Pontificia Universidad Católica de Chile
Facultad de Economía & Administración
Instituto de Economía

Problem Set # 3

Jose Carlo Bermúdez

jcbermudez@uc.cl

Econometric Theory II

Professor: Tomás Rau

June 9th, 2024

Summary

This report includes my solution to the problem set # 3 for the Econometric Theory II graduate course. The first section of the report offers a theoretical framework for matching analysis. The second part of the document implements causal analysis using Propensity Score Matching (PSM) and Differences-in-Differences (DiD) models. The replication package is available at my GitHub account.

Contents

1 Theoretical Problem

Suppose you want to estimate the ATT from a non-random intervention. For this purpose, you have a sample of n_1 treated and n_0 controls. Let I_0 be the index for control units and I_1 the index for treatment units. Thus, you observe $y_i = y_i^1 d_i + (1 - d_i)y_i^0$, where d_i is a treatment dummy. Let us suppose that you are trying to implement a Matching strategy. For every treatment unit i, you would like to find a counterfactual. A general approach is approximating the individual counterfactual as $\hat{y}_i^0 = \sum_{j \in I_0} \lambda(i, j) y_j$, where $\lambda(i, j)$ is a weight.

1. Show that the ATT can be written as follows:

$$\tau_{att} = \frac{1}{n_1} \sum_{i \in I_1}^{n_1} y_i - \frac{1}{n_1} \sum_{j \in I_0}^{n_0} \bar{\lambda}(j) y_j \tag{1}$$

where $\bar{\lambda}(j) = \sum_{i \in I_1}^{n_1} \lambda(i, j)$ represents the sum of all weights using the control j.

Proof: The expression is obtained from the definition of ATT.

$$\tau_{att} = \mathbb{E}[y_i^1 - y_i^0 | d_i = 1]$$

$$\tau_{att} = \mathbb{E}[y_i^1 | d_i = 1] - \mathbb{E}[y_i^0 | d_i = 1]$$

By analogy principle, we can find expressions for $\mathbb{E}(.)$, but for the second term (y_i^0) we do not directly observe the counterfactual, thus, we employ its approximation ($\hat{y}_i^0$):

$$\tau_{att} = \frac{1}{n_1} \sum_{i \in I_1}^{n_1} y_i - \frac{1}{n_1} \sum_{i \in I_1}^{n_1} \hat{y}_i^0$$

$$\tau_{att} = \frac{1}{n_1} \sum_{i \in I_1}^{n_1} y_i - \frac{1}{n_1} \sum_{i \in I_1}^{n_1} \left(\sum_{j \in I_0}^{n_0} \lambda(i, j) y_j \right)$$

$$\tau_{att} = \frac{1}{n_1} \sum_{i \in I_1}^{n_1} y_i - \frac{1}{n_1} \sum_{j \in I_0}^{n_0} \left(\sum_{i \in I_1}^{n_1} \lambda(i, j) y_j \right)$$

$$\therefore \tau_{att} = \frac{1}{n_1} \sum_{i \in I_1}^{n_1} y_i - \frac{1}{n_1} \sum_{j \in I_0}^{n_0} \bar{\lambda}(j) y_j$$

■

2. There are many different types of weights. Refer to the structure of these weights $\lambda(i, j)$ for the case of Abadie and Imbens (2006), the case of Inverse Probability Weighting (some of its versions), and the case of Díaz, Rau, and Rivera (2015).

- Abadie & Imbens (2006) propose an estimator based on matching imputation, where the matching process is with replacement, allowing each unit to be used as a match more than once. The weights they provide can be used to compute a weighted difference in means that is equal to the matching imputation estimator. There are two cases depending on the treatment effect to be estimated:
  - When it comes to Average Treatment on Treated (ATT) each control unit i receives a weight equal to $\lambda_M(i)$, where $\lambda_M(i) = \sum_{l=1}^{N} \mathbb{1}\{i \in J_M(l)\} \frac{1}{\#J_M(l)}$, where $J_M(l)$ is the set of units matched (M) to unit l, and $\#J_M(l)$ is the size of that set.
  - For the Average Treatment Effect (ATE) each unit i receives a weight equal to $1 + \lambda_M(i)$ with $\lambda_M(i)$ as defined above.

- Díaz et al. (2015) implement the same weighting process as in Abadie & Imbens (2006) but with an optimal selection of the $\#J_M(l)$, which they call Bilevel Optimization Problem (BLOP). They solve the following unconstrained optimization problem:

$$\min_{(\lambda_1,\ldots,\lambda_{J_m(l)})\in\Lambda_m} ||X_i - \sum_{j=1}^{J_m(l)} \lambda_j X_j||$$

  when there are more than one solution, the method selects the solution which minimizes

$$\sum_{j=1}^{J_m(l)} \lambda_j ||X_i - X_j||$$

- Finally, for the inverse probability weighting it first estimates a discrete binary choice model (probit or logit) using information from both sets of matched ($m \in M$) and unmatched ($u \in U$) units using information from covariates. Thus, the weight λ is a fitted probability from the model estimated for both treatment ($d_i = 1$) and control ($d_i = 0$) arms. Thus:

$$\lambda_M(i) = \hat{\mathbb{P}}(d_i = 1)$$

3. A friend tells you that it is better to use $\lambda(i,j) = \frac{1}{n_1}$. Write the population analog of the ATT that results from replacing the new $\bar{\lambda}(j)$ in Equation 1 and discuss the possibility of removing the selection bias in this case.

This is quite similar to the exercise we saw in ayudantía # 6. Let's replace the new weight in the expression above.

$$\tau_{att} = \frac{1}{n_1}\sum_{i\in I_1}^{n_1} y_i - \frac{1}{n_1}\sum_{j\in I_0}^{n_0}\left(\sum_{i\in I_1}^{n_1}\frac{1}{n_1}\cdot y_j\right)$$

$$\tau_{att} = \frac{1}{n_1}\sum_{i\in I_1}^{n_1} y_i - \frac{1}{n_1^2}\sum_{j\in I_0}^{n_0}\left(\sum_{i\in I_1}^{n_1} y_j\right)$$

$$\tau_{att} = \frac{1}{n_1}\sum_{i\in I_1}^{n_1} y_i - \frac{1}{n_1^{\cancel{2}}}\sum_{j\in I_0}^{n_0}\left(\cancel{n_1}\cdot y_j\right)$$

$$\tau_{att} = \frac{1}{n_1}\sum_{i\in I_1}^{n_1} y_i - \frac{1}{n_1}\sum_{j\in I_0}^{n_0} y_j$$

Notice that the second expression is the sum of the counterfactual units for $j \in I_0$, thus we can sum for the n_0 control units:

$$\tau_{att} = \frac{1}{n_1}\sum_{i\in I_1}^{n_1} y_i - \frac{1}{n_0}\sum_{j\in I_0}^{n_0} y_j$$

$$\therefore \tau_{att} = \bar{y}_1 - \bar{y}_0$$

The expression above is a "naive" matching estimator. Our friend's suggestion is not that good because representing $\lambda(i,j) = \frac{1}{n_1}$ yields a biased estimator of the ATT. This bias results because in observational studies, treatment, and control groups often differ in observable and unobservable characteristics that may influence outcomes. The naive estimator for the ATT calculates the difference in means of outcomes between treated and control groups without adjusting for these differences. In that case, the resulting estimator will be biased due to confounding variables (factors that affect both the probability of receiving treatment and outcomes) or selection bias (units receiving treatment may not be comparable to those not receiving it). Then, we would rather rely on any of the three weighting methods discussed in question # 2.

