Pre-Analysis Plan: 2. Relative Returns to Fields of Study in Sweden*

Adam Altmejd

2017-04-27

Contents

1 Introduction **1**
1.1 Hypothesis 2

2 Analytical Framework **2**
2.1 Causal Identification 3
2.2 Instrumentation 3

3 Data **4**
3.1 Variable Definitions 4
3.2 Sample Selection and Construction 5

References **7**

1 Introduction

The purpose of this project is to replicate Kirkeboen et al. (2016) using Swedish data and also to evaluate how well these results hold up when identified by an admission lottery rather than the quasi-randomization that occurs around the admission cutoff.

Kirkeboen et al. (2016) study the relative returns to different fields of study and institutions in Norway. They find large variation in returns to different fields of study, even after controlling for institutions and peer quality, but the increase in income from studying the same subject at a better institution is rather small. Moreover, they show that one needs detailed information about applicants' complete rankings to accurately be able to estimate relative returns. Like the original authors, we have complete application data and will be able to

*Stockholm School of Economics, adam@altmejd.se

recreate the study in the Swedish setting. In addition, in the Swedish system those applicants who are exactly at the cutoff are admitted through a lottery. We will thus be able to compare results from a true randomization to estimates produced using regression discontinuity and evaluate the validity of the original technique.

There is a growing literature using regression discontinuities at grade cutoffs to study the causal effects of education. Apart from Kirkeboen et al. (2016), also Hastings et al. (2013) do something similar on Chilean data. They do not have access to complete applications however, and their identification relies on the assumptions about constant returns across less preferred choices that Kirkeboen et al. (2016) test and reject. Ketel et al. (2016) use an admission lottery to estimate returns, but only for medical school. The tie breaking lotteries in Swedish university applications have previously been exploited by Öckert (2010) to estimate financial returns, but he only uses admissions during 1982 and does not study treatment heterogeneity across fields.

This document is a pre-analysis plan (PAP), registered in a public repository before the author has been given access to the data set needed for analysis. Since the aim of this project is a replication, I will follow the research design outlined in the original paper as closely as possible. It is still hard however to plan for all contingencies without having seen the data. If I for any reason need to deviate from this plan or from the original authors' strategy I will clearly state so.

Below, I will just describe the data that I will be using briefly, for a more detailed overview see the introduction.

1.1 Hypothesis

I expect to find similar results to the original paper. Returns to fields will vary a lot, also over differences in less prefererd fields. Wages being somewhat lower in Sweden, the distribution will look somewhat compressed but relative returns should be similar. I do not expect to find any large differences between the two estimation techniques.

2 Analytical Framework

Let $D_i \in \mathbb{J} = \{0, 1, 2, \ldots\}$ be individual i's education among a finite set of mutually exclusive and collectively exhaustive alternatives (fields of study). For each field $j \in \mathbb{J}$, $d_{ij} \equiv \mathbb{I}(D_i = 1)$ is a dummy variable that is equal to 1 if i has completed field j. We can then express i's earnings as

$$y_i = \beta_0 + \beta_1 d_{i1} + \beta_2 d_{i2} + \ldots + \varepsilon_i.$$

Where each β_j coefficient is the relative return to field $j \in \mathbb{J}$ compared to the reference field 0 (which could be the field of "no education"). Obviously, trying to estimate earnings with just education attainment like above would yield biased estimates. Instead I will use two sources of (quasi) random variation, regression discontinuity and an admission lottery. Details of these two processes can be found in the data description.

2.1 Causal Identification

For each course, some students have scores that put them just above the admission threshold while others end up just below. As has been argued in many studies, when admission thresholds vary year by year, being close to receiving an offer makes admission as good as random and can thus be used as a regression discontinuity. Selecting these students thus gives us a sample where the admission to a certain field of study is basically exogenous.

The second source of variation is from the admission lottery. In the Swedish admission system, because grades are rounded, after the e.g. 45 best students have been selected, there could be 10 students with the exact same GPA. To allocate the five remaining spots between these 10 students, a simple lottery is employed. 5 students are randomly admitted and the rest are deferred to the next choice in their ranking.

