Pre-Analysis Plan: 1. Introduction*

Adam Altmejd

2018-04-27

Contents

1 Introduction

Our choice of education is one of the most consequential we make. Not only does it govern future employment possibilities, it also has a large impact on financial returns (Kirkeboen et al. 2016), affects our choice of partner (Eika et al. 2017; Mare 1991), and our overall life style. In Europe, where college programs are often field-specific from the start, choosing correctly becomes even more important. Our college years are indeed formative, also for our personality. But education choices are uninformed (Hastings et al. 2016; Pekkala Kerr et al.

*Stockholm School of Economics, adam@altmejd.se

2015; French and Oreopoulos 2017), and seem to be controlled by availability heuristics and information asymmetries rather than careful premeditation. In such situations, where clear and objective information is not readily available, we tend place trust those peers who have more experience.

The purpose of the projects presented here is to evaluate the consequences of education choices from many different perspectives. To estimate causal effects of education choices we will use two sources of random variation in admissions. We will exploit the usage of tie-breaking lotteries, and the quasi-random admission of students around the score treshold.

In this document, we begin by describing the data that will be used throughout all projects. When writing this plan, we have yet to gain access to the data and all descriptions are based on information given by the collaborating agencies on the characteristics of the data sets.

In the accompanying pre analysis plans we dive deeper into a number of different research questions that will be analyzed using this data. We will (ch. 2) use the data to study financial returns to fields of study and replicate the results from Kirkeboen et al. (2016) in a Swedish setting, (ch. 3) analyze how education choices of individuals affect the behavior of their younger siblings, (ch. 4) study how college peers are influenced by the skill composition of their class, and (ch. 5) evaluate the impact of education on financial decision making and stock market participation.

In each project, we will present many different specifications and planned strategies to study each research question. We will however clearly state which of these are the main tests of the hypotheses, where the p-values can be interpreted at face value, and which should rather be seen as supplementary or exploratory, needing new data to confirm any possible findings. This disposition is especially useful because we am expecting to receive more data at a later stage, making it possible to test also these exploratory studies in a rigorous way. A second reason for this is to decrease the number of hypotheses tests, since we will have to correct the significance level depending on the number of hypothesis tests that are being performed.

1.1 Pre-Analysis Plan

The research projects are presented as pre-analysis plans (PAP), that will be registered in a public repository before the author has been given access to the data set needed for analysis. The data will be delivered by Statistics Sweden (SCB) at the earliest on October 4th, 2017. Registered PAPs are often used in experimental research to reduce the researcher's degrees of freedom. By registering the plan before generating any data (i.e. running the experiment), the researcher can avoid the risk of finding false positives through more or less conscious data mining, and any results (null results or not) become more credible (Rubin 2007; Olken 2015).

It is difficult to plan for all contingencies when working with empirical data, the generation of which is beyond the author's control. We have yet to see the data, and have not had the possibility to do any exploratory analysis. Normally, such efforts would guide the empirical research. Decisions like how to exactly define variables and how to handle missing data would be dealt with after knowing more about the exact characteristics of the problem. But it is difficult to draw the line between such exploratory work and actual data mining.

By pre registering this report I fully commit to following the plan outlined below. It is far from impossible that I will run into unexpected obstacles, perhaps requiring me to change course and not follow this plan. But should that happen I will state all such deviations and give arguments to why they are necessary, yielding at worst projects where the pre-analysis plan was not followed, which is arguably still preferable to not having one to start with.

The first version of this pre analysis plan was registered before the authors had access to any data[1]. Since then, we have ordered supplementary data from the Swedish National Archives including all applications between 1992 and 2005, and some parts of the plan have been updated. Changes to the plan can be tracked through our Github Repository. Extensions to this plan made after the first data was analyzed will therefor only be evaluated on the second data set. This new data set will also provide a robustness measure for the initial analysis. This updated plan will be registered on the OSF before we are given access to the second data set. The updated version only includes a minimal number of changes, trying to keep the plan true to its original state, even though the project has developed in several new directions since then. The most important of these changes and direction have been included here, but many others have not, and are referred to as exploratory in the papers.

