



Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

Methodological considerations for a quasi-experimental approach to skill reallocation and destruction

Holm, Jacob Rubæk

Publication date:
2015

Document Version
Other version

[Link to publication from Aalborg University](#)

Citation for published version (APA):

Holm, J. R. (2015). *Methodological considerations for a quasi-experimental approach to skill reallocation and destruction.*

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy

If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Methodological considerations for a quasi-experimental approach to skill reallocation and destruction

Jacob R. Holm. Department of Business and Management, Aalborg University

Version of: 11 April 2016

Abstract

This paper is a methodological background note for the paper Holm et al. (2017). The research in Holm et al. (2017) required a number of methodological considerations, which turned out to be of a more general character and too detailed to include in the paper itself. Therefore they are presented here. The discussions in the current paper highlight the difficulties encountered when estimating the effects of factors on individual level wage – an area where endogeneity is arguably ubiquitous. The main messages are: use propensity score matching with care, evaluate balance systematically and be careful when assuming common support.

Introduction

In the paper Holm et al. (2017) we study the effect of a number of variables on an outcome variable based on wage data. These variables are all binary and they can be considered as treatments: we are interested in the effect on a person's wage of undergoing each of these treatments. The treatments are:

1. Direct = 1 if the person moves directly between jobs without being unemployed or out of the labour market in-between jobs.
2. Spin-off = 1 if the person's new job is at a spin-off firm

3. Unrelated = 1 if the person's new job is in an industry, which has a negative skill relatedness value in relation to the previous industry
4. Moved = 1 if the person has moved his/her residence from one municipality to another in-between leaving one job and entering another
5. New education = 1 if the person has acquired a new education of a similar or higher level compared to his/her previous education in-between leaving one job and entering another

The treatments as listed above are not completely self-explanatory but the above descriptions do suffice for the purpose of this note. In Holm et al. (2017) we study the percentage point change in wage that persons experience when leaving one of four specific shipyards and entering a new job. We are thus focussing on the instantaneous effect of the treatments and effects in the longer term may in principle differ considerably (Ham et al., 2009). For additional details see Holm et al. (2017).

If it is assumed that agents are rational, that they are optimizing their wage and that they are informed (or, at least, not wrong on average) about the effects of the treatments, then these five treatments are all endogenous by assumption. These assumptions are so common within economics that we must make our results robust to them, regardless of our own stand, in order to reach a wide audience.¹

The chosen solution has been to pre-process the data using matching, to exploit the structure of the data to construct a difference-in-difference estimator, and to estimate the average treatment effect on the treated (ATT). Estimating the ATT rather than the more general average treatment effect (ATE) has been chosen since the latter entails the estimation of an additional counterfactual, which relies on assumptions that are not reasonable in our data.

¹ An anonymous reviewer of an early version of Holm et al. (2017) should be thanked for emphasizing this.

Detailed discussion

Due to the possible endogeneity of the treatments we estimate their effects in a quasi-experimental setting. This entails pre-processing the data to create treatment and control groups for the analysis of each treatment. Any variable that affects both the outcome (wage in the new job) and the probability of treatment must be identically distributed in the treatment and control groups; that is, the data must be balanced between the groups. The vector of variables upon which we balance is referred to as X . The process of iteratively applying different matching methods and checking balance is tedious so with the computer power available to researchers in combination with developments in algorithms for machine learning it is a job for a computer rather than a person. This is the idea behind genetic matching (Sekhon, 2011; Diamond and Sekhon, 2013): based on a treatment variable to delimit the groups and a vector of variables upon which to balance, X , genetic matching maximizes balance. Genetic matching is related to the relatively popular propensity score (PS) matching (Caliendo and Kopeinig, 2008), and PS matching is a limiting case of genetic matching if a PS is included in X .² PS matching will result in balance asymptotically if the model used to estimate the PS is the true model. If the model for the PS is incorrectly specified then PS matching may even worsen balance (Diamond and Sekhon, 2013); a problem which was observed in early analyses for Holm et al. (2017). If data are not balanced after PS matching has been applied then the researchers may either add more observations or re-specify the PS model. Adding more data is often not feasible and with genetic matching we can be said to automate the process of re-specifying the model.

Estimator

We are interested in comparing the effects of the five treatments across shipyards, but we will estimate the ATT rather than the ATE. The preference for ATT over ATE is based on a risk that the

² The standard approach to computing a PS is by estimating the probability of receiving treatment in a logistic regression with X as the covariates.

assumptions required for estimating ATE do not hold. Let $w_k(0)$ be the wage³ of worker k if he is not treated and let $w_k(1)$ be his wage if he is treated. In order to estimate the expected wage for treated workers, $E(w(1))$, it is not sufficient to observe the wage of workers that were actually treated; it is additionally necessary to know the counterfactual wage that untreated workers would have received if they had been treated, and vice versa for $E(w(0))$. Let $s_k = 1$ if k is treated. Thus if $s_k = 1$ then $w_k(1)$ is observed and $w_k(0)$ is a counterfactual wage. In the following the subscript k is dropped. $ATE = E(w(1)) - E(w(0))$ while $ATT = E(w(1)|s = 1) - E(w(1)|s = 0)$. Thus for ATE we need two counterfactuals: the wage for the treated in case they were not treated, and the wage of the untreated in case they had been treated; while for ATT we need only the former. In order to construct the counterfactual $E(w(1)|s = 0)$ we must assume that $w(0) \perp s|X$. This is the conditional independence assumption (CIA). In words the CIA states that, conditional on X , the outcome for the untreated must be independent of treatment assignment. This implies that X must contain all variables that affect both assignment to treatment and the outcome for the untreated. In order to estimate ATE the CIA would also need to hold for the outcomes for the treated, $w(1)$, and X would need to also include all variables that affect the outcome for the treated (Caliendo and Kopeinig, 2008).