4. Let us suppose that you decide to implement Kernel Matching a la Heckman et al. (1998) so we define $\lambda(i,j) = \kappa(|e(X_i) - e(X_j)|/h)/\sum_{j\in I_0}\kappa(|e(X_i) - e(X_j)|/h)$. $\kappa(.)$ is a Kernel function, $e(X_m)$ is the propensity score evaluated in X_m, and h is the bandwidth. Suppose that you are interested in balancing the variables "post-matching" for every unit in the treatment group i, the weighted average of the X_j in the treatment group will be the closest as possible. ¿How would you choose the bandwidth? Hint: You can propose an objective function to be minimized that improves the balance following Díaz et al. (2015).

Following Díaz et al. (2015), an appropriate objective function to measure the balance between groups post-matching is the weighted average of the absolute differences in covariates between the treatment and control groups. Specifically, we can consider the sum of the standardized mean differences (SMD) between the covariates of the treated and untreated groups. The SMD for a covariate X_k is defined as:

$$\mathrm{SMD}_k = \frac{|\bar{X}_{k,treated} - \bar{X}_{k,control}|}{\sqrt{\frac{1}{2}\left[Var(X_{k,treated}) - Var(X_{k,control})\right]}} \tag{2}$$

Where $\bar{X}_{k,\text{ treated}}$ is the weighted mean of the covariate X_k in the treated group and $\bar{X}_{k,\text{ control}}$ is the weighted mean of the covariate X_k in the control group. The weighted average of X_k in the control group is:

$$\bar{X}_{k,\text{ control}} = \sum_{j\in I_0}\lambda(i,j)X_{k,j}$$

Where:

$$\lambda(i,j) = \frac{\kappa\left(\left|e\left(X_i\right) - e\left(X_j\right)\right|/h\right)}{\sum_{j\in I_0}\kappa\left(\left|e\left(X_i\right) - e\left(X_j\right)\right|/h\right)}$$

To choose the bandwidth h that minimizes the imbalance, we can follow these steps:

(a) Define the Objective Function: The objective function ($g(h)$) we want to minimize is the sum of the SMDs of all covariates:

$$\mathrm{g}(h) = \sum_k \frac{|\bar{X}_{k,treated} - \bar{X}_{k,control}|}{\sqrt{\frac{1}{2}\left[Var(X_{k,treated}) - Var(X_{k,control})\right]}} \equiv \sum_k \mathrm{SMD}_k$$

(b) Implement an optimization algorithm:

- We start by defining $[\underline{h}, \bar{h}]$, an arbitrary grid for h.
- We arbitrarily define a functional form for κ (i.e. a Gaussian kernel).
- We set a tolerance level ξ.
- While convergence is not achieved, we do: For each value of h in the grid, perform matching, and calculate the weighted means of the covariates X_k in the control group, compute Equation 2 for each covariate X_k, and compute $\hat{g}(h)$.
- Repeat until $||\hat{g}(h+1) - \hat{g}(h)|| \leq \xi$.
- Thus, our optimum bandwidth is defined as $h^* \equiv \min_{h\in[\underline{h},\bar{h}]} ||\hat{g}(h+1) - \hat{g}(h)||$.

5. You are interested in obtaining the standard error of your estimator. Use the variance identity formula to propose an estimator for the variance of ATT. You can assume that y_i is i.i.d.

For simplicity, let us define:

$$\tau_{att} = T(y_i) - C(y_i, \lambda)$$

Where:

$$T(y_i) \equiv \frac{1}{n_1} \sum_{i \in I_1}^{n_1} y_i$$

$$C(y_i, \lambda) \equiv \frac{1}{n_1} \sum_{j \in I_0}^{n_0} \bar{\lambda}(j) y_j$$

The variance for the average treatment on the treated is defined as follows:

$$V(\tau_{att}) = V[T(y_i) - C(y_i, \lambda)]$$

$$V(\tau_{att}) = V[T(y_i)] + V[C(y_i, \lambda)] \tag{3}$$

Now, we need to find analytical expressions using the formula for the variance identity.

- First, for $V(T)$:

$$V[T(y_i)] = \mathbb{E}[V(T|\lambda)] + V[\mathbb{E}(T|\lambda)]$$

  But we know that the expected value for the treated does not depend on weights (λ), thus we have unconditional moments:

$$\begin{aligned}
V[T(y_i)] &= \mathbb{E}\{V[T(y_i)]\} + V\{\mathbb{E}[T(y_i)]\} \\
&= \mathbb{E}\left\{V\left(\tfrac{1}{n_1}\textstyle\sum_{i\in I_1}^{n_1} y_i\right)\right\} + V\left\{\mathbb{E}\left(\tfrac{1}{n_1}\textstyle\sum_{i\in I_1}^{n_1} y_i\right)\right\} \\
&= \mathbb{E}\left\{\left(\tfrac{1}{n_1}\right)^2 \textstyle\sum_{i\in I_1}^{n_1} V(y_i)\right\} + V\left\{\tfrac{1}{n_1}\mathbb{E}\left(\textstyle\sum_{i\in I_1}^{n_1} y_i\right)\right\} \\
&= \mathbb{E}\left\{\tfrac{1}{n_1^2}(n_1\sigma^2)\right\} + V\left\{\tfrac{1}{n_1}(n_1\mu)\right\} \\
&= \mathbb{E}\left\{\tfrac{\sigma^2}{n_1}\right\} + \cancelto{0}{V\{\mu\}} \\
\therefore V[T(y_i)] &= \tfrac{\sigma^2}{n_1}
\end{aligned}$$

