Methods for Observational Studies using Data from Massive Open Online Courses

by

Justin Helbert

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Bachelor of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2016

Author .... Signature redacted ....
Department of Electrical Engineering and Computer Science
January 29, 2016

Certified by .... Signature redacted ....
Kalyan Veeramachaneni
Research Scientist
Thesis Supervisor

Accepted by .... Signature redacted
Dr. Christopher Terman
Chairman, Department Committee on Graduate Theses

Methods for Observational Studies using Data from Massive Open Online Courses

by

Justin Helbert

Submitted to the Department of Electrical Engineering and Computer Science
on January 29, 2016, in partial fulfillment of the
requirements for the degree of
Bachelor of Science in Computer Science and Engineering

Abstract

Measuring the effect of a course component in online classes present an opportunity to use propensity score methods. Propensity score methods aim to balance the effect of self-selecting biases and other confounding variables that arise in observational studies like this, as each student decides what components they engage in throughout the course. This method is applied to an edX course, 6.002x, to estimate the effect of attempting homework and other assessments on students' final exam performance.

Thesis Supervisor: Kalyan Veeramachaneni
Title: Research Scientist

Contents

List of Figures

Chapter 1

Introduction

Many educational courses at the collegiate level are offered each semester to a new set of students. With the growing prevalence of public online educational courses, these online formats have certain advantages. Lectures can be divided into a sequence of shorter focused segments. Students can submit both optional and graded assignments and receive immediate feedback. While much of a course's content generally stays consistent either in a traditional or online setting, instructors continually look to improve the course. With online courses comes the ability to analyze much more detailed data on every interaction students have with the course material. Each assignment submission and video interaction can be tracked and measured. This allows more quantitative investigations of specific components in the course,

A general process of measuring effectiveness is measuring the impact of varying a component, or treatment, on a desired outcome, a process called causal inference. In this educational context, assignments such as homework, exercises, and labs are intermediate treatments throughout the course with a goal of helping students ultimately master the content, which could perhaps be assessed by exams.

The goal of this thesis is to use observational data from online courses to study the effect of individual components in a course on exam performance. Courses such as edX's MIT 6.002x generate observational data that can be used for this, although this type of dataset requires different methods than experimental studies. The observational study technique used is propensity score matching, which accounts for the nature of online courses where students self-select the amount they interact with any course components. These results can provide insight about treatment effects and inform choosing more effective in-depth experimental studies. This information may be useful to instructors and course researchers when deciding if experimental studies are worthwhile for future course runs. Current edX capabilities for experimental studies are Content Experiments, where course staff can give two groups different courseware components.

1.1 Observational versus Randomized Studies

In both randomized and observational studies, a researcher investigates the effect of a treatment T on an outcome Y. In observational studies, treatments are chosen for observed reasons (such as participant self-selection), while in randomized experiments researcher assigns treatment groups.

Randomized controlled trials are considered the ideal for estimating the effects of treatments on outcomes. Dividing participants randomly into treatment groups ensures that treatment status will not be affected by any characteristics of the subjects. This allows the effect of treatment on outcomes to be estimated by directly comparing the difference in outcomes between treated and untreated subjects. For a randomized controlled trial, the treatment T must be decided before participants engage with it, in order to separate subjects beforehand into the control and treatment groups.

In an observational study, a researcher does a posthoc analysis of the resulting data, and had not altered what occurred. The subjects, instead of the researchers, decide if they will receive the treatment, resulting in selection biases on which subjects decide to take any given treatment.

For example, students who choose a homework problem may do better on a final exam compared to those that do not attempt the homework. This effect is not necessarily solely due to the homework problem's effectiveness, but also as the type of students who choose to do the homework might have characteristics that both make them more likely to perform better on the exam and more likely to choose to do homework.

Researchers using observational education data are not able to use randomization techniques to eliminate the effects of these confounding variables. Therefore, methodology for treatment effect estimation in observational studies must account for confounding variables, referred to as covariates. The effect of these confounding variables can be minimized as much as possible but likely not completely eliminated; therefore these observational study results are more suited as indicators rather than conclusive explanations.