However, there are variables in X , which take on values in their observed range that could arguably make one or more treatments irrelevant. Most prominently, a worker with a university level education will in practice not acquire a new education of similar or higher level between two jobs. Thus in the analysis where we estimate the effect of the “new education” treatment, we have workers in the data for whom the probability of treatment is zero, entailing that the counterfactual $E(w(0)|s = 1)$ cannot be estimated. This problem is known as a lack of common support

³ As explained below the dependent variable is not wage level but percentage change in wage from one job to another, i.e. $w() = \log(\text{New wage}) - \log(\text{Old wage})$.

(Caliendo and Kopeinig, 2008) and entails that we cannot estimate the ATE. The problem is arguably smaller for our other treatment variables but in order to keep the analyses relatively simple we settle on estimating the ATT in all instances.

To increase the plausibility of the CIA in our case, we ensure that all variables in X are time invariant or measured before treatment is administered – i.e. before the closure of the shipyard. This ensures that no variables in X are affected by treatment. In order to increase the plausibility of the CIA even more we exploit the fact that we have data on all workers both before and after treatment to apply the conditional difference-in-difference estimator. Thus unobserved individual specific effects are differenced out and cannot be considered missing from X (Ham et al. 2011). This entails that the outcome variable in our analyses is the change in (log) wage and the effect of a treatment will be expressed as additional wage growth in percentage points. As we are not estimating the ATE we are not estimating the expected effect of treatment. Instead we are estimating the average gross gains from treatment that actually materialised, the ATT.

Data and balance

We estimate the effects of each treatment in each of the four shipyards with the exception of the spin-off treatment in the case of the B&W shipyard as only 0.26% were treated. In each case we match on 11 variables: two dummies for education, two dummies for occupation, a dummy for gender, a dummy for long tenure, a dummy for full time employment and continuous variables for age (in years) work experience (in years) and wage at the shipyard (log of constant USD). The eleventh variable is the PS estimated from a logistic regression of the first ten variables on the treatment in question. Following Diamond and Sekhon (2013) we use the linear predictor rather than the fitted probabilities. We use replacement when conducting the matching as this leads to better balance and hence less bias, though it does not necessarily increase the efficiency of estimates (Diamond and Sekhon, 2013). We specify the treatment so that the pre-matching treatment group is

always smaller than the pre-matching control group and then match with a ratio of 1:M meaning that we have M controls for each treated observation. A higher value of M will decrease balance but it will provide more data for estimates. The data in the control group are weighted so that the two groups have identical size regardless of M. Our strategy for matching is to start out with M=3. If it is not possible to achieve data balance with 1:3 matching we decrease M. If we achieve balance but the estimated ATT is not significant we increase M. If we are able to achieve balance and find a significant estimate for ATT with 1:3 matching we nevertheless explore at least one other value of M for robustness. When assessing whether balance has been achieved or not we perform two tests on each of the 11 variables. In addition to the paired t-test for equal means, we test the entire distribution by comparing Q-Q plots for the control and treatment groups. We employ a bootstrapped Kolmogorov-Smirnov test with 1000 bootstraps to test whether it can be rejected that the maximum distance between the curves is zero. The lowest p-value for this battery of tests is presented along with the ATT estimates and only when the lowest p-value exceeds 0.1 do we consider the result to be balanced.

Conclusions

In order to avoid endogeneity when estimating the effects of various treatments on wage we use a quasi-experimental approach where the data is initially pre-processed to create treatment and control groups. Genetic matching is preferred over PS matching since genetic matching is more general and automates the process of achieving balanced data.

When constructing the control group we use replacement and a ratio of 1:3 relative to the treatment group. We alter this ratio and repeat the matching to explore ranges where balance can be archived, or where a significant estimate for the ATT can be found. We choose not to estimate the ATE as the data does not allow us to make the necessary assumption.

Acknowledgements

While my colleagues C. R. Østergaard and T. R. Olesen and I have contributed equally and substantially to the paper Holm et al. (2017), I take full responsibility for any errors or omissions in the present methodological background note.

References

Caliendo, M. and Kopeinig, S. (2008) “Some practical guidance for the implementation of propensity score matching”. *Journal of Economic Surveys*, 22(1):31-72.

Diamond, A. and Sekhon, J. S. (2013) “Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies”. *The Review of Economic and Statistics*, 95(3):932-945.

Ham, J. C., Li, X. and Reagan, P. B. (2011) “Matching and semi-parametric IV estimation, a distance based measure of migration, and the wages of young men”. *Journal of Econometrics*, 161(2):208-227.

Holm, J. R., Østergaard, C. R. and Olesen, T. R. (2017) “Destruction and reallocation of skills following large company closure”. *Journal of Regional Science*, 57(2):245-265.

Sekhon, J. S. (2011) “Multivariate and propensity score matching software with automated balance optimization: The Matching package for R”. *Journal of Statistical Software*, 42(7):1-52.