- Secondly, for $V(C)$:

$$\begin{aligned}
V[C(y_i,\lambda)] &= \mathbb{E}[V(C|\lambda)] + V[\mathbb{E}(C|\lambda)] \\
&= \mathbb{E}\left\{V\left(\tfrac{1}{n_1}\textstyle\sum_{j\in I_0}^{n_0} \bar{\lambda}(j) y_j|\lambda\right)\right\} + V\left\{\mathbb{E}\left(\tfrac{1}{n_1}\textstyle\sum_{j\in I_0}^{n_0} \bar{\lambda}(j) y_j|\lambda\right)\right\} \\
&= \mathbb{E}\left\{\left(\tfrac{1}{n_1}\right)^2 \textstyle\sum_{j\in I_0}^{n_0} \bar{\lambda}(j)^2 V(y_j|\lambda)\right\} + V\left\{\tfrac{1}{n_1}\textstyle\sum_{j\in I_0}^{n_0} \bar{\lambda}(j)\mathbb{E}(y_i|\lambda)\right\} \\
&= \mathbb{E}\left\{\tfrac{\sigma^2}{n_1^2}\textstyle\sum_{j\in I_0}^{n_0} \bar{\lambda}(j)^2\right\} + \underbrace{\cancelto{0}{V\left\{\frac{\mu}{n_1}\sum_{j\in I_0}^{n_0} \bar{\lambda}(j)\right\}}}_{\bar{\lambda}\text{ are fixed}} \\
\therefore V[C(y_i,\lambda)] &= \tfrac{\sigma^2}{n_1^2}\textstyle\sum_{j\in I_0}^{n_0} \bar{\lambda}(j)^2
\end{aligned}$$

Replacing expressions for $V[T(y_i)]$ and $V[C(y_i, \lambda)]$ in Equation 3, we arrive at:

$$V(\tau_{att}) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_1^2} \sum_{j \in I_0}^{n_0} \bar{\lambda}(j)^2 \tag{4}$$

Finally, the standard error for our estimator of the ATT is given by:

$$SD(\tau_{att}) = \sigma \cdot \sqrt{\frac{1}{n_1}\left(1 + \frac{1}{n_1} \sum_{j \in I_0}^{n_0} \bar{\lambda}(j)^2\right)} \tag{5}$$

Where $\bar{\lambda}(j) = \sum_{i \in I_1}^{n_1} \lambda(i, j)$

6. A cousin of your friend suggests using Bootstrap to calculate the standard error. Discuss the feasibility of finding an appropriate estimator for this standard error. Does your response depend on the choice of $\lambda(i, j)$?

   It all depends. For instance, Abadie & Imbens (2008) show that the standard bootstrap is, in general, not valid for matching estimators, even in the simple case with a single continuous covariate where the estimator is root-N consistent and asymptotically normally distributed with zero asymptotic bias. So in the context of our naive estimator for the τ_{att} standard bootstrapping methods would not work. However, the choice of $\lambda(i, j)$ definitely has implications when trying to apply bootstrapping techniques. Abadie & Imbens (2008) also argue that applying standard bootstrapping techniques on the weighting method proposed by Heckman et al. (1998) –discussed in question d) above– can yield valid inferences because these estimators are asymptotically linear.

2 Matching

In this section, I discuss and replicate some of the results reported in the paper *"Human Capital and Industrialization: Evidence from the Age of Enlightenment"* (Squicciarini & Voigtländer, 2015).

2.1 Summary of the Paper

Motivation of the paper. Nowadays, economic literature portrays education as one of the main predictors of contemporaneous development. However, when it comes to periods before the Century XIX education often goes unnoticed as a key driver in economic growth. Some of the reasons are that education was usually intended as a mechanism to pursue sectarian/partizan goals, or because countries with high levels of education were often facing weaker economic performance than non-literate countries. Using the mainstream economic wisdom that is based on literacy as a measure of the average worker's skill would yield misleading conclusions on how relevant education was in early industrialization. Nonetheless, the Industrial Revolution was not directly born as a collective initiative but as the result of a small specific group of innovative people instead, so it was potentially driven by the top-end of the skill distribution.

Main research question and framing. The paper tackles a thriving question: Does the distinction between upper-tail and average skills reinstate the importance of human capital during the first Industrial Revolution? This question is quite relevant because while average worker skills raise productivity for a given technology, upper-tail knowledge (entrepreneurs) fosters the adoption of new manufacturing technology, which is particularly useful to understanding why technological progress is rapid. Predictions derived from this wisdom are twofold. On one hand, as technological growth becomes more rapid, a larger local knowledge elite is associated with higher income, manufacturing employment, and growth. On the other hand, the paper's intuition would predict that worker skill levels raise wages in both sectors (and thus income in the cross-section), but not growth.

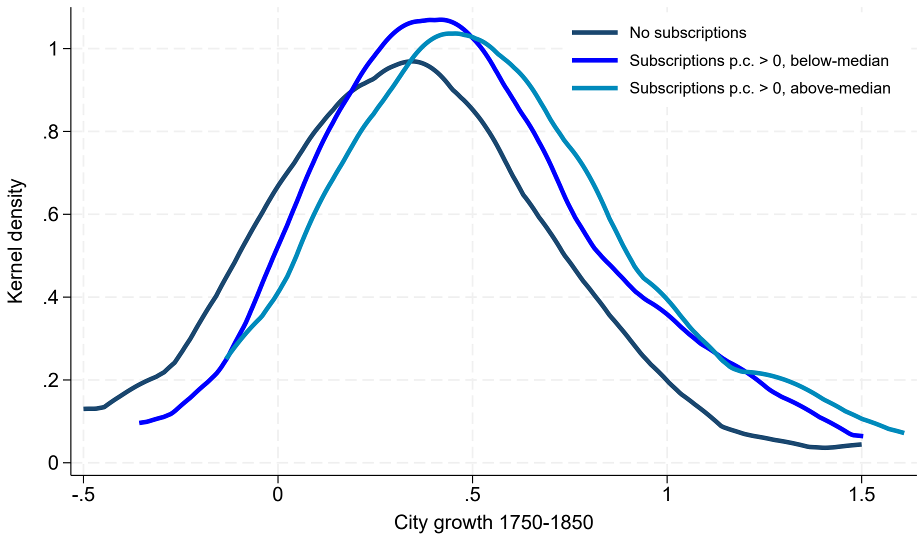
Results. The paper offers several compelling findings. The authors use the density of subscribers to the Great Encyclopédie as a proxy for local scientific elites and find that this density is uncorrelated to literacy. They also found that cities with higher subscribers' density grew faster after 1750, but there was a small or no effect before 1750. Using data at the department level, they also find that departments with higher subscriber density had higher income and industrialization in mid of the mid-19th century, but not before 1750. In addition, using firm census data they found that wage effects were strongest in innovative modern sectors. Finally, they conclude that literacy is positively associated with income in the cross-section before and after 1750, but not with growth.