For observational studies, different choices of T and Y can repeatedly be run retroactively on any subset of treatments given the appropriate data was collected. Therefore observational studies are effective for preliminary indicators of the effect of T on Y; conclusions may be used to direct a more specific study later [Lawallen], such as a randomized experiment like edX A/B tests.

This thesis therefore investigates the effect of various course components on exam grades, and it uses an observational study method that controls for the self-selecting nature of students in online courses. This method is called propensity score matching, which is explained in-depth starting in the next chapter and throughout the thesis.

1.2 Challenges

- The overarching challenge in study of observational data involves determining an effect by comparing a treated population with an untreated population, where these are not randomized by the researcher but instead self-selected by the subjects which makes them not directly comparable. Therefore, to make any statistically valid conclusions, these non-comparable populations must be transformed into comparable populations. The technique used for this is the propensity score methodology.

- The propensity score method requires the ability to account for covariates that either affect or capture a student's propensity to self-select a treatment. The availability of detailed MOOC data makes capturing many covariates possible.

There are various considerations, such as the following, which are addressed in Chapter 4, Variable Extraction.

  - Not all covariates are objectively quantifiable. Some, specifically assignment grades, have a definite value. However, variables for other course components such as lecture / tutorial videos, such as how long each was watched, are not as clearly measured and must be estimated. These estimated values vary based on the methodology used.

- The propensity score methodology is not an effective solution for all possible studies. Therefore the ability to measure its effectiveness is critical to determining when it is valid. One example of its shortcomings are when the treatment and control populations greatly differ for a given treatment. When looking at which students attempt optional exercises in 6.002x, many students either attempt the vast majority of exercises, or a very small number of exercises. This leads to a very small subset of comparable students for propensity score matching on a given exercise, and therefore non-conclusive results on estimating the treatment effect of an exercise. Propensity score methods should have metrics to indicate when they are effective, and when either other factors must be accounted for or if propensity matching is not applicable for a desired study.

- A great potential of observational data is the ability to retroactively run many different trials, varying different combinations of treatment, outcome, and covariate variables. The software intrastrucure in conducting must be designed to fully allow this capability.

1.3 Contributions

- Created a software pipeline which extracts variables for each student on lecture videos, tutorial videos, homework, labs, exercises, and exams. In this study on 6002x Fall 2012 around 250 features were extracted per student. For quantifying each student's video engagement, the method detailed in Chapter 4 approximates how long a student watches a given video.

- Applied propensity score methodology detailed in "Some Practical Guidance for the Implementation of Propensity Score Matching" by Caliendo, Kopeinig 2005, to edX observational data for MIT 6.002x. This technique generates a propensity score for each student and creates comparable populations to more accurately estimate treatment effects for various course components.

- The methodology as applied to each treatment is evaluated by statistical techniques for amount of remaining self-selection biases and standard error. These give quantifiable metrics for the accuracy and level of bias for the resulting treatment effect estimation, answering both the questions of the effectiveness of the balancing and if balanced enough, the uncertainty of the treatment effect estimate. This results can indicate to the researcher if their results are sufficiently unbiased, if capturing more covariates further reduces bias, or if this specific study is not a good application for the propensity score method.

- This propensity score method is applied to three exercises, one lab, and three homework problems in a chosen week. Significant results are reached for homework problems and labs, and the propensity matching process is shown to be not as effective for the optional exercises. The software infrastructure is set up for a researcher to repeatedly choose their desired features to extract, covariates, treatment and outcome variables by specifying the corresponding module tags for either this choice of 6.002x or in application to another course.

1.4 Thesis Outline

The thesis is organized into the following chapters:

- Chapter 2 explains the propensity score method, its various forms, and metrics for evaluating its effectiveness when applied to observational studies.