The code for the project is available in this github repository. Following the history of edits one can clearly see the changes made since this pre analysis plan was published.

2 The Application System

The Swedish university application system is centralized. Applicants apply to both multi-year programs that yield degrees at different levels and individual courses that can span as little as a few months in the same application, ranking all these items (below referred to as choices) in order of preference. University credits are measured in ECTS, with 30 ECTS corresponding to one semester of full time studies. A typical bachelor's degree program consists of 180 ECTS over 3 years. In each application round, students can at most be admitted to 45 ECTS per semester, which means that one cannot start two full-time programs at

[1]The plan was registered at OSF.io on 2017-10-03 and can be accessed via this link: https://osf.io/rj6t7/.

the same time. But nothing prohibits the student from applying and registering for a second program at a later stage.

Students are allocated to choices using a serial dictatorship mechanism. The original serial dictatorship can can easily be shown to be both pareto efficient and strategy-proof (see e.g. Svensson 1999). Since there are multiple admission groups for each choice pareto efficiency is no longer certain in this setting, but it is still a weakly dominant strategy to reveal true preferences in most cases. The admission decision of a student is in no way dependent on their ranking. The ranking of choices only affects the ability to accept an offer after it has been given. The only threat to strategy proofness is that the set of choices submitted is limited to 20. Students therefor have an incentive to strategically place options with high admission probability at the end of their ranking, to avoid ending up with no offer at all. However, most applicant do not even submit lists that long, making the truncation somewhat irrelevant.

Once each semester, applicants submit a list where they rank their choices in order of preference. Some courses and programs also start at other dates, but these are fairly uncommon and will be excluded from the analysis. For each choice, applicants are ranked by their score in those admission groups (AG) that they are eligible for. There are several such groups, each using a different score to rank applicants. While many students are admitted in groups based on some version of high school grade averages, others make use of a standardized test similar to the SAT (Högskoleprovet) or go through special interviews that are conducted for certain programs. Students are automatically ranked in all admission groups that they are eligible for. Importantly, if a student chooses to write the Högskoleprovet, they are eligible for admission in that group on top of their GPA group. However, a student can never be eligible for admission in more than one GPA group.

Each university decides how many students to admit on grades, tests and through other means. Among e.g. the grade-based admission groups, the size of each group is determined by the relative number of eligible applicants. For example students with grades from an older system, and those who have supplemented their grades with post high school classes are admitted in different groups. If there are twice as many applicants with old grades, their admission group will be twice as large as the one for students with grades from the new system.

If there are more applicants than slots only the best ranked students are admitted. For example, an engineering program in Stockholm could have 50 slots in one AG. When more than 50 students that were eligible for admission in the group apply, only the 50 with the best GPA are admitted.

The process consists of two admission rounds. In the first round, going through rankings from highest to lowest, students are admitted to the first 45 ECTS where their score is high enough. Applications with lower preference ranking get automatically withdrawn. One must then decide to accept the admission and also whether or not to stay on the waiting list for preferred choices. In the

second round, students are again only admitted to the top 45 ECTS. This means that choosing to stay on the waiting list gives the option to be admitted to a preferred choice, but if that happens, the previous admission to a less preferred choice will be rescinded. After the first round, it is important to withdraw from any waiting list if preferences have changed. For example an applicant could already have found an apartment in a city where his or her second choice is located, having decided to move there he or she should not stay on the waiting list for preferred choices in other cities. Applicants on the waiting lists can be admitted up until a few days into the start of a new term, but after the results of the second round are finalized, staying on the waiting list does not mean that less preferred choices are automatically withdrawn given a late acceptance. Offers given after round 2 are administered locally by the universities themselves, and are accepted by simply showing up for class.