Main conclusion. The paper points out an important message for economic development: upper-tail knowledge is crucial, likely via affecting productivity/adoption of modern technology.

2.2 City Growth Density

Regional income data are not available for early modern Europe, hence, the city population is a widely used proxy for economic development. In Figure 1 I report the distribution for the growth of the population in French cities between 1850 and 1750. The figure suggests that cities with subscriptions to the Encyclopèdie grew more rapidly compared to those cities without any subscriptions. This is suggestive evidence that economic dynamism was higher during the Industrial Revolution in the context where upper-tail knowledge was present.

Figure 1: Subscriptions and City Growth, 1750–1850



Note: The figure shows the Kernel density of city population growth over the period 1750–1850 for three subsets of French cities: 108 cities without Encyclopèdie subscriptions, as well as 43 (42) cities with subscriptions and below-median (above-median) subscriber density.

2.3 Empirical Strategy

The authors exploit their panel (at the city and department level, respectively) using OLS and matching estimations. For the latter, they use the nearest neighbor method using 3 neighbors for the definition of the control arm. A treated group is defined as those cities with at least one subscriber but matching on population because larger cities are more likely to have at least one subscriber. Hence, the authors compare cities with and without subscriptions of similar size. The main assumptions behind their empirical strategy are unconfoundness and overlap.

The first assumption implies that, to capture a causal effect of the presence of scientific elites on economic growth, city characteristics after the industrialization period must be independent of the adoption of subscriptions at the regional level. This assumption is plausible and feasible because the authors disentangle between the fact that knowledge elites foster the adoption of new technology and therefore growth—especially when the technological frontier expands quickly (as in the context of industrialization)–from the average human capital that may affect the productivity of any given mode of production (and thus income levels), but not the adoption of new technology (and thus not growth).

The second assumption implies that for given any observable, French cities could be potentially observed with and without at least one subscriber to Encyclopèdie. I think that the authors are able to control for several confounders such as human capital, geographical advantages, or cultural characteristics, thus, these are observables that are not left into residuals of the models, making plausible the assumption that for any covariate a French city could be potentially observed the treatment with at least one subscription to Encyclopèdie. Maybe, out of the assumptions, this is the most difficult to meet, because there are confounders that mechanically trigger the presence of subscribers, such as the size of the population, but In my belief, this assumption also holds because they normalize all the outcomes of interest using population size, and also match treated and control cities using population.

2.4 Sample Balance

Table 1 replicates the Table I reported in the paper of Squicciarini & Voigtländer (2015). Results come from independent regressions on several covariates using the log of the density of subscribers (normalized by population) as the dependent variable. It is always important to test for sample balancedness to check whether treated and control arms are the same in a wide set of observables except in the treatment. This is why the authors test if other town characteristics vary systematically with Encyclopèdie subscriptions. The authors argue that these results confirm the intuition that few city characteristics that vary systematically with subscriber density are those that one should expect if subscriptions reflect the size of the local knowledge elite. I think this conclusion is true, but not to a full extent because there are potential unobserved factors that are associated with both city size and city density as shown by the sign of the coefficient for population which turns out to be negative for the subsample of cities with at least one subscriber.

Table 1: Correlations With Subscribers Density

	(1) All (no controls)	(2) Subs>0
Panel A: Baseline controls		
ln(population 1750)	0.374***	-0.234**
	(0.073)	(0.110)
Atlantic Port	0.081	-0.222
	(0.207)	(0.213)
Mediterranean Port	0.022	-0.129
	(0.276)	(0.223)
Navigable River	0.422**	-0.167
	(0.202)	(0.210)
Non French speaking	-0.376**	-0.719**
	(0.146)	(0.297)
Panel B: Early knowledge controls		
University	1.030***	0.316
	(0.186)	(0.194)
Printing press	0.712***	0.203
	(0.170)	(0.182)
ln(books printed 1500)	0.171***	0.039
	(0.061)	(0.066)

	(3) All (no controls)	(4) All (controls)	(5) Subs>0
Panel C: Worker skills			
Literacy 1686†	0.551	0.905	0.026
	(0.593)	(0.650)	(0.777)
Literacy 1786†	0.290	0.358	0.054
	(0.345)	(0.324)	(0.383)
School rate 1837†	0.363	0.313	0.200
	(0.350)	(0.335)	(0.412)
Panel D: Additional controls			
ln(STN books density)	0.242***	0.194***	0.062
	(0.041)	(0.050)	(0.056)
Pays d'election	0.046	0.137	0.048
	(0.129)	(0.128)	(0.205)
ln(pre industrial density)†	-0.027	-0.033	-0.546
	(0.830)	(0.829)	(0.846)
ln(distance coal)	-0.173**	-0.104	0.025
	(0.083)	(0.083)	(0.106)
ln(nobles density)†	0.233	0.253	0.736***
	(0.278)	(0.301)	(0.249)

Note: This table replicates Table 1 of the Squicciarini & Voigtländer (2015) paper. The dependent variable is subscriber density. Variables with the symbol † are taken from a dataset at the department level. Regressions in the column titled "All (no controls)" do not include any control, the column titled "All (control)" includes baseline and early knowledge controls, and those columns titled "Subs>0" only include observations with positive subscriptions.

2.5 Matching Results and Interpretation

Table 2 replicates matching results as in the Table II of the Squicciarini & Voigtländer (2015) paper. The coefficient displayed in the table corresponds to the ATT, using the method of nearest neighbor for three neighbors. Columns (1)-(4) show results for a variety of specifications for the period of industrialization (1750–1850); column (1), includes the full sample and matches by initial population; column (2) excludes the 10 percent smallest and largest cities in 1750; columns (3) and (4), introduces geographic latitude and longitude as additional matching variables. Columns (5) and (6) are Placebo tests for the pre-industrialization period (1700–1750). For our period of interest, we get statistically significant coefficients. The economic interpretation for the first four columns is as follows:

- **Column (1).** Conditional on population size (our matching variable), French cities with at least one subscription to Encyclopèdie faced an increase in population growth by ≈ 0.15 log points with respect to cities without any subscriber.
- **Column (2).** Conditional on population size (our matching variable) and only including the bottom and top 10% of the growth distribution, French cities with at least one subscription to Encyclopèdie faced an increase in population growth by ≈ 0.16 log points with respect to cities without any subscriber.
- **Column (3).** Conditional on population size, latitude, and longitude (our matching variables), French cities with at least one subscription to Encyclopèdie faced an increase in population growth by ≈ 0.27 log points with respect to cities without any subscriber.
- **Column (4).** Conditional on population size, latitude, and longitude (our matching variables) and only including the bottom and top 10% of the growth distribution, French cities with at least one subscription to Encyclopèdie faced an increase in population growth by ≈ 0.16 log points with respect to cities without any subscriber.

