PPHA 31002: Homework 4

Professors Frank, Moskowitz & Vellekoop

Harris School of Public Policy

Due by December 5, 2025 at 11:45pm CT

Instructions

Please *upload* the following by 11:45 PM (CT) on the due date:

1. Your write-up through the **PDF** Gradescope portal – Your write-up should be formatted in a clean and professional way; it should be organized, clearly labeled, and fully legible. Additionally, you must always accurately tag the location of each question in your PDF write-up using Gradescope. *Note: the answers that you upload to Gradescope will be the answers that are graded.* You are responsible for ensuring that you have uploaded the most up-to-date and most complete file.

2. Your R script through the **R-Script** Gradescope portal – This script should be able to reproduce all answers you included in your write-up that were generated using R, including numerical quantities, graphs/figures, etc. The script should be organized and appropriately commented so that a skilled R user can easily follow your code. Note that this script should be comprehensive (meaning that it produces everything included in your write-up that you generated in R) and executable (meaning that, after changing the working directory to the appropriate location, it should run all the way through without triggering errors).

Your PDF must contain the answers to all questions, including numerical answers, written explanations, and relevant graphs and figures; your grader will not look to your R script for your answers. In other words, you should consider the R script to be a supporting document for a technical reader, but the reader is not expected to reference your script to find an answer to a question. You must *show your work.*

Grading Policy: To ensure that we can expeditiously return assignments back to students, only some questions will be graded based on correctness while the remaining questions will be graded based on completion. In other words, you will receive full credit for questions graded based on completion if you make a good-faith attempt to answer the question and show the work that generated your answer, *even if your answer is incorrect.* We will randomly select a subset of questions to be graded for correctness; students will not be told in advance which questions will be graded for correctness. With that in mind, you should strive to complete all of the questions correctly. Given that not all questions are graded for correctness, it is imperative that you look closely through the posted solutions, as a high score on an assignment may only indicate complete answers rather than correct answers.

Collaboration Policy: You are permitted to work together on this assignment in small groups of 2-4 students. All students are required submit their own write-ups to Gradescope. Note *your* write-up and code should only reflect *your* understanding of the material. As such, these should be written in *your own words.* If your write-up contains identical language to another person in the course (including a member of your group), it is considered a violation of the school's academic integrity policies. Please indicate at the top of your write-up the names of students with whom you collaborated.

Pen & Paper Exercises

Is Hand Sanitizer Effective? Evidence from One Study

Hand disinfection is frequently recommended for preventing transmission of viruses. In particular, alcohol-based hand sanitizer has been found in a laboratory setting to have both immediate and persistent ability to inactivate rhinovirus (RV), one of several viruses that cause "the common cold." To determine effectiveness in a natural setting, a total of 212 volunteers were assigned at random either to a treatment group, which used alcohol-based hand sanitizer every three hours or after each hand washing, or to a control group, which was asked to use routine hand washing but not hand sanitizer.[1] Here are the data from the study indicating the count of individuals from the treatment and control groups who had a rhinovirus infection during the 10-week study period:

	RV Infection	
	Yes	No
Treatment Group	49	67
Control Group	49	47

1. Calculate the proportion of the treatment group who had an infection. [1pt]
2. Calculate the proportion of the control group who had an infection. [1pt]
3. Calculate the difference between the proportion of people who had an infection in the treatment group and the control group. [1pt]
4. Conduct a formal hypothesis test examining whether the treatment group had a *lower* infection rate than the control group. Be sure to report all of the steps of the test. [3pt]
5. Suppose you are a public health official in the middle of a pandemic. You have a sworn duty to only recommend evidence-backed treatments to reduce the spread of viruses. A reporter asks you if the evidence supports alcohol-based hand sanitizers as being effective at preventing disease. What do you say? [2pt]
6. Suppose you learn that there was noncompliance in the study. Of the 116 individuals assigned to the treatment group, 93 used hand sanitizer. Of the 96 individuals assigned to the control group, 20 used hand sanitizer. Estimate the intent-to-treat effect as well as the complier average treatment effect based on this information. If it is not possible to estimate a particular quantity, briefly explain why that is so. [2pt]

Does Tylenol Cause Autism?