2.1 Admission Lotteries

All students are randomly assigned a lottery number for each of their applications at the start of the application process. The number is used to break ties at admission cutoffs. In an AG with 50 spots, there could be 10 students with the exact same GPA after the 45 best have been admitted. To allocate the five remaining spots between these 10 students, the lottery number is used. In this case the 5 students with the lowest numbers are admitted and the rest are put on the waiting list, in order based on their lottery numbers.

That a student fails the first-round lottery does not mean he or she is not admitted at a later stage. In some cases, all the students who were initially randomly put on the waiting list will be offered a spot, and no actual randomization will have occurred. In other cases, a few extra students will be admitted in the second round, but exactly who will still be random.

Students that are strongly motivated to study a specific field could interpret a failed lottery (or rejection due to being just below the threshold) as a signal that they were really close and should reapply next year. Using the random variation from a student's final lottery participation as identification is therefor potentially endogenous, as those who re-apply could be more ambitious students. It also happens that students apply for individual courses sometimes unrelated to the field associated with their degree. An Economist might have studied Art History before starting their degree program. Using the first lottery could thus also be incorrect. To get around these issues we will use the first lottery to a degree program as our source of random variation.

Tie-breaking lotteries are used throughout the admission system at all kinds of different grade levels. It's usage is much more common in certain admission groups, however. The `BGII` admission group contains all students who have somehow supplemented their high school degree. Some have retaken high school classes to improve grades, others have extended their e.g. math knowledge to be

eligible for engineering studies. This also means that it happens there are more applicants with a perfect score than there are spots, and thus that all admitted students are randomized. This is quite common for medicine, psychology and some other highly selective programs.

I will identify treatment and control groups in the lottery subsample as follows. For a specific admission group (AG), excluding those that received an offers in other groups, the sample consist of those that (a) have the exact same score as the person in their AG with the lowest score who was still admitted, and (b) are in an AG where at least one person with that score was not admitted (after the second round is finalized).

2.1.1 Alternative tie-breakers

For certain years, schools, and/or programs, lotteries were substituted with other methods to break ties. For example, in the last few years, medical programs usually break GPA ties with HP scores. Similarily, during the mid 2000's many schools chose the applicant by gender (prioritizing whichever gender was underrepresented) when ties occured. In these cases, the lotteries were either weighted by the the relative number of students of that gender, or a 50/50 selection rule was employed, where if multiple spots were allocated for students with that score each gender received half, no matter the proportion of applicants. Fortunately we are able to identify these cases and assign the correct admission probabilities. Any applicant that is not in a lottery since because of the use of the HP as tie breaker is removed from the sample.

2.2 Discontinuities at Admission Cutoffs

For each choice, 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, the fact that the exact admission threshold is moving around and that one is so close to receiving an offer, makes admission as good as random. This creates a different source of quasi-random variation that we can use to identify causal effects of education choices. Compared to the previous method, the sample size when also those slightly below and above the actual tie are included is substantially larger.

Similarly to before, we will use the score of the last admitted student as the cutoff value, and since students might apply multiple times I will only look at their first application that includes a degree program. If nothing else is stated in the individual pre analysis plans, we will use standard optimal bandwidth selection as defined by Imbens and Kalyanaraman (2012).

2.3 The Admission Algorithm in More Detail

The actual centralized admissions algorithm works as follows. After all admissions have been submitted, admission group sizes are calculated. For a specific program, a university has decided on a number of slots to offer to students in GPA-based admission groups (e.g. 3 in total), say 60. If there are 100 eligible students in group A, 200 in group B and 300 in group C, group A gets 10 slots, group B 20, and group C 30.

All applicants applying to the program are then ordered by score in all AGs that they are eligible for. For each choice, they are also given a lottery number. In each AG ordering, whenever two consecutive applicants have the same score, they are instead ordered by their random number.