- Chapter 3 provides an overview of the selected course of study, MIT 6.002x. Background is provided on the overall course, as well as the specific topic of study chosen, Second Order Circuits. Specific course components and their available datasets are outlined, resulting in an overall "feature matrix" that encompasses all variables needed for each student to conduct the propensity score method.

- Chapter 4 describes in more detail the process used for extracting variables from the edX data source to generate the feature matrix.

- Chapter 5 applies the propensity score method in Chapter 2 using the feature matrix described in chapters 3/4,. using a specific homework problem as an example.

- Chapter 6 displays and interprets the results from repeating the preceding propensity score method to selected labs, exercises, and homework problems in the Second Order Circuits section of the course.

- Chapter 7 concludes with putting these results in context for how they can inform the improvement of educational courses.

Chapter 2

Propensity Score Methods for Observational Studies

Studies investigating treatment effects aim to measure the effect of an independent variable (the treatment) on a dependent variable (the outcome). For example,

> *Do students who choose to attempt a homework problem perform better on a final exam compared to those that do not attempt the homework ?*

In this case the treatment variable is a binary value of whether a homework problem is attempted, with the goal to measure its effect on the outcome variable of exam grades. When observational data like that from online courses is used, researchers do not have control over which participants receive the treatment, in this case which students choose to do the homework problem. Therefore the difference in outcomes is not solely due to the homework problem's effectiveness, but also as the group of students who choose to do that particular homework can have characteristics or tendencies which make them more likely to do better on the exam than those that do not do the homework. Therefore observational studies must separate the treatment's effect from characteristics, called confounding variables, that affect both a student's likelihood, also called their *propensity*, to self-select a treatment and exam performance.

Students may have varying propensities of interacting with each component in the course, and the one of interest is their propensity to choose the treatment, the homework problem. Variables related to homework from previous weeks and lecture videos in the current week are examples of *covariates*, which either affect or capture a student's propensity to self-select a treatment. Building a model such as a logistic regression on these covariates allows estimation of each students propensity score: the probability a student will choose the treatment based on any measured covariates up to that point.

To minimize the effect of the confounding variables on the estimation of the treatment's effect, propensity score matching aims to compare the treated population with a new control population that has a similar distribution of student likelihoods; this works to make balance the effects confounding variables might have between the two treatment groups.

Although there are various methods for creating these balanced populations, the most straightforward and common technique is matching each member of the treatment population to the member in the control population with the closest propensity score. This attempts to generate two groups with very similar propensity score distributions, meaning the students in the newly generated "matched treatment group" are no more likely to choose the treatment than in the new "matched control group, and allows for a direct comparison in outcomes to isolate and estimate the treatment effect, by reducing the effects of the self-selection bias.

This chapter will be structured as follows:

- Section 2.1 will go into more detail into the calculation of the propensity score
- Section 2.2 will go through various matching algorithms for creating balanced treatment group populations using the propensity scores.
- Section 2.3 will analyze the matching quality including the procedure's effectiveness in reducing self-selecting bias across all covariates
- Section 2.4 will estimate the effect of the treatment using the balanced treatment groups

- Section 2.5 estimates the statistical variance of the treatment effect's value

Key terms used in this paper are:

Assessment: a homework, lab, exam, or exercise with a grade. Quantity of submissions and grades for these assessments are candidates for treatment and outcome variables, as well as covariates.

Video: Lecture or Tutorial videos. Measurements for length of time students spend on each video are included as covariates.

Module: The general term for a single unit in the courseware, such as an assessment or video.

Treatment variable: The independent variable of interest

Outcome Variable: The dependent variable of interest. A researcher wants to determine the effect of the treatment variable on the outcome variable.

Propensity score (p) : the probability a given student chooses the treatment. Its value ranges from 0 to 1.

Covariates $C_1 ... C_n$: The set of variables that either affect or capture to some degree a student's propensity to self-select a treatment.