Tylenol and Autism

On September 22nd of this year, the Food and Drug Administration (FDA) announced that it "initiated the process for a label change" on acetaminophen that would include a warning indicating that *"the use of acetaminophen by pregnant women may be associated with an increased risk of neurological conditions such as autism and ADHD in children."* Acetaminophen is a pain reliever and fever reducer. It is

[1] R.B. Turner et al., "A randomized trial of the efficacy of hand disinfection for prevention of rhinovirus infection," *Clinical Infectious Diseases*, 54 (2012), pp. 1422-1426.

commonly sold under the brand name Tylenol, and it is called paracetamol in many regions of the world. In this exercise, we will examine statistical analyses from one influential study published in the *Journal of the American Medical Association* (*JAMA*).[2]

The *JAMA* study uses detailed microdata from Sweden, where they track and record social, economic, and health data on the entire population. The population under study includes all children born in Sweden from July 1, 1995 to December 31, 2019 (to get a feel for the data, Sweden currently has a population of about 10.6 million). We will ignore the very important issue of measurement error in usage of acetaminophen and simply assume that it is measured accurately. The source of information for the autism diagnosis data is Sweden's National Patient Registry. The following table reports the number of children by their autism diagnosis status and by maternal usage of acetaminophen during the pregnancy. These numbers are taken directly from Table 1 in the study.

	Autism Diagnosis	
	Yes	No
Exposed to Acetaminophen	5,912	179,997
Not Exposed to Acetaminophen	62,672	2,232,216

7. Calculate the proportion of children diagnosed with autism who were exposed to acetaminophen in utero and the proportion of children diagnosed with autism not exposed to acetaminophen in utero. Based only on those two proportions (i.e., only comparing those proportions without further analysis), can we reach a conclusion as to whether there is a *true* difference in the incidence of autism diagnoses between children exposed and not exposed in utero to acetaminophen? You can assume that these data represent a sample from a broader population. Briefly explain in 1-2 sentences. [1pt]

8. Formally conduct a **two-sided** hypothesis test that there is no difference in the proportion of children diagnosed with autism between the group exposed and not exposed to acetaminophen. Be sure to report all of the steps of the hypothesis test. [3pts]

9. Please elaborate on your finding using the concepts of practical significance (sometimes, also referred to as economic significance) and statistical significance. [2pts]

10. We have thus far examined the association between acetaminophen exposure in utero and autism. As we have emphasized throughout the course, the presence of a correlation does not imply a causal relationship. This STAT News article summarizes the evidence on the possibility of a causal link between acetaminophen and autism. After reading the article, describe one possible source of *selection into* acetaminophen usage by pregnant women, which would make the comparison between the acetaminophen-exposed group and the non-exposed group *invalid*.[3] [1pt]

11. One possible criticism on the statistical analysis so far is that no other characteristics have been taken into account that may also affect autism diagnoses. For example, in the same data set it is

[2] Ahlqvist, V.H., Sjöqvist, H., Dalman, C., Karlsson, H., Stephansson, O., Johansson, S., Magnusson, C., Gardner, R.M. and Lee, B.K., 2024. Acetaminophen use during pregnancy and children's risk of autism, ADHD, and intellectual disability. *JAMA*, 331(14): 1205-1214, link

[3] It might be useful to consider the example we discussed in lecture of reemployment services. In that setting, we noted that individuals who have a high motivation to find a job might select into receiving reemployment services, while individuals with a low motivation to find a job opt out. For this question, you need to describe one source of *selection into* acetaminophen usage, which could also affect autism risk.

shown that women who are older are more likely to use acetaminophen during pregnancy. The age of the mother has been suggested as one potentially important risk factor for autism.