The centralized admission consists of two rounds. Students that are initially not admitted could receive an offer in the second round. Even among those that do not, some are admitted by the university at an even later stage, at times a few weeks into the semester. Using the GPA of the last admitted student as the cutoff would be inappropriate, since this student could potentially have had a lower GPA just because higher ranked students already started programs in different cities and declined their late offers. I will thus use the score of the last admitted student in the centralized system as the cutoff.

Moreover students often apply more than once, especially if they are not admitted to their first choice. Because there could be systematic differences between those that only try once and those that keep re-applying, I will only use the first lottery/discontinuity that a student is subject to when applying to a degree program.

2.2 Instrumentation

Estimating how earnings are affected by random admission to a field gives us the intention to treat effect. But since students drop out and re-apply, it will be hard to interpret estimates produced by such a model. Instead, we will focus on an instrumental variables approach, using the random admission as an instrument for completing a field. Furthermore, since it is true in Norway (as Kirkeboen et

al. (2016) show) it is highly likely that treatment effects are heteregoenous by next best field. Consider two groups that both have medicine as their preferred choice but one group has business as their second choice and the other has biology. One reason for prioritizing differently is that their expected relative returns to could be different. One group is perhaps more likely to take up a leadership position, while the other would go into medical research. As was argued by Kirkeboen et al. (2016), and discussed in the introduction, we then need to estimate the effects separately by next best field.

Let $Z_i \in \mathbb{J}$ be the field that the student was randomly admitted to (either through lottery or discontinuity). Each instrument $z_{ij} = \mathbb{I}(Z_i = j)$ is equal to 1 if the applicant has been randomly admitted to field j. Observe that for each individual, only at most one d_j, z_j will be active. Let k be the field that the student would choose if he or she was not admitted to field j. For each next best field k and individual i, the second stage is

$$y_i = \sum_{j \neq k} [\beta_{jk} d_{ij}] + X_i \gamma_k + \lambda_{jk} + \varepsilon_{ik}.$$

And for each such second stage, we need one first stage

$$d_{ij} = \sum_{j' \neq k} [\pi_{j'k} z_{ij'}] + X_i \psi_{jk} + \eta_{jk} + u_{ijk},$$

for each field $j \neq k$. With a set of control variables X_i in the baseline model to gain accuracy. The β_{jk} are the coefficients of interest, return to field j when having field k as next best alternative. In the first stage, π_{jk} can be thought of as the fraction of students randomly admitted to field j' who still end up going to field j. When $j = j'$ this is the rate of compliers, while all other $\pi_{j \neq j'}$ are fractions of always takers from each other field of study. The variables λ_{jk} and η_{jk} are sets of fixed effects from preferring field j to field k, capturing the variation in preferences between individuals. To gain precision (and following Kirkeboen et al. (2016)), we will estimate the system of equations jointly for all next-best fields, allowing for separate intercepts for preferred field and next-best field (i.e., $\lambda_{jk} = \mu_k + \theta_j$ and $\eta_{jk} = \tau_k + \sigma_j$).

3 Data

3.1 Variable Definitions

The outcome variable is income 8 years after the year when the student was admitted to the program.

As in Kirkeboen et al. (2016), we will model educational attainment as a dummy variable $d_{ij} = \mathbb{I}(D_i = j)$ taking the value 1 if individual i has finished a *degree* in field j.

Last, the controls included are gender, cohort and age at application, as well as the running variable when employing the regression discontinuity and fixed effects for preferred and less preferred field.

It is possible that the distribution of applicants that participate in application lotteries is different from the ones that are subject to regression discontinuities. For example, many of the admission lotteries are for medical school where top grades are required but where there is still oversubscription. To get the correct comparison margin we will also look at a subset of the regression discontinuity sample that has the same distribution of fields as the lottery sample.

There is a multitude of options for how to categorize higher education. The SCB variable SUN2000Grp has 97 fields, while the coarser one only has 2 for higher education. Kirkeboen et al. (2016) create their own categories: `science, business, social science, teaching, humanities, health, engineering (bsc), technology (msc), law, medicine`. We have classified the SCB variable SUN2000Grp into their categories, available in an attached spreadsheet.

3.2 Sample Selection and Construction

Before joining the application data to the individual registry we will perform a number of operations to properly identify the correct treatment margins. We follow the description given by Kirkeboen et al. (2016) through email correspondance closely.