Confounding A study has confounding variables and is therefore biased if the measure of the treatment T on the outcome Y is also influenced by some unmeasured variable(s) Z, where Z is related to both T and Y. Accounting for covariates via propensity score balancing aims to remove confounding effects from the study.

Treatment Population: All participants who *select* the treatment.

Control Population: All participants who *do not select* the treatment.

Matched Treatment Group: The subset of students from the treatment population used after matching.

Matched Control Group: The subset of students from the control population.

Average Treatment effect on the Treated (ATT): The estimated effect of the treatment on the outcome for all students who receive the treatment.

Variance Approximation: Variation (also called standard error) that accounts for all factors throughout the procedure including the normal sampling variation and estimation of the propensity score.

2.1 The Propensity Score

A student's propensity score p is the probability that a student will choose the treatment T

$$p = Pr(T = 1|C_{1...n}) \tag{2.1}$$

given all covariates $C_1...C_n$. A logistic regression is a common model for calculating this. The inputs are the covariates for each student as the features array, and the treatment labels for each student: a 1 if they received the treatment or a 0 otherwise. From this training data, the model calculates optimal weights for each covariate to minimize the error in predicting each student's label. The model will then, given an input of covariates, output the propensity score, a cnotinous variable between 0 and 1 that a student with those covariate features will receive the treatment. All covariates must be either static or occurring before the treatment to ensure they are independent of the treatment choice. In an educational course this includes covariates related to content occurring chronologically before the treatment.

2.2 Balancing with Propensity Score Matching

Attempts to balance effects of individual covariates between treatment groups leads to the curse of high dimensionality: as the number of covariates (dimensions) to consider grows, the amount of data needed to fill the space grows exponentially and becomes ultimately infeasible. Instead, the effect of confounding variables on outcome can be best reduced by balancing the aggregate effect of all covariates, as captured by the propensity score.

Once a propensity score for a given treatment is generated for each student, new treatment and control populations are generated with similar propensity score distributions: for each student in the treated group, a synthesis of one or more students in the untreated group with a similar propensity score is added to a new "matched control group". Rosenfield and Ruben demonstrate that if the propensity score distributions between the new treatment and control populations are very similar, the distribution of covariate variables is much more similar between these new groups as well. For example, if homework attempts in the preceding week is a significant covariate, then after propensity score matching the newly matched treatment and control groups will have similar distributions of homework attempts from the previous week, and similarly for all other covariates. The extent of remaining bias per covariate is a metric for the effectiveness of balancing, and this process is detailed later in the following section "Measuring Latent Bias Per Covariate".

2.2.1 Choosing A Balancing Method

The following balancing methods using propensity scores are compared in Austin 2011 and Caliendo, Kopeinig 2005.

- **Nearest Neighbor Matching.**

  1. For each participant in the treated group t_i find the participant in the control population, c_j with the closest propensity score, $min|p(t_i) - p(c_j)|$ for the all of the members t_i in the treated population.

  2. If this difference is less than a small threshold, called a *caliper*, add t_i to to the new matched treatment group and c_j, to the new matched control group. Otherwise, this pairing is not included in the new treatment groups.

  3. If doing this with replacement (which is encouraged), c_j remains a candidate for all following matches. Otherwise, remove it from the set of candidates. If matching without replacement, the order in which control group participants are considered should be randomized so it doesn't have a pre-determined effect on the matches.

  4. Repeat this for all t_i in the control group. At the end, the matched treatment group and the matched control group will have very similar propensity score distributions.

  Matching with replacement allows an untreated individual to be matched to multiple treated individuals. Matching with replacement is encouraged, which increases the average quality of matches and decreases biases (Caliendo Kopeinig 2005). This is of particular interest with data where the propensity score distribution is very different in the treatment and control group; in this case it also can lead to a much more significant sample size of matched populations. The *caliper*'s value as set by the researcher is the maximum allowed propensity difference between a match and therefore reduces bad matches.