One of the analyses in the *JAMA* article attempts to use regression to control for observable differences between mothers who select into using acetaminophen and those who do not[4]. You will learn more about using covariate adjustment for the purposes of causal inference in Stats II (i.e., using regression to attempt to control for differences between treated and control units). In other words, these are OLS estimates generated from a multiple regression of a binary dependent variable (autism diagnosis) on a binary explanatory variable of interest (acetaminophen exposure in utero), controlling for a host of observable characteristics. Consider the estimated coefficient below from this regression (coefficients for all the background variables are estimated, but not shown):[5]

Coefficient	Estimate	Estimated Std. Error	t-statistic
$\hat{\beta}_{acetaminophen}$	0.0009	0.00054	1.67
Sample size (n): 2,480,797			

The estimated coefficient on acetaminophen represents the predicted difference in the probability of an autism diagnosis for children exposed to acetaminophen relative to the group of children who were not exposed, controlling for various observable characteristics.

Construct a 95% confidence interval (you can use 1.96 as the critical value given the enormous sample size, as the appropriate critical value from t distribution would closely approximate 1.96) for $\widehat{\beta}$ reported above, and determine based on the confidence interval whether you would reject the null hypothesis that $\beta = 0$ for $\alpha = 0.05$. [1pt]

12. The *JAMA* study does not have the benefit of random assignment to treatment, but the authors do implement a sibling design as part of the empirical analysis. The STAT News article describes the analysis: *"When it comes to Tylenol and autism, (...) [the researchers] tried to deal with this [selection] issue by looking at cases in which a mother took acetaminophen in one pregnancy, but not another. The idea was that if genetic or other environmental factors were leading to an increase in autism risk, both siblings would be exposed to the same risk."* In other words, if there is an unobserved factor that influences autism diagnoses *and* this factor is genetic or familial, siblings would have the same exposure to this factor and thus have the same risk of an autism diagnosis. In this research design, one can think of the group of siblings exposed to acetaminophen in utero as something akin to a treatment group and the group of siblings who were not exposed in utero as a control group. This analysis is able to hold constant many genetic and familial factors. Below we report the estimated coefficient from the sibling-design regression, where all of the comparisons are made *within* siblings (who have the same parents). Coefficients for all of the background variables are estimated, but not shown. The estimated coefficient reported below represents the predicted change in the probability of an autism diagnosis for a sibling exposed to acetaminophen in utero relative to a sibling who was not exposed in utero.

[4]See the note to Figure 3 in JAMA for a list of all of these controls. The researchers have access to a large list of variables for the Swedish population.

[5]The actual analysis in the paper uses a hazard model. The study reports absolute risk difference as well as a hazard ratio. We are reporting the estimates for the absolute risk difference. The interpretation of the two is different, and both the statisical technique and interpretation are well outside the scope of Stats I. We will make sure to note where the differences between the two are important.

Coefficient	Estimate	Estimated Std. Error	t-statistic
$\hat{\beta}_{acetaminophen}$	0.0002	0.0008	0.25
Sample size (n): 780,838			

Interpret the regression coefficient in words. Is the coefficient statistically significant from zero? What is your conclusion [2pt]

Concluding Thoughts: Taken together, while there is evidence of an *unadjusted* association between acetaminophen and autism, the magnitude of this estimated relationship erodes as we adjusted for observable differences and eventually compared within siblings. The lack of strong evidence for a positive effect, correlational or causal, however, is not the same as there being evidence of no effect. This presents a challenging problem for health officials making recommendations about pain reliever use during pregnancy, as all over-the-counter alternatives are generally considered unsafe during pregnancy.[6]

In other words, the *precautionary principle* might push a regulatory agency to issue a warning if they think there is a non-negligible degree of risk of harm from using the product. In particular, it might be reasonable for a physician to recommend against use to a patient for situations like a low-grade fever or very mild pain. On the other hand, in this particular setting, there are also potential serious harms that could result from the policy. First, because there is no safe alternative available, the saliency of this new warning might result in more pregnant women using these unsafe alternatives. Second, there is not yet any evidence of a causal linkage. So, the harms from use remain only potential harms. There are actual harms in the form of pain and suffering that will occur if pregnant women do not use acetaminophen. And, third, there is strong evidence of serious risk to the mother and fetus from high fevers. In such circumstances, acetaminophen is the only over-the-counter option available to treat high fevers. If pregnant women interpret the warning/guidance to discourage use broadly, they may not use it when they have a high fever, which would result in serious risk of harm.