The system then admits applicants one by one. Starting with the AG that is farthest away from its ideal size, group C, the highest ranked applicant in that group is given an offer and removed from the ordering for all other groups. Since there are now 29 more students to admit to group C, it is still the group farthest away from its ideal size, so also the second student is admitted to C. After the 11th student is admitted however, group B will be further away and the 12th applicant will be admitted in that group instead. This process continues until all seats are filled. Students that do not receive offers stay on the waiting list for round 2 with the same queue number as they were initially assigned.

When the process is finished many students will have multiple successful admissions. It is only now that the applicant's own preference ranking plays a part. For each applicant, only the first 45ECTS to which he or she is successfully admitted are registered. All offers below are withdrawn and the complete admission process is repeated to fill the now vacated slots. This process is repeated until it produces no new vacated slots.

When the first round is finished applicants are shown their offers and asked to accept offers and/or to stay on the waiting list for preferred offers. Since some students will reject their initial offers preferred spots will again be vacated and the admission process from round 1 is repeated in round 2. Students that decided to stay on the waiting list for preferred choices have their applications for these choices re-evaluated, and should they be admitted, their previous offer to a less preferred choice is withdrawn. When all free seats are assigned, the finalized student roster is sent to the universities along with the ordered waiting list. In some cases, not all students show up, and the institution can decide to contact students on the waiting list to fill these spots.

3 Constructing the data set

During the construction of the data set several choices have to be made that could impact results. Some choices are specific to each paper, however others

are related to the application system and apply to all projects. These data set construction decisions are listed here.

- Students apply to courses and programs in the same process. However, we are currently not able to identify the field of individual courses. Instead, as was specified already above we only look at applications to programs and remove applicants who are only admitted courses even though they have a degree program in their choice list.
- We then remove all applications to courses, yielding a data set where each applicant is admitted to at least one program.
- However, in some cases applicants are admitted to multiple programs (can happen if one program is e.g. half time). We then only look at the highest prioritized program. For example, an applicant could have a lottery for a program but also be deterministically admitted to a higher prioritized choice. We would not include these applicants in the data set.

When looking at quasi random assignment, it is possible that an applicant is close to multiple cutoffs. In our first pre analysis plan, we planned to only include the highest prioritized randomization. To increase the number of observations we will also implement a version where we try to use all application margins. An applicant can be close to many thresholds in an RD design, or participate in many lotteries. In the first case, each margin is a pair of preferred and next-best choice that can be used as a separate observation. In the second, the applicant could lose multiple lotteries. Each such lottery can then be included in the data set. In both these cases, standard errors will be clustered by individual.

For each choice, the applicant can only be allocated a spot in one admission group, even when he or she is above the cutoff in multiple groups. To identify the treatment margin, we need to collapse the admission groups. We do so by finding out which group the applicant was furthest away from the cutoff (in terms of standardized score). Using this distance to the cutoff as the running variable. When the applicant is above the cutoff in multiple groups, we still only include them in the local estimation if the highest distance to a cutoff is within the bandwidth. When no application is above the cutoff but multiple ones are exactly at the cutoff (i.e. the applicant participates in multiple lotteries) we look if any produced a successful admission and include that one. When all scores are below the cutoff we use the score that is the highest.

4 Causal Estimation

As described above, we have two sources of exogenous variation into different education alternatives that we can use to overcome the well-documented confounder of selection into education. Using this ranodmization in a statistical regression model, we can measure the causal effect of random admission on different outcomes. However, for many questions the treatment of interest is not

admission, but rather some component of the subsequent education. The estimates from the regression are intention to treat-effects. To get to the treatment effect of education we need to use instrumental variables.

The instruments need to satisfy requirements about exclusion, independence, full rank, and monotonicity to have proper identification and to be able to interpret the coefficients as local average treatment effects of the compliers (LATE). Obviously, random assignment means that the instrument is independent and it seems very unlikely that the exclusion restriction should not be fulfilled (through what other channels could randomization affect the outcome variables?). While a substantial proportion of students who fail the lottery still manage to eventually study their preferred field, these are all always takers. Since application rankings are statements of preferences it seems very unlikely that there would be defiers – students who do not complete the education if they are admitted but would have done so if they were not. Thus also the assumption about monotonicity should be fulfilled.