- **Propensity Score Stratification**
  partitions the range of the propensity score into a set of intervals and calculates the impact within each interval by taking the mean difference in outcomes between treated and control observations. For example if the common support region was 0.1 to 0.9, stratification into four intervals would yield four treatment effect estimations for participants with propensity intervals 0.1-0.3, 0.3-0.5, and 0.7-0.9. This method is advantageous when wanting to study how the treatment effect varies based on propensity score. For example, the stratification method could indicate if attempting a homework problem has a different effect for students who are unlikely to do homework compared to those that are more likely. While this method can lead to more specific conclusions, some evidence suggests that it is not always as effective in removing systemic differences between populations as nearest neighbor or the following weighted matching techniques.

- **Inverse probability of treatment weighting**
  can be a useful alternative when it is difficult to directly obtain data samples (either covariates or whole participants) from a target population. Possible difficulties include time, ethical concerns, and missing data. Each participant in the treated group is weighted by the inverse of their propensity score $1/p$, or 1/(1 - p) for the untreated participants. This accounts for each participants self-selecting bias by giving more weight to samples that are underrepresented due to missing data, and therefore aims to generate populations representative of the overall population for both the treatment and control groups.

- **Kernel matching** is a process where a match is constructed for each person in the treated group from a weighted combination of multiple similar people in the control group, weighted by the closeness of their propensity score to that of the treated person. These methods lead to lower variance because more information is used for each. However, it also increases the likelihood of bad matches being used by using matches in the control group with a further propensity score difference.

2.3 Assessing Balancing Quality

With a goal of balancing the treatment and control populations, using each of these following methods measures the similarity of treated and untreated subjects.

Overlap / Common Support

The most straightforward accepted overlap approach is a qualitative visual analysis of the density distribution of the propensity score in both groups. The number of nearest-neighbor matches, without replacemen,t and using a caliper is also an indicator of propensity distribution overlap between the two treatment groups. A more involved quantitative metric involves estimations of each group's density distribution.

Measuring Latent Bias per Covariate

in the matched sample includes a comparison of the means of continuous covariates (or the distribution of the categorical covariates) between treated and untreated participants.

One suitable indicator for assessing the balance of distributions for the marginal distribution of each covariates is the standardized bias (SB) suggested in Rosenbaum and Rubin 1985. For each covariate C

$$SB = \frac{100 * \bar{C}_{treatment} - \bar{C}_{control}}{\sqrt{0.5 * (V_{treatment}(C) + V_{control}(C)))}} \qquad (2.2)$$

where $\bar{C}$ is the mean covariate value for each group and V(C) is the group's variance.

For binary variables,

$$SB = \frac{\hat{b}_{treated} - \hat{b}_{control}}{\sqrt{\frac{\hat{b}_{treated}(1-\hat{b}_{treated}) + \hat{b}_{control}(1-\hat{b}_{control})}{2}}} \tag{2.3}$$

where $\hat{b}_{treated}$ and $\hat{b}_{control}$ denote the mean of the binary variable in treated and untreated subjects.

In most empirical studies a bias reduction below 5% is considered sufficient.

2.4 Treatment Effect Estimation

At this point, propensity scores have been calculated and used to generate balanced populations, and the quality of balancing has been measured, which fulfill the requirements to get a valid estimation of the treatment effect. There are two possible metrics for any treatment effect: ATE is the average treatment effect for the whole population, and the ATT, average effect of treatment on only those subjects who ultimately choose the treatment. Propensity score methods focus on generating ATT results. After matching, the ATT (average treatment effect on the treated) is found by simply taking the difference in means of outcome variables between the two matched populations.

2.5 Standard Error of Treatment Effects

With a treatment effect value, the final step is to measure how good this value is, in terms of statistical variance, or standard error. The estimated standard error of the treatment effect includes the variance due to the estimation of the propensity score, the coverage of the common support, standard sampling error. Both bootstrapping and variance approximation (Lechner 2002) aim to measure the desired standard error that accounts for all of these components in the procedure.