Has Banning Neonicotinoids in the European Union Lowered Agricultural Yields?
Intensive agriculture uses a variety of chemicals, from fertilizers to pesticides. There are concerns that these chemicals have adverse health effects on people and nature. Neonicotinoids are a class of insecticides that has drawn attention due to their potential harmful effects on non-target insects (in other words, the insects that are not crops pests)—namely, bees and other pollinators.

As a results of numerous studies linking neonicotinoids to declines in bee populations, there have been growing calls to limit or to completely ban the use of this class of insecticides. In 2018, the European Union (EU) announced that starting in 2018, they are expanding a previous restriction on the use of neonicotinoids to cover all field crops.[7] The agricultural lobby strongly opposed this new policy arguing that there are no good and economically viable alternatives to neonicotinoids. Farmers argued that agricultural production would decline as a result.[8]

A team of researchers is working on testing the environmental and agricultural consequences of the EU 2018 neonicotinoids ban. As part of their analysis, they produce a simple descriptive regression analysis that aims to quantify the overall change in agricultural production in the years after the ban, relative to the years before the ban.

[6] Indeed, in his recent notice to physicians, the FDA Commissioner stated: *"This consideration should also be balanced with the fact that acetaminophen is the safest over-the-counter alternative in pregnancy among all analgesics and antipyretics; aspirin and ibuprofen have well-documented adverse impacts on the fetus."*

[7] You can read more about the decision and the background here.

[8] You can read more about the opposition and the arguments against this policy here.

To do so, the researchers obtain annual data on the total production and the total planted area of a variety of crop groups in different administrative regions in France.[9] For each crop group, they calculate the yield (the amount produced divided by the planted area), and construct a crop yield index.[10,11] As a simple first step, the researchers use data from 2015 to 2022, and estimate the following regression model:[12]

$$(\text{Crop Index})_{it} = \alpha + \beta(\text{Post-2018})_t + \varepsilon,$$

where the crop yield index is the dependent variable, and the explanatory variable of interest is a binary indicator that is coded = 1 for observations in which the 2018 ban on neonicotinoids was in effect.

The researchers present you with the following regression table that estimates the above model using the data we described earlier:

Coefficient	Estimate	Estimated Std. Error	t-statistic
$\hat{\beta}_{Post-2018}$	-0.091	0.028	-3.24
Sample size (n): 756			

While the researchers scratch their heads as how to interpret the results, they ask for your help to determine whether they can reject that their findings are likely to be observed just due to random chance. For simplicity, you tell them you are also going to focus only on a **two-sided test**.

13. State the null and alternative hypotheses for a test that seeks to evaluate whether the prediction for the crop index changes after the 2018 EU neonicotinoids ban. [1pt]

14. Using the critical value approach, can you reject the null hypothesis at the 5% significance level? **Throughout the question, use $n-2$ as the number of degrees of freedom.**[13] [1pt]

15. Calculate the p-value of observing the t-statistic of -3.24. Based on this p-value alone (as in do not use the critical threshold), can you reject the null hypothesis at the 5% significance level? [1pt]

16. Construct a 95% confidence interval for the coefficient, $\hat{\beta}_{Post-2018}$. The general form of a confidence interval for a regression coefficient is $\hat{\beta} \pm t_{\alpha/2,df} \times \hat{SE}$. Using only this 95% confidence interval alone, determine whether you (fail to) reject the null hypothesis at the 5% significance level? [2pt]

How Do Minimum Wage Increases Affect Employment?

Among the many topics economists study, few have attracted as much sustained attention—or generated

[9] The data are publicly available here.