Furthermore, Kirkeboen et al. (2016) show that in unordered choice situations such as this, if the treatment effect is heterogeneous by next-best field, we also need to make an assumption about the irrelevance of certain choice margins and estimate the model separately by next-best field. Doing so will give us proper estimates of the LATE. Such differences can only be correctly estimated if we can differentiate between the two treatment groups. Whether or not this heterogeneity is important is an empirical question, the answer to which will vary between research questions. It is however important to note that we actually have a large number of treatment margins. Each randomization is between a preferred and next-best choice, and the variation over how this are chosen is not random. This heterogeneity of treatments will be explored in all projects presented here.

5 Data

The data consists of two parts. Registry data on the Swedish population, that includes yearly registries of all individuals from 1970 and onwards. This data is linked to two application registries. One using data archived at the Swedish National archives from years 1992–2005. In 2005, the application system was rebuilt and a new government authority started managing it. We also have data from the new system on applications from 2008–2017.

Currently, all Swedish university applications are managed by Universitets- och Högskolerådet (UHR) through their online portal Antagning.se. The application process is described in detail above. The application data will be sent directly from UHR to SCB who will match it to the Swedish population registry and connect all individual level variables before removing any identifying information.

During 1992–2005, applications were managed by Verket för Högskoleservice,

VHS. While VHS do not exist anymore, their data has been archived by the Swedish National Archives who will share a linked but anonymized version with us.

In the application data, the unit of observation is year × semester × applicant × program × institution × admission group and the variables are described in the table below. Since VHS/UHR only tracks applicants until the second round is finished, they do not know for certain if the applicant has actually started. We will therefor match the application data to the SCB university registry data to see if the applicant is actually registered.

The UHR data has the following structure:

Variable	Description
Year	2008-2016
Personal ID	Individual id, used to match with SCB data
Admission_Round_ID	Fall/Spring semester
Course_Offering_ID	Unique id for the course/program
Education_Org_ID	Unique id for the institution
Rank	Rank in the application, from 1 (best)
Qualified	1 if the applicant is qualified
Sel_Criterion_ID	Admission group
Score	Score in the AG
Result round 1	Queue pos. r1, 0 if admitted, NA if not in queue
Result round 2	Queue pos. r2, 0 if admitted, NA if not in queue
Registered	1 if student actually started the program (from SCB)

The data from VHS is divided into a complex structure and that we have yet to gain access to. We will extract the same variables as above from this registry, however.

The individual data from SCB is yearly and includes the complete population registry since 1970, with family links for parents and siblings. It is linked to the application data using the personal identification numbers before it is anonymized.

Variable	Description/LISA variable name
Year	1970-2016
Personal ID	Unique individual id
Family ID	Unique id for all individuals in one family
Parent	1 if parent in family
Birth year	
Municipality	
Gender	
Country of birth	FodGrEg

Variable	Description/LISA variable name
Family position	FamStF
Education	Sun2000Inr, Sun2000Grp
Degree year	ExamAr
Degree municipality	ExamKommun
Student status	StudDelt
Occupation info	SSyssStat, SyssStatJ, Ssyk4, SsykAr, SsykKalla
Income	KU1-4Ink, KU1-4YrkStalln, ForvInk, ForvInkNetto
Wage	LoneInk, LoneInkJ, DekLon
Student aid	Stud, StudTyp
Consumption weight (family)	KonsViktF, KonsViktF04
Disposable income	DispInkKE, DispInkKE04, DispInk, DispInk04

The population registry is complemented with data on schooling results, etc. The complete list of variables can be found in the attached excel file.

