the Medicaid expansion using graph and evaluate whether this plot raises any concerns about prior trends violations.

Guide to answer:

```
iplot(twfe_event)
```



- The plot should be self-explanatory, and the students should describe what the figure shows for example, the effect estimates before and after the treatment. Is there evidence of trend differences before treament?

Description of variables and names

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

Variable	Description
`stfips`	State identifier
`year`	Year identifier
`dins`	Share of low-income childless adults with health insurance
`yexp2`	Year when a state expanded Medicaid coverage under the Affordable Care Act