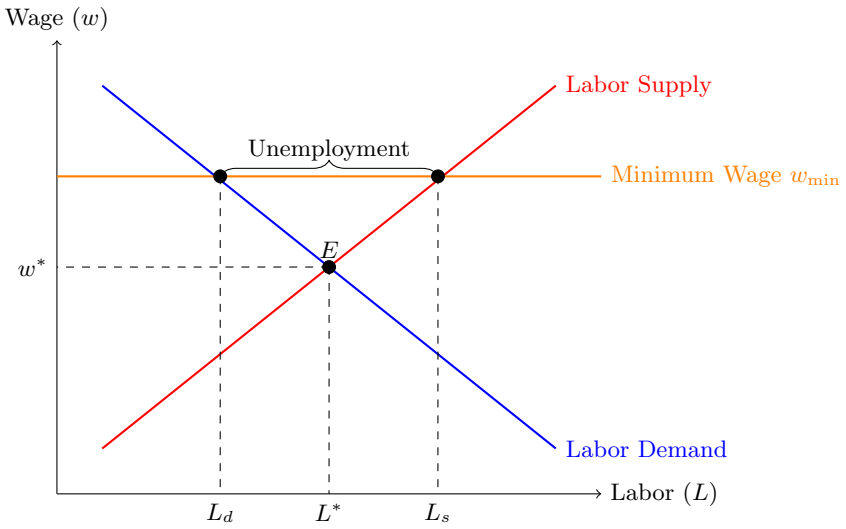
[10] They construct an index of crop group yields by standardizing each crop group. In other words, for each crop group value, they subtract the sample mean, and divide by the sample standard deviation. Effectively, this transforms each yield value to a z-score. For each administrative unit and year pair, the researchers take the mean of the z-scores for all the crop groups. This provides a single annual index at the local administrative unit level.

[11] This exercise lumps together crops that might benefit from pollination and crops that do not benefit from pollination. While that is an interesting dimension to explore, the first step here is just to establish whether the researchers can reject the null hypothesis that crop yields are the same after the ban. The benefit of using such an index is that it resolves the issue that different crops have different yield baselines. It also allows to run one regression and run one hypothesis test instead of multiple hypotheses tests where the researchers might end up making a Type-I error because they are testing a large number of crop groups.

[12] The model includes some additional variables we do not list here. Those variables simply subtract the mean level of the outcome across all years for each administrative unit. Meaning, you can think of calculating the mean for the first administrative unit for all the years in the data, then for the second administrative unit, and so on. Then, you subtract those means from the annual values for each administrative level. There is a way to do this quickly in a regression, but we did not cover it in Stats I. You can ignore all of these details when answering the questions, but just know the regression is doing all of that in the background, and you are once again, looking at numbers that are derived from an analysis on **real world data**.

[13] We use $n-2$ because the regression is estimating two parameters: intercept, and slope.

as much debate—as the effect of minimum wage increases on employment. Standard labor demand theory, which you encountered in microeconomics this quarter, predicts that raising the wage above the competitive equilibrium level level should reduce the quantity of labor demanded by firms, and therefore decrease employment. Yet decades of empirical research have produced mixed results: some studies find little to no employment effect, while others uncover evidence consistent with the textbook prediction.



A seminal study by David Card and Alan Krueger (1992) compared fast-food employment in New Jersey—where the minimum wage increased—to employment in Pennsylvania, where it did not. They found no evidence that the minimum wage increase reduced employment in New Jersey. Their paper is widely regarded as one of the earliest and most influential applications of natural experiments in economics, and it continues to serve as a benchmark in empirical research on the minimum wage. You can read more about the Card and Krueger study, as well as explore accompanying replication code in R, in this blog post.

In April 2024, California raised its minimum wage to $20 per hour for fast-food workers employed by large chains. Some research has found little to no effect on employment (see this summary), while other work has documented slower employment growth in California's fast-food sector relative to the rest of the country.[14]

In the questions below, we will use data summarized by Jeffrey Clemens, Olivia Edwards, and Jonathan Meer in their working paper, "Did California's Fast Food Minimum Wage Reduce Employment?"[15] Using a graphical summary of a set of regression coefficients, we will evaluate the evidence on how employment in California changed following the April 2024 policy implementation.

17. State the null and alternative hypotheses for a two-sided hypothesis test that evaluates the mean levels of employment in California versus the rest of the United States, after the 2024 increase in minimum wages in the fast food sector. State your hypotheses as statement on the mean employment mean. No need to state them as population regression slopes. [1 pt]

[14] A key limitation of the study referenced by the California government is that it relies on data only through July 2024, which is roughly three months after the policy went into effect, providing a relatively short window for evaluation.

[15] Working papers are pre-prints that have not yet completed the peer-review process. Their findings may evolve prior to publication. Working papers are common in many fields, including economics, because they allow researchers to disseminate results, solicit feedback, and build scholarly visibility while formal peer review proceeds, often over several years.