The conclusion is that French cities with subscriptions grew ≈ 0.15–0.27 log points faster (relative to an average city growth rate of 0.37 log points) than those of comparable size without subscriptions.

Table 2: Matching Estimation by City Size and Location

	Period 1750–1850				1700–1750	
	(1) All	(2) 10–90 pct	(3) All	(4) 10–90 pct	(5) All	(6) 10–90 pct
$\mathbb{1}[Subs > 0]$	0.146**	0.155**	0.267***	0.163**	0.087	0.063
	(0.07)	(0.07)	(0.07)	(0.07)	(0.06)	(0.05)
Population	✓	✓	✓	✓	✓	✓
Location	✗	✗	✓	✓	✓	✓
Observations	177	154	167	144	129	110

Note: This table replicates Table 2 of the Squicciarini & Voigtländer (2015) paper. The dependent variable: log city growth over the indicated period.

2.6 Matching & OLS

Table 3 replicates matching results as in the Table VIII of the Squicciarini & Voigtländer (2015) paper. This exercise is a robustness check and is important because tests for alternative proxies of upper-tail knowledge. Other reasons are:

- Comparing the two approaches can reveal if differences in results are due to model specification or each method's ability to balance covariates.
- If results differ, it may indicate that the linearity assumption in the OLS model is not appropriate for the data.
- Comparing the results can help identify if either model is particularly sensitive to model specification.
- If results are similar between the two methods, it can increase confidence that selection bias is not strongly influencing the results. If they differ, it may suggest the presence of omitted variables affecting OLS results.
- Comparing both methods can provide cross-validation of results. If both methods reach similar conclusions, there is greater confidence in the validity of the findings. If they differ significantly, it may indicate the need for further exploration and possibly additional methods to address the discrepancies.

The data on pre-1750 scientific societies allows the authors to address the possibility of reverse causality: since the Quarto edition was printed in 1777–1779, initial industrial growth between 1750 and 1780 may have raised the demand for the Encyclopèdie. Both propensity score matching (Panel A) and OLS estimation using member density (Panel B) confirm the main results of the paper: cities with pre-1750 scientific societies grew significantly faster during French industrialization.

Table 3: Scientific Societies, Descriptions Des Arts et Metiers, and City Growth

	Pre-1750 scientific societies			Desc. Arts et Metiers	
	(1)	(2)	(3)	(4)	(5)
	1750–1850		1700–1750	1750–1850	
Panel A: Matching estimation					
$\mathbb{1}[x > 0]$	0.204**	0.193**	0.028	0.193**	0.140
	(0.098)	(0.095)	(0.083)	(0.082)	(0.096)
Population	✓	✓	✓	✓	✓
Location	✗	✓	✓	✗	✓
Observations	185	175	136	177	167
Panel B: OLS					
ln[density(x)]	0.285***	0.295**	0.041	0.533***	0.460**
	(0.083)	(0.126)	(0.098)	(0.167)	(0.208)
$\mathbb{1}[x > 0]$		-0.010			0.048
		(0.130)			(0.086)
Controls	✓	✓	✓	✓	✓
R^2	0.34	0.34	0.22	0.32	0.32
Observations	158	158	118	166	166

Note: This table replicates Table 8 of the Squicciarini & Voigtländer (2015) paper. The dependent variable is the log city population growth over the period indicated in the table header. OLS destinations, reported in Panel B are weighted by the initial population of the respective period.

2.7 Critiques to the Paper

This is a great paper and a significant contribution to the literature on economic development. However, in my opinion, there are two concerns:

1. Regarding the internal validity of the results.
    - I think that the authors are not very clear about the criteria for selecting the nearest neighbor estimator, or the number of neighbors. This is relevant in the sense that they are working with data at the city level, thus, the number of neighbors and the weighting method for control cities is quite relevant because it could yield imperfect matching and thus, residual bias. They do not address this concern at all.
    - They do not push the estimations beyond propensity score matching. They could have exploited another method to get sharper causal effects claiming a smaller number of assumptions (as required when using matching methods). For instance, I think they could have implemented instrumental variables to address omitted variable threats. For instance, the Encyclopèdie was printed on a very specific type of paper known as "laid" paper, was imported from countries like Great Britain, and was typically made from linen rags. Thus any variation related to prices or shocks that affected the laid paper may give a nice instrument of a variable that directly affects the production of any additional printed version and thus the subscription to Encyclopèdie that is uncorrelated with French city's growth. This is only an example of how the internal validity (in terms of causal interpretation) could have been improved in the paper.
2. Regarding the external validity of the results.
    - The results are not necessarily extendable to other contexts due to the geographic and temporal scope of the study. The authors focus on France and specific cities. Industrialization in different contexts may have been driven by a wider array of factors, and a broader comparative analysis could provide a more comprehensive understanding of the role of human capital in industrialization.

3 Differences-in-Differences

3.1 Summary of the Paper

This paper uses a randomized field experiment to study changes in the sexual behavior of Kenyan teenagers in response to information on HIV risk. The author measures the responsiveness of teenagers to HIV information and compares their responses along both the risk avoidance and the risk reduction margins. In terms of the empirical strategy, the author is able to estimate the ATE and ATT. For the first one, she implemented a standard difference (SD) in means model, and because the program was randomly assigned she interpreted the difference as the ATE. Secondly, the author exploits control cohorts for students enrolled in grade 8 in 2004 to estimate a difference-in-differences (DD) model.

The author compares estimations from the SD and DD models. The DD model is applied in the context where the cohort data is available. Comparing the single-difference to the difference-in-differences estimates is useful for two reasons. First, if the randomization of the RR program assignment was not perfect, the difference-in-differences will adjust for potential pre-existing random differences in means between RR treatment and RR comparison schools. Second, the difference-in-differences allows the inclusion of school fixed effects, which allows for control for unobservable school characteristics that enter the equation in an additive way.

The author relies on several identification assumptions. She argues that since the programs were randomly assigned the expected value of unobservables conditional on the RR or the TT program is equal to zero (independence). She also relies on the assumption of parallel trends and assumes zero spillovers across treated and controlled teenagers. In terms of the plausibility of the assumptions, the parallel trends are quite reasonable since both, treated and control teenagers are quite similar at baseline. The one that is not plausible is the absence of spillovers, and the author does not address this issue.

3.2 Sample Balance

In this question, we are asked to discuss Tables 1 and 2 of the paper. Table 1 displays summary statistics on knowledge and behavior among adolescents in the study area, while Table 2 reports balance test between groups in terms of school characteristics and outcomes for pre-program cohort.