5.1 VHS Incompleteness

For some years in the VHS data, applications to certain schools are not included. Indeed, Stockholm University used their own admission system until the mid 2000s, and so did Stockholm School of Economics. While there is no way for us to know if an applicant actually applied to any of these schools, we have data on their degrees and thus whether they eventually finished. We will therefor remove any applicants from our sample that received degrees from schools that are not in the VHS data.

5.2 Including courses in the data set

Currently we have no way of identifying the field categorization of courses, which is why we drop them from the sample. While some courses are studied separately, as in the example of an Economics major taking a class in Art History, many students create degrees from accumulating a set of courses. It thus makes sense to include those students who have been (quasi) randomly admitted also to courses, if we can create appropriate field assignments. There is a small probability that this will be possible using the VHS data set, and in that case we will include also those applications.

5.3 Detailed Variable Definitions

- **Siblings and parents**: I plan to use SCB's definition of a family with some minor alterations. In their definition, everyone who lives at the

same address share a family ID number. For my purposes, I don't want the children to become part of a new family when they move from home, so I plan to use the family ID that they were assigned when they were still young. Moreover, families break up and are recreated in different constellations. I plan to use the following definition:
  - All biological siblings (sharing one or two parents)
  - Adoptive siblings that share one or two parents
  - To also include "extra" siblings, from e.g. previous marriages, I also include everyone else who is younger than the applicant and shares family ID at some point before the age of 16 for at least 3 years.
- **Field of study**: 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, and I will use the same; `science, business, social science, teaching, humanities, health, engineering (bsc), technology (msc), law, medicine`. The exact classification of SUN codes into these broader categories can be found in the attached spreadsheet.
- **Institution**: To decrease the number of institutions somewhat, I will only look at institutions with more than 300 registered students. All other I will pool on the regional (Landsting) level. If there are two 100-student schools in Stockholm, they would both go into the "Stockholm - other schools" institution.
- **City**: Captures fixed effects common for all institutions within a city. When schools lie in neighboring cities (e.g. Malmö and Lund) between which students commute, the cities are counted as one category.

References

Eika, Lasse, Magne Mogstad, and Basit Zafar. 2017. *Educational Assortative Mating and Household Income Inequality*. Working Paper 682. Federal Reserv Bank of New York. doi:10.3386/w20271.

French, Robert, and Philip Oreopoulos. 2017. "Behavioral Barriers Transitioning to College." *Labour Economics*, May. doi:10.1016/j.labeco.2017.05.005.

Hastings, Justine, Christopher A. Neilson, Anely Ramirez, and Seth D. Zimmerman. 2016. "(Un)Informed College and Major Choice: Evidence from Linked Survey and Administrative Data." *Economics of Education Review*, Access to higher education, 51 (April): 136–51. doi:10.1016/j.econedurev.2015.06.005.

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

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.

Mare, Robert D. 1991. "Five Decades of Educational Assortative Mating." *American Sociological Review* 56 (1): 15–32. doi:10.2307/2095670.

Olken, Benjamin A. 2015. "Promises and Perils of Pre-Analysis Plans." *The Journal of Economic Perspectives* 29 (3): 61–80. doi:10.2307/43550121.

Pekkala Kerr, Sari, Tuomas Pekkarinen, Matti Sarvimäki, and Roope Uusitalo. 2015. *Post-Secondary Education and Information on Labor Market Prospects: A Randomized Field Experiment*. 9372. IZA Discussion Paper Series. IZA. http://legacy.iza.org/en/webcontent/publications/papers/viewAbstract?dp_id=9372.

Rubin, Donald B. 2007. "The Design Versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials." *Statistics in Medicine* 26 (1): 20–36. doi:10.1002/sim.2739.

Svensson, Lars-Gunnar. 1999. "Strategy-Proof Allocation of Indivisible Goods." *Social Choice and Welfare* 16 (4): 557–67. doi:10.1007/s003550050160.