18. ***Note*: This question will only be graded on completion.** The figure below is reproduced from Clemens et al. (2025). The specific method used to generate the plot is more advanced than what we cover in Stats I (you will learn it in Stats II). For present purposes, you only need the following intuition: each circular point represents the estimated change in employment (in percentage terms) for fast-food workers in large chains in California, relative to (i) fast-food workers outside California, and (ii) other labor sectors both inside and outside California (each comparison made relative to September 2023). Values near zero indicate that employment levels in California are evolving similarly to employment levels outside California. Negative values indicate that employment in California is falling behind the trend observed outside California. The vertical lines around each point represent 95% confidence intervals.

    Based on the confidence intervals for the estimate six months after September 2023, can we reject the null hypothesis you stated earlier at the 5% significance level? [2pt]

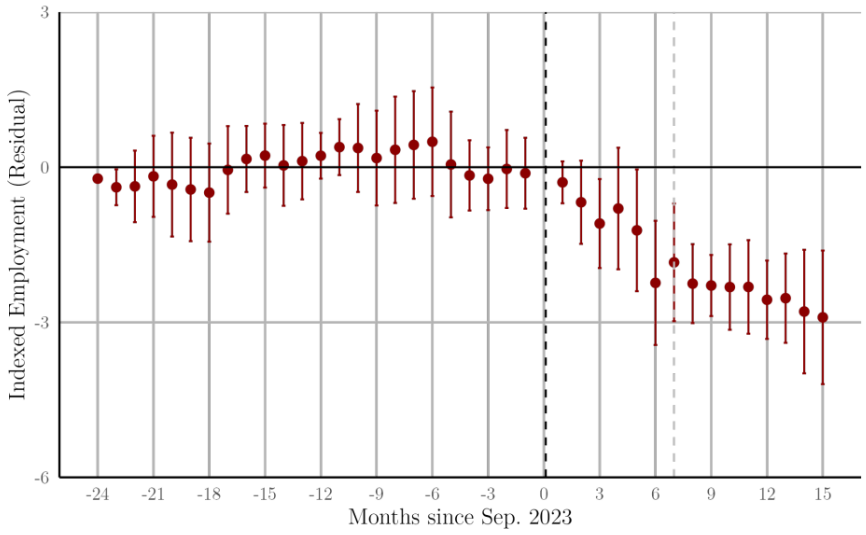


Figure A6: **Full-service restaurant triple-difference event-study estimates in California versus all other states and non-full-service industries (September 2021–December 2024).** This figure presents coefficient estimates from a triple-difference event-study regression of indexed full-service restaurant employment (NAICS 722511) for California relative to all other states and relative to non-minimum-wage-intensive industries, using QCEW data. All series are indexed to 100 at September 2023, and estimates are obtained from Equation 4. We first detrend each geography-industry series by removing the 24-month compound-monthly growth trend over September 2021 – September 2023. Then, we adjust the detrended series for seasonality by subtracting month-of-year fixed-effects estimated on September 2021 – September 2023 data. The zero event-time point corresponds to September 2023 (the AB 1228 announcement). Confidence intervals are constructed using cluster robust standard errors which, for reasons discussed in the main text, are insufficiently conservative in this setting and thus do not underlie the inference procedures we describe and utilize in the main text.

Source: "Did California's Fast Food Minimum Wage Reduce Employment? Jeffrey Clemens, Olivia Edwards, and Jonathan Meer." NBER Working Paper No. 34033, July 2025.

Data Exercise

Randomized Controlled Trial: How Has Prohibiting Criminal-Record Questions Affected Labor-Market Discrimination?