In terms of Table 1, a thing that comes to my attention is the fact that only less than half of girls are aware that condoms can prevent HIV infection or prevent pregnancy. Even more, only 29 percent of girls and 25 percent of boys knew that older men were more likely to be HIV positive than teenage boys. These numbers might indicate a lack of information among the group of teenagers under study.

On the other hand, when it comes to Table 2 some facts arise. Except for class size, which is lower on average in RR treatment schools, all other differences in pre-treatment school characteristics are small and we can not reject the null hypothesis of null differences between treatment and control groups. on the other hand, the sample is less balanced when it comes to the long-run schooling status of teenagers. Schools in the RR treatment group are significantly less likely to allow their students to repeat grade 8 compared to schools in the RR control group. What's more, schools in the TT treatment group are significantly less likely to see their students go on to secondary school than schools in the TT comparison group. Hence, we cannot argue that both groups are exhaustively equal in terms of baseline covariates. These pre-existing differences will bias us against finding an effect since the opportunity cost of getting pregnant is lower for out-of-school girls than it is for in-school girls, who will not be allowed to stay in school while pregnant.

3.3 Replicating Baseline Results

In this question, we are asked to replicate the table 2 of the paper and explain the role this exercise has in the paper. The author estimates different specifications to get the causal effect of both programs: treatment 1 Teacher Training (TT) on the national HIV/AIDS curriculum for primary school, and treatment 2, the Relative Risk Information Campaign (RR). Thus the table plays the role of reporting Average Treatment Effects (ATE) and Average Treatment Effects on the Treated (ATT) of both programs using an OLS difference-in-means model and a Differences-in-Differences approach, respectively. Results are displayed in Table 4 and mirror those reported by Dupas (2011), except for the margins derived from the Probit model in column (2). This is because the author computes the marginal effects using an obsolete command called "margfx", but I implement the in-built command "margins". Although results are qualitatively the same. Interpretations are as follows:

- **Has started childbearing.** Columns from (1) to (4) display estimates for the treatment effects of the RR and TT programs. Column (1) is a standard difference in the means model estimated using OLS. Since the program was randomly assigned, the coefficient for "RR information" corresponds to the ATE of the program. Thus, the RR information reduced the incidence of childbearing by 1.5 percentage points among treated girls relative to girls in the comparison group. The effect is also estimated using a Diff-in-Diff approach, reported in columns (3) and (4). Now, our parameter of interest is the interaction between RR information and the 2004 cohort. The point estimate is quite large compared to the one obtained with our OLS model, but also standard errors are quite larger. In contrast, the TT program had no impact on the incidence of childbearing.
- **Has started childbearing, according to marital status.** Columns (5) to (8) display the treatment effects of both programs on childbearing but are split according to marital status. We get that the RR information program decreased childbearing in the treated RR group by ≈ 1 and 3 percentage points with respect to the non-treated arm. Notice that, again the TT information program did not had any effect on childbearing.

Table 4: Probability That Girls Have Started Childbearing

	Has started childbearing				Has started childbearing, unmarried		Has started childbearing, married	
	(1) SD OLS	(2) SD PROBIT	(3) DD OLS	(4) DD-FE OLS	(5) SD OLS	(6) DD OLS	(7) SD OLS	(8) DD OLS
RR information	-0.015*	-0.014	0.006		-0.009**	0.015	-0.005	0.011
	(0.008)	(0.009)	(0.013)		(0.004)	(0.010)	(0.006)	(0.012)
RR information × 2004 cohort			-0.024	-0.020		-0.027**		-0.017
			(0.016)	(0.016)		(0.011)		(0.013)
TT on HIV/AIDS curriculum	0.006	0.007	0.008	0.003	0.006	0.006	-0.000	0.002
	(0.007)	(0.006)	(0.006)	(0.018)	(0.004)	(0.004)	(0.005)	(0.005)
Mean of dependent variable	0.054	0.054	0.054	0.054	0.021	0.021	0.033	0.033
Observations	5,988	5,988	10,968	10,968	5,988	10,968	5,988	10,968
Control cohort included	✗	✗	✓	✓	✗	✓	✗	✓
Individual characteristics	✓	✓	✓	✓	✓	✓	✓	✓
Primary school characteristics	✓	✓	✓	✗	✓	✓	✓	✓
Primary school fixed effects	✗	✗	✗	✓	✗	✗	✗	✗

Note: This table replicates Table 3 of the Dupas (2011) paper.

3.4 Childbearing & Partner's Age

In this question, we are asked to replicate table 4 of the paper and explain its role. Recall we are studying the impact of two different programs on teenage sexual behavior. We already know that the RR did decrease the incidence of childbearing, but now we are interested in understanding if there was any change in the composition of the partner's selection. This is super relevant because these results will inform the "intensive" margin, that is: whether the teenagers had low or high-risk intimacy, with high risk being defined as the relationships with older guys and lower risk with younger guys. Results are reported in Table 5[1] and mirror those reported by Dupas (2011), except for the marginal effects derived from the Probit model, again due to the differences in the command implemented to get the results. But the intuition remains immaculate. The analysis in this section will therefore not tell us anything about how the RR information affected girls' partner selection overall. However, it will inform us on how the RR information affected girls' selection of partners with whom to have unprotected relationships. The interpretation is as follows:

- **Age difference between a teenager and her partner.** Columns (1) and (2) display estimates for the age difference between the respondent and her baby's father. Results from the standard difference (SD) in means and Diff-in-Diff models suggest that the RR information program had a negative and statistically significant effect on the age difference for treated girls vis-a-vis non-treated group by $\approx$ 1.7-2.6 years.
- **Age gap>5.** In columns 3–5, the dependent variable is a dummy indicating whether the baby's father is more than 5 years older than the teenage girl. The coefficient for the ATE turned out to be negative and statistically significant. Thus, we can argue that the RR information program decreased by $\approx$ 22 percentage points off of a mean of 49% in the control group, which is equivalent to a significant decrease of 45%.
- **Age gap>10.** In columns 6–8, the dependent variable is a dummy indicating whether the baby's father is more than 10 years older than the teenage girl. The coefficient for the ATE turned out to be negative but we can not reject the null hypothesis of zero effects, even though we are still able to capture an ATT effect from the Diff-in-Diff model for a decrease in $\approx$ 23 percentage points, which is statistically significant at the 5%.

In all cases, we do not find any evidence on the effect of the TT program on the age gap. Moreover, the coefficients estimated from the Diff-in-Diff model are larger (in absolute terms) with respect to those we obtained using the simple difference model. The author argues that "This seems driven by the fact that the coefficients for being in an RR information school (but not in the RR cohort) are large and positive. This suggests that the RR program might have had negative spillovers onto nontreated students in the RR treatment schools. Indeed, the control cohort available is a younger cohort (the seventh graders of 2004). This cohort could have been indirectly and negatively affected by the RR information program if the "sugar daddies" newly turned down by informed eighth graders decided to try their luck with seventh graders instead" (p22).