- **Bootstrapping** involves repeating a complete retrial of the procedure, from the first steps of the estimation (including the propensity score calculation), where the randomization step (when matching without replacement) provides varying results. The procedure is repeated N times to give N bootstrap samples and N estimated average treatment effects. The distribution of these treatment effects is an approximation for the sampling distribution and therefore also approximates the variance of the treatment effect estimate (Caliendo, Kopeinig 2005)

- **Variance Approximation by Lechner** For the estimated treatment effect obtained via the Nearest Neighbor Matching method, the following formula is applied:

$$\tau_{ATT} = \frac{1}{N_1} Var(Y|T = 1) + \frac{(\sum_{j \in I_0}(w_j)^2}{(N_1)^2} * Var(Y|T = 0) \qquad (2.4)$$

  where T is 1 for treated individuals, 0 for untreated; Y is the value of the outcome variable; N_1 is the number of matched treated individuals. w_j is the number of times individual j from the untreated group has been used for all matched students in the untreated group I_0. This equation accounts for matching to be performed with replacement, and its results in practice vary little from the bootstrapping method and is much more efficient (Lechner 2002).

2.6 Procedure Selections for this Thesis

This thesis chooses these components to calculate the following values:

- propensity scores for the treatment variable are estimated with a logistic regression.

- two treatment groups $m_{treated}$ and $m_{control}$ with balanced propensity score distributions, using nearest-neighbor matching with replacement and a caliper.

- ATT treatment effect estimate of the difference in the outcome variable between $m_{treated}$ and $m_{control}$

- Bias remaining for each covariate after balancing, with comparisons to each covariate's bias before balancing

- Lechner's Variance Approximation. Due to matching with replacement, randomized ordering in matching is not applicable, which reduces the value of bootstrapping.

Chapter 3

Data Description

The previous chapter outlined steps for estimating a treatment effect by using propensity score methods, which require a treatment variable, outcome variable, and covariates. The next step is to determine these variables in the context of the desired observational study.

> *This study focuses on a specific week in the 6.002x course, Second Order Circuits, and aims to measure treatment effects for attempting various homework, labs and exercises on final exam problems involving second order circuits.*

This chapter starts with background information on both 6.002x as a whole and specifically Second-Order Circuits, including descriptions of the various course modules. V Variables related to these modules form the set of candidates for the treatment, outcome, and covariates, and at the end of this chapter a feature matrix will define the necessary variables for all desired effect estimations in Second Order Circuits. The next chapter, Feature Extraction, describes the implementation process for generating all these values.

3.1 6.002x Background

6.002x Circuits and Electronics is an introductory STEM course, one of the first requirements for the 6-1 electrical engineering major curriculum at MIT. edX's stated prequisites for 6.002x are that students "should have a mathematical background of working with Differential Equations and a physics background through AP level Electricity and Magnetism". MIT's prerequisite classes for 6.002 are 8.02 Physics II: Electricity and Magnetism, and 18.03 Differential Equations.

Throughout 6.002x, various problems combine conceptual knowledge of circuits with solving differential equations. One primary purpose of this thesis is to use the 6.002x course as an example study to develop the propensity score methodology for edX data. Results here are not intended to be conclusive, but rather to inspire methods of investigating educational courses going forward. Potential topics of educational studies of which 6.002x has highly relevant content of interest could include student's mastery of different types of learning, from conceptual understanding to mathematical problem solving. Different types of course modules as described in the following section could complement each other or provide options to students with varying preferences. The sequential design of the course where early weeks build the foundation for material in later weeks provides an opportunity to study how various students accelerate with, keep up, or fall off the designed pace. This thesis involves data on variables related to all of these topics. While not striving an in-depth conclusive results, it works on developing a groundwork for propensity score methodology that could spark directions for future studies in any of these directions.

3.2 Course Content Description

There are 14 weeks of material, and each week's course content is composed of various modules. The term module refers to a single unit of any of the following:

- **Lecture videos** take the form of an ordered sequence of multiple videos, typically two sequences per week. Individual videos often focus on a single concept, application, or solve an example, and are on the order of 5-10 minutes long each.