In this assignment, we will "replicate" a study evaluating how policies that prohibit employers from asking about applicants' criminal histories affect racial discrimination in labor markets.[16] The study, "Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment," by Amanda Agan and Sonja Starr, was published in 2018 in the Quarterly Journal of Economics. To get you started, here is abstract of the paper:

> "Ban the Box" (BTB) policies restrict employers from asking about applicants' criminal histories on job applications and are often presented as a means of reducing unemployment among black men, who disproportionately have criminal records. However, withholding information about criminal records could risk encouraging racial discrimination: employers may make assumptions about criminality based on the applicant's race. To investigate BTB's effects, we sent approximately 15,000 online job applications on behalf of fictitious young, male applicants to employers in New Jersey and New York City before and after the adoption of BTB policies. These applications varied whether the applicant had a distinctly black or distinctly white name and the felony conviction status of the applicant. We confirm that criminal records are a major barrier to employment: employers that asked about criminal records were 63% more likely to call applicants with no record. However, our results support the concern that BTB policies encourage racial discrimination: the black-white gap in callbacks grew dramatically at companies that removed the box after the policy went into effect. Before BTB, white applicants to employers with the box received 7% more callbacks than similar black applicants, but BTB increased this gap to 43%. We believe that the best interpretation of these results is that employers are relying on exaggerated impressions of real-world racial differences in felony conviction rates.

19. Read the first three pages of the paper and provide a high-level summary in 5-7 sentences. Your summary should address the following points: what a "Ban the Box" (BTB) policy does, one reason why the policy could succeed in achieving its goal, and one reason why it might fail. You do not need to include all of the details; focus on the main ideas. [2pts]

    First, we will load the data file `hw4_data.csv` into R. This data set contains information from the fictitious resumes used in the study: whether the applicant received a response, the race signaled by the resume, and whether the application was submitted before or after the implementation of the Ban the Box (BTB) policy.[17] The dataset includes the following variables:[18]

Name	Description
`response`	`=1` if the fictitious applicant received a response
`post`	`=1` if resume was sent after the enactment of the BTB policy
`black`	`=1` if the resume signaled a Black applicant

20. Each row is a different fictitious resume. How many resumes are there in the sample? What share of these resumes received a positive response (i.e. a value `=1` in the `response` variable)? [2pt]

21. We are going to start our analysis with two different two-sample t-tests. Each test assesses whether there is a difference in the `response` outcome between fictitious Black and white applicants. One test compares the difference in response for resumes sent before the BTB policy, and the other test

[16] We use the term "replicating" loosely: you will conduct a simplified version of the authors' analysis, but one that preserves the core logic of their approach.

[17] For more details on how the study signaled a white or Black name, see Section II.C on page 201 of the paper (page 11 of the PDF).

[18] This is a simplified version of the data. The full dataset is available at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VPHMNT.

uses the resumes sent after the BTB policy. For each test, conduct a two-sided t-test comparing using the `response` variable as the outcome.

Can you reject the null that the proportion of responses is the same across the Black and white resumes before the BTB policy? What about after the BTB policy? Write down the null and alternative hypotheses (you only need to do this once). Then for each test, report the p-value, the decision to reject the null or not, and your interpretation of the result of the test. *Throughout the exercise, use a significance level of* $\alpha = 0.05$. [3pt]

22. We will get very close to replicating the result reported in Table 4, column 2, in the paper:

TABLE IV

EFFECTS OF THE BOX ON RACIAL DISCRIMINATION: DIFFERENCE-IN-DIFFERENCES

	(1)	(2)	(3)	(4)	(5)
Box × white	−0.030**	−0.036**	−0.033**	−0.027**	0.002
(*White* × *pre, column* (5))	(0.015)	(0.014)	(0.014)	(0.013)	(0.014)
White	0.032***	0.044***	0.040***	0.123	0.022**
	(0.012)	(0.013)	(0.012)	(0.132)	(0.009)
Box	0.015	0.003	−0.002	−0.345**	−0.016
(*Pre, column* (5))	(0.024)	(0.015)	(0.013)	(0.139)	(0.017)
N	7,245	3,712	4,794	4,794	7,476
Controls	Yes	Yes	Yes	Yes	Yes
Center FE	Yes	No	Yes	Yes	No
Chain FE	No	No	No	Yes	No
Post × chain FE	No	No	No	Yes	No
White × chain FE	No	No	No	Yes	No
Box variation	Cross-section	Temporal	Temporal	Temporal	None
Sample	Pre-BTB	Box remover -balanced	Box remover -full	Box remover -full	Other empl. balanced