[1]For this exercise we had to adjust the variable "age_father_27" which was available to us with the year the father was born for cases when the father was older than 56 years old. Thus we had to fix it by substracting 2005 (the year when the survey was implemented) minus the year the rather was born.

Table 5: Age Gap between Girls Who Have Started Childbearing and Their Partner

	Age difference between teenage girl and her partner		Age gap > 5 years			Age gap > 10 years		
	(1) SD OLS	(2) DD OLS	(3) SD OLS	(4) SD PROBIT	(5) DD OLS	(6) SD OLS	(7) SD PROBIT	(8) DD OLS
RR information	-1.685***	1.070	-0.224*	-0.230**	0.157	-0.064	-0.094	0.166**
	(0.609)	(0.817)	(0.116)	(0.103)	(0.121)	(0.061)	(0.070)	(0.084)
RR information × 2004 cohort		-2.576**			-0.351*			-0.229**
		(1.048)			(0.190)			(0.109)
TT on HIV/AIDS curriculum	-0.708	-0.331	0.074	0.102	0.026	-0.076	-0.065	-0.030
	(0.720)	(0.451)	(0.081)	(0.074)	(0.060)	(0.058)	(0.053)	(0.037)
Mean of dependent variable	5.91	5.91	0.49	0.49	0.49	0.16	0.16	0.16
Observations	120	250	134	134	278	134	134	278
Control cohort included	✗	✓	✗	✗	✓	✗	✗	✓
Individual characteristics	✓	✓	✓	✓	✓	✓	✓	✓
Primary school characteristics	✓	✓	✓	✓	✓	✓	✓	✓

Note: This table replicates Table 4 of the Dupas (2011) paper.

3.5 Event Study

In this section, I will dive into some of the nitty-gritty of the event-study methodology.

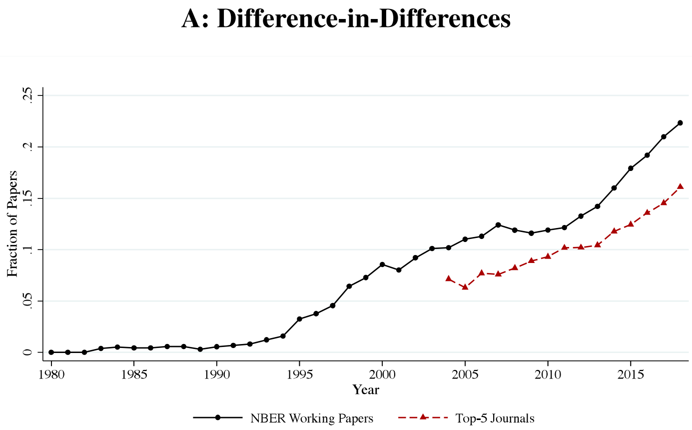
Why and where does the need to use models of this type arise? Surprisingly for me, the event study method was first developed in financial economics by Fama et al. (1969). This seminal work laid the groundwork for event studies by examining how stock prices adjust to new information, specifically analyzing the effects of stock splits. They developed a methodological framework for assessing the impact of specific events on stock prices. They established the concept of an event window and an estimation window to measure abnormal returns. Beyond financial economics, the event study has also been implemented in other branches such as labor economics (Jacobson et al., 1993). These types of models are usually applied in context when the researcher is interested in estimating dynamic treatment effects.

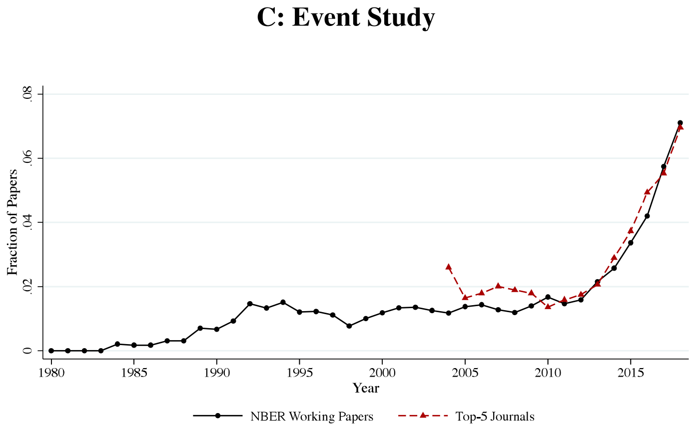
Differences-in-Differences vs. Event-Study. As shown in Figure 2, frontier economic research has increasingly adopted both methods as its main empirical strategy. Miller (2023) state that event-study models are "typically estimated in a reduced-form treatment effects context". Indeed, the event study specification is a generalization of a standard two-way fixed effects difference-in-difference specification.

Moreover, Miller (2023) also extends the connections between event-study and difference-in-differences models. He argues that "event study models fit within a family of related models that rely on a parallel trends assumption for identification of causal effects. All of these employ panel fixed effects (or a simplified version, such as dummies for post and treated unit) as key control variables". Table 6 summarizes some related approaches within this family. The first column labels the approach; the second column indicates the relevant estimating equation, and the third and fourth columns identify the relevant data structure.

Identification assumptions in an Event-Study. To estimate a causal effect using the event study methodology, several identification assumptions must be met. These assumptions are crucial to ensure

Figure 2: Use of DiD and Event Study in Econ Literature





Note: This figure was taken from Currie et al. (2020).

Table 6: **Comparisson between Difference-in-Difference and Event-Study**

	Model	Estimation Equation	Event Date Variation	Never-treated group(s)
1.	2 x 2 Difference-in-Difference	DiD	N/A	Yes
2.	2 x T Difference-in-Difference	ES, DiD	N/A	Yes
3.	N x T Difference-in-Difference	ES, DiD	Common	Yes
4.	N x T Generalized DiD	DiD	Varying	Optional
5.	Event Study, Timing based	ES	Varying	No
6.	Event Study, DiD style	ES	Common	Yes
7.	Event Study, Hybrid	ES	Varying	Yes

Note: Table retrieved from the Online appendix in Miller (2023).

that the estimation of the event's effect is valid and that the results can be interpreted as causal. The main assumptions are as follows:

- **No Anticipation Assumption.** It is assumed that individuals or units of analysis do not anticipate the event and therefore do not change their behavior before the event in response to their anticipated knowledge. If units anticipate the event and adjust their behavior accordingly, this can contaminate the causal effect estimates.