- **Exercises** are optional questions interspersed in lecture video sequences. They are related to the material just covered and are often in the form of either multiple choice or numerical answers. Once submitted, they provide immediate feedback on correctness and can provide explanations, but do not count towards a student's overall course grade.

- **Homework problems** are due at the end of each week and count in the overall grade.

- **Tutorial videos** are videos that demonstrate how to solve specific problems. These are not part of the lecture sequence, instead a supplementary part of each week.

- There are two **exams**. There is a midterm after Week 8, and a final exam after Week 14. Each has six questions.

- A course **textbook** is provided, with relevant sections referred to throughout each section of the course. This study does not cover students' usage of these text materials, as detailed student textbook activity was not available for the course in this study.

3.3 Second Order Circuits

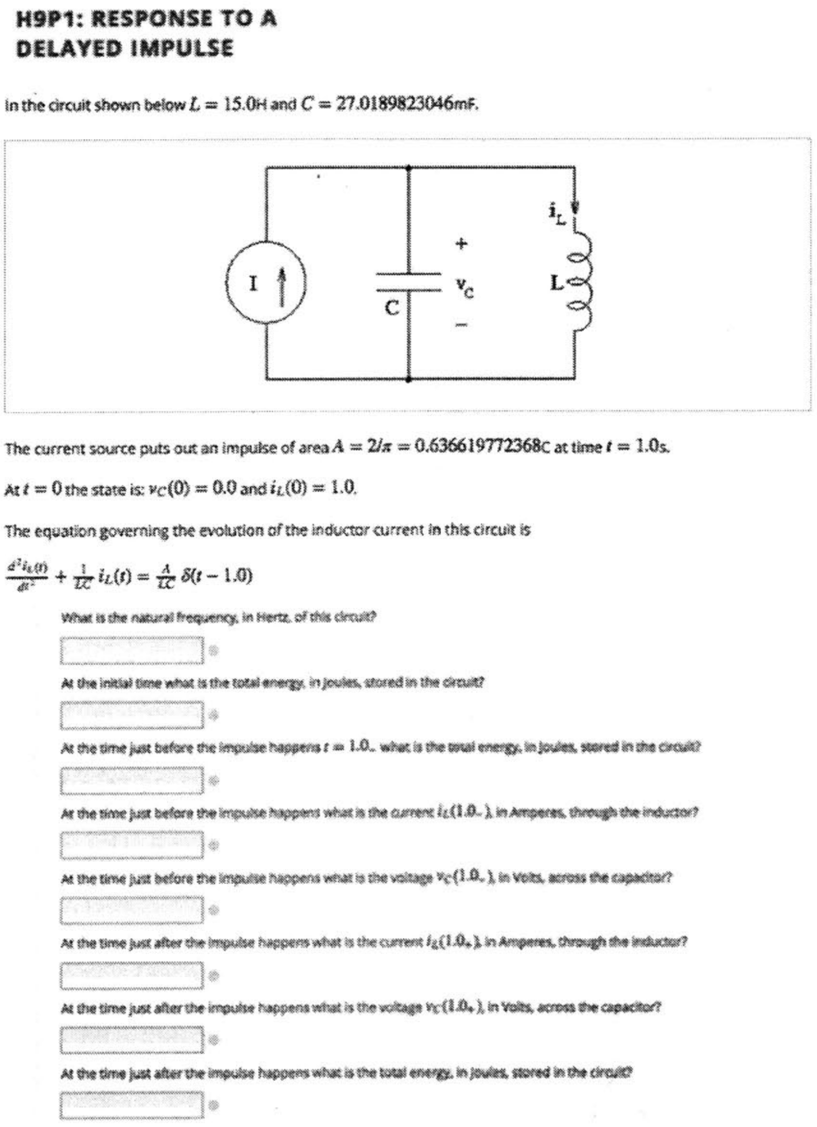
As stated above, a primary purpose in deciding the part of the course to conduct is the extent that it is related other other weeks in the course, as well as the breadth of skills its content covers. Second Order Circuits in Week 9 is a core section in the middle of 6.002x that directly builds on many concepts earlier in the course. Weeks 6 and 7 focus on first order circuits as prerequisite material for second order circuits. Solving an electric circuit involves conceptual understanding of resistor, capacitor and inductor circuit components and resulting circuit interactions, which have been covered in previous chapters. Second order circuits builds on this by involving two independent energy storage components, such as both a capacitor and an inductor. Therefore this chapter covers solving the new second order differential equations that describe these circuits, as well as resulting properties like damped circuits and oscillating voltages / currents that arise from these second order systems.