Notes. Standard errors clustered on chain in parentheses. Dependent variable is whether the application received a callback. Box removers are stores that had the box in the pre-BTB period and removed it after BTB. "Box removers-balanced" consists of box remover stores to which we sent exactly four applications, one white/black pair in each period. Fixed effects can include geographic center, chain, post × chain, and white × chain, and are included as indicated; note that because of the inclusion of interacted fixed effects in column (4), the white and box coefficients are not meaningful. Controls are whether the applicant had a GED (versus regular high-school diploma) and whether he had an employment gap. Box variation indicates the source of variation in the box variable: "Cross-section" means the variation comes from a comparison of box and nonbox stores in the preperiod; "Temporal" means the variation is pre- and post-BTB, triggered by the implementation of the BTB policy. In the last column, which is shown as a comparison point, there is no box variation; the pattern over the same time period is shown for companies that did not change their job applications. *10%, **5%, and ***1% significance level.

To replicate this result, we will make a small adjustment to the exact regression specification we ask you to run, but it is capturing the most important features of the main analysis presented in the paper. First, you will need to create a new variable which is the interaction term of post and

Black. You simply need to create a variable that multiplies `post` by `black`. Then run a linear regression model of the following form:

$$\text{response}_i = \alpha + \beta_1 \text{Black}_i + \beta_2 \text{post}_i + \beta_3 \text{post}_i \times \text{Black}_i + \varepsilon_i$$

Report the estimated coefficients from the regression. Note that you should obtain results that are close, but not identical to those in Table 4, column 2 above (the main difference is with the estimated standard errors, but you can ignore those differences). You should arrive at a nearly identical value to the one reported in the first row of estimated coefficients reported in the second column. [4pt]

23. This study used data from an experiment that sent out fictitious resumes to employers in New Jersey and New York City. Do you think we can easily generalize the results from this study when discussing how a similar policy would perform in San Francisco or Washington DC? Discuss briefly in 1-2 sentences. [2pt]

24. **This question will only be graded on completion.** A similar study was conducted more recently. You can read a brief summary of that study here. Academics and non-academics responded very differently to this study, less so to the results of the study, but more to its method of sending out fictitious resumes, and the practice of revealing the callback rates by company. You can see some of the discussion here, and here. Briefly discuss whether this type of study presents an ethical issue with respect to its method. Do you think we gain enough important knowledge from this kind of research such that these studies should still be conducted despite these concerns? [1pt]

Simulation Exercise

Sample Size and Rejecting the Null Hypothesis

We are going to run the following code to see how sample size and true differences in the data generating process lead to our ability to detect or fail to detect those differences when using hypothesis tests. The following summarizes what the simulation is going to do:

(i) Construct a population with some baseline outcome
(ii) Define the true treatment effect of some intervention
(iii) Define the sample size we draw from the population
(iv) Draw a random sample from the population
(v) Randomly assign units into treatment and control
(vi) Add the real treatment effect to the baseline outcome
(vii) Run a t-test for the post-treatment outcome between treatment and control
(viii) Repeat many times, and summarize the distribution of the p-values

Each time we run the following code, we will have a **real** difference between the treatment and control groups. The question is whether we will always manage to detect the difference, and reject the null hypothesis of no difference between the population means even when the size of the effect is small, and/or when the sample size is small.

25. Run the code provided in `simulation_treatment_effects.R`. We will first use the baseline values of the true effect and the sample size, as provided in the code (0.5 and 50, respectively). Based on the distribution of p-values (use the `hist()` function at the end of the code), if you were to collect just one sample, should you feel confident that you will reject the null hypothesis of no difference in the population means at a significance level of 5%? (Including the graph of the distribution of p-values is optional.) [1pt]

26. Increase the sample size to 500, and then to 5,000, and re-run the simulation for each sample size. How does your answer to the previous question change for each new sample size value? [1pt]

27. Reset the sample size value to 50. Increase the size of the true effect from 0.5 to 1 and then to 5, and re-run the simulation for each value. How does your answer to the original question change? [1pt]

Take a moment to think about and reflect on how many of the things you completed during this assignment that would have been entirely incomprehensible to you merely 10 weeks ago. We have covered *a lot* of material, and you have learned *a lot*. Congratulations!