- **No Confounding Assumption.** This assumption implies that no other events or policies are occurring simultaneously with the event of interest that could affect the dependent variable. The presence of concurrent events could confound the estimation of the effect of the event of interest.

- **Exogeneity of the Event.** It is assumed that the event is exogenous with respect to the units of analysis, meaning that the event is not endogenous or caused by the variables being analyzed. The event must be independent of the factors affecting the dependent variable.

- **Parallel Trends Assumption.** This assumption is crucial for event studies in a difference-in-differences (DiD) design. It is assumed that, in the absence of the event, the treated and untreated units would have followed parallel trajectories in their outcomes. This means that any difference in outcomes between the treated and control groups is due solely to the event. Mathematically, this can be represented as:

$$E\left[Y_{it}(0) \mid D_i=1\right]-E\left[Y_{it}(0) \mid D_i=0\right]= \text{ constant}$$

  where $Y_{it}(0)$ denotes the potential outcome for unit i at time t without the event, and D_i is an indicator for the treatment group.

- **Conditional Independence Assumption.** It is assumed that, after controlling for certain covariates, the treatment is independent of the potential outcome. This assumption is particularly relevant when using matching or regression methods to control for observable characteristics. Formally:

$$(Y_{it}(0), Y_{it}(1)) \perp D_i \mid X_i$$

  Where X_i represents the covariates.

- **Linearity and Additivity.** This assumption, though less commonly discussed, implies that the effect of the event is linear and additive in the specified model. In practice, this can be relaxed by using nonlinear or interaction models if necessary. The linearity and additivity can be represented as:

$$Y_{it}=\alpha+\beta D_{it}+\gamma X_{it}+\varepsilon_{it}$$

  where Y_{it} is the outcome, D_{it} is the treatment indicator, X_{it} are covariates, and ε_{it} is the error term.

Are event-study assumptions stronger regarding DiD? Event studies may have stronger assumptions if the no anticipation assumption is difficult to meet, as any anticipation can bias the results. The need for a clearly exogenous event can be a stronger challenge compared to some implementations of DiD where controlling for covariates can be more robust. Moreover, if event studies focus on very short time windows, this can make the parallel trends assumption easier to meet compared to the longer periods required in DiD. In sum, while both methodologies share many assumptions, some of the specific assumptions of event studies may be considered stronger in certain contexts, particularly regarding the anticipation of the event and its exogeneity. However, the strength of these assumptions largely depends on the specific context of the study and the quality of the available data.

3.6 Event-Study in the Context of the Paper

In the context of the paper, one of the main conclusions is that teenagers are responsive to risk information. Still, their sexual behavior is more elastic on the intensive than on the extensive margin. That is, they change even more their partner's decisions and not their sexual activity overall (having sex or not). All these results are maybe valid in the short-term. Thus, a natural question is related to the long-term effects of HIV risk information on labor or birth control, family planning, or medical care outcomes. For instance, is it that the treated group goes more often to the doctor afterward to check their health in general compared to non-treated arm? or did treatment groups face better educational or labor outcomes in the long run?

Even though the questions mentioned above are interesting, we cannot answer them empirically because the data availability does not allow it. We would need a follow-up survey or administrative records from all teenagers in the study (treated or not) for more than one year before and after the program. We would be interested in estimating the following model:

$$y_{it} = \alpha_i + \lambda_t + \sum_{\tau=-k}^{k} \beta_\tau D_{i,t+\tau} + X_{it}\Gamma + \varepsilon_{it}$$

Where y_{it} is our long-term outcome of interest (educational performance, medical care, birth decisions, labor earnings). α_i is an individual fixed effect that controls for idiosyncratic unobserved characteristics of the teenagers, λ_t is a time fixed effect to control for time-specific shocks like the economic cycle or the age of the individuals, $D_{i,t+\tau}$ is a dummy that takes the value of 1 if the teenager was exposed to the RR program and zero otherwise, $\tau \in [-k, k]$ represents the timing and $\beta_k = 0$ when $k = -1$, thus the causal effect of the RR program is captured by β_τ with respect to the baseline moment $k = -1$. X_{it} is a vector of controls for individual characteristics. ε_{it} is an error term. Standard errors should be clustered at the individual level.

Some of the expected results would be that being exposed to the RR information program would positively affect the take-up of health care in the long run. Also, it could be expected that treated teenagers face a substantially different pattern in birth control decisions, as being exposed to the RR program makes treated teenagers even more aware of the implications of good sexual health.

3.7 Critiques to the Paper

Concerns related to the internal validity of the empirical estimations:

- The author does not tackle the spillovers in the empirical estimations. This is a plausible threat when trying to estimate the causal effects of the program because the RR campaign may encourage all teenagers to share the information or discuss it. Also, there is no clear evidence that the design of the program took into account potential bias that would result from spillovers.

Concerns related to the external validity of the results:

- Even though this is compelling evidence of the effects of information campaigns on the risk of HIV contagion is a paper that is valid in contexts where sexual education is low but even more, the health system might be less efficient, thus, it is not obvious that results could be directly comparable to other contexts when the health system is more present and efficient.

References

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, *74*(1), 235–267. 3

Abadie, A., & Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, *76*(6), 1537–1557. 7

Currie, J., Kleven, H., & Zwiers, E. (2020). Technology and big data are changing economics: Mining text to track methods. *AEA Papers and Proceedings*, *110*, 42–48.
URL https://www.aeaweb.org/articles?id=10.1257/pandp.20201058 18

Dupas, P. (2011). Do teenagers respond to hiv risk information? evidence from a field experiment in Kenya. *American Economic Journal: Applied Economics*, *3*(1), 1–34. 15, 16, 17

Díaz, J., Rau, T., & Rivera, J. (2015). A matching estimator based on a bilevel optimization problem. *The Review of Economics and Statistics*, *97*(4), 803–812. 3, 5

Fama, E. F., Fisher, L., Jensen, M. C., & Roll, R. (1969). The adjustment of stock prices to new information. *10*(1), 1–21. 17

Heckman, J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. *The Review of Economics Studies*, *65*(2), 261–294. 5, 7

Jacobson, L. S., LaLonde, R. J., & Sullivan, D. G. (1993). Earnings losses of displaced workers. *The American Economic Review*, *83*(4), 685–709.
URL http://www.jstor.org/stable/2117574 17

Miller, D. L. (2023). An introductory guide to event study models. *Journal of Economic Perspectives*, *37*(2), 203–30.
URL https://www.aeaweb.org/articles?id=10.1257/jep.37.2.203 17, 18

Squicciarini, M. P., & Voigtländer, N. (2015). Human capital and industrialization: Evidence from the age of enlightenment. *The Quarterly Journal of Economics*, *130*(4), 1825–1883. 8, 10, 11, 12