1. Keep only applications to degree programs.
2. Remove invalid applications, where the student is not qualified or where something is missing (after imputing application scores from the applicants other applications in that admission group and year).
3. Drop applicants admitted in special quotas (apart from grades from high school and "Folkhögskola", as well as Högskoleprovet), then collapse admission groups, keeping the group where the difference between the score and the cutoff is the largest ($\max\{\text{GPA} - \text{cutoff}\}$).
4. Keep only the first observed application period where tha applicant applies to a degree program.
5. Keep only applicants with no higher education (degree) when applying.
6. Aggregate choices into fields of study and collapse rankings to get relevant choice margins (between different fields). I.e. if a student applied to med school in two different cities as their preferred choice, then to three engineering schools, and last to a business school; collapse into (j) medicine, (k) engineering, (l) business. When collapsing follow the same procedure as when collapsing admission groups, keeping the most successful outcome.

7. Keep those observations where there is (quasi) randomization on the correct margin.
   (a) When using the natural experiment, keep only those that participate in a lottery that if they win would put them in preferred field j and if they loose in field k.
   (b) For the regression discontinuity approach, this translates to cases where the applicant is predicted to be offered j (k) but would have been offered k (j) if their score was lower (higher).
8. When there are multiple relevant randomization margins:
   (a) In the lottery, keep the first that the applicant wins. I.e. if the applicant is in a lottery for field j and looses only to participate in a lottery for k that he wins, keep the j/k margin. If he would loose the second lottery, and instead be admitted to l, use the k/l margin.
   (b) Similarly for the RD approach; if the applicant is predicted to be admitted to field k, but with slightly lower grades would be admitted to l and with slightly higher to j, keep the j/k choice margin.

This will yield a data set of applicants that have been admitted to field j rather than field k, with one observation per individual. I will then join this data to the individual characteristics data set from SCB. The final data set will contain one observation per applicant.

The above description follows the procedure in Kirkeboen et al. (2016) as closely as possible. We will also however try to ameliorate their process. First and foremost, they use non-local estimation, including all applicants that are on a margin in the sample, no matter how far away they are from the cutoff. We will compare their approach with an actual regression discontinuity design approach where only applicants close to the cutoff are included. Compared to their approach, when we have identified all admission margins according to the algorithm above, we then drop those applicants that have a score too far away from the cutoff, selecting bandwidth and weights using the procedure in Imbens and Kalyanaraman (2012).

Second, we will try to make use of more of the variation in the data. For one thing, we have a sample of applicants that are exactly at the cutoff, out of which a random sample will be offered spots. This lottery sample can be joined to the discontinuity sample for an increase in power. As described in the introduction, we will also include each applicant multiple times, once for each admission margin within the bandwidth. An applicant can potentially be just below several cutoffs, but can only be just above one, the one to which they are actually admitted. Since the applications are collapsed by field, however, only a small fraction of applicants will have more than two different fields in their application. When using multiple admission thresholds, standard errors will be clustered by applicant.

References

Hastings, Justine, Christopher A. Neilson, and Seth D. Zimmerman. 2013. *Are Some Degrees Worth More Than Others? Evidence from College Admission Cutoffs in Chile*. Working Paper 19241. National Bureau of Economic Research. doi:10.3386/w19241.

Imbens, Guido, and Karthik Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *The Review of Economic Studies* 79 (3): 933–59.

Ketel, Nadine, Edwin Leuven, Hessel Oosterbeek, and prefix=van der family=Klaauw given=Bas. 2016. "The Returns to Medical School: Evidence from Admission Lotteries." *American Economic Journal: Applied Economics* 8 (2): 225–54. doi:10.1257/app.20140506.

Kirkeboen, Lars J., Edwin Leuven, and Magne Mogstad. 2016. "Field of Study, Earnings, and Self-Selection." *The Quarterly Journal of Economics* 131 (3): 1057–1111. doi:10.1093/qje/qjw019.

Öckert, Björn. 2010. "What's the Value of an Acceptance Letter? Using Admissions Data to Estimate the Return to College." *Economics of Education Review* 29 (4): 504–16. doi:10.1016/j.econedurev.2009.12.003.