While not the focus of this study, second order circuits also provides the foundation for applications later in the course. These include Impedance (Week 10), a characteristic of second order circuits, and Week 11 covers how second-order circuits can be used as filters.

Week 9, Second Order circuits has

- 38 lecture videos divided into two sequences
- 2 supplementary tutorial videos
- 18 exercises
- 3 homework problems
- 1 lab

The following graphic illustrates a second order circuit homework problem, the one used as an example in Chapter 5: Procedure.



H9P1: RESPONSE TO A DELAYED IMPULSE

In the circuit shown below $L = 15.0$H and $C = 27.0189823046$mF.

The current source puts out an impulse of area $A = 2/\pi = 0.636619772368$C at time $t = 1.0$s.

At $t = 0$ the state is: $v_C(0) = 0.0$ and $i_L(0) = 1.0$.

The equation governing the evolution of the inductor current in this circuit is

$$\frac{d^2 i_L(t)}{dt^2} + \frac{1}{LC}\, i_L(t) = \frac{A}{LC}\, \delta(t - 1.0)$$

What is the natural frequency, in Hertz, of this circuit?

At the initial time what is the total energy, in Joules, stored in the circuit?

At the time just before the impulse happens $t = 1.0_-$ what is the total energy, in Joules, stored in the circuit?

At the time just before the impulse happens what is the current $i_L(1.0_-)$, in Amperes, through the inductor?

At the time just before the impulse happens what is the voltage $v_C(1.0_-)$, in Volts, across the capacitor?

At the time just after the impulse happens what is the current $i_L(1.0_+)$, in Amperes, through the inductor?

At the time just after the impulse happens what is the voltage $v_C(1.0_+)$, in Volts, across the capacitor?

At the time just after the impulse happens what is the total energy, in Joules, stored in the circuit?

Figure 3-1: Week 9 Homework Problem 1: Response to a delayed impulse

3.4 Overall Dataset Statistics

6.002x has multiple runs of course data available starting with Spring 2012. This study used the Fall 2012 semester. Students in this course, as common with online courses, interact with the course with widely varying intentions and levels of engagement; although some may not complete the course due to its difficulty, many may never intend to complete it in the first place. An advantage of the large number of people who can interact with online courses allows focusing on just a fraction of them and still have sample sizes in the thousands. There were:

- 106,825 users who had some interaction with the 6.002x content that semester.
- Of these, 17,380 had at least 1 submission to a problem in the course.
- 4294 students attempted the midterm, and
- 3269 attempted the final exam. As this contains the majority of student who attempt the midterm, engagement with the midterm is a strong indicator for continuing to the course's completion.

This study focuses on the student population defined as the 3,269 students who either earned at least 1 point on the final exam or attempted the second order final exam questions; this narrows the scope of the study to focus on only those students who engage with the complete course. Although this definition of students of interest is somewhat arbitrary, as discussed in the results section, this definition's variations only have a marginal impact on the effect estimation that does not detract from the overall significance. With the student population and relevant course material determined, the relevant variables can be extracted.

3.5 Extracted Feature Matrix

The following variables are extracted for 6.002x from MOOCdb. The criteria for inclusion in this matrix is as follows: variables related to all modules in Week 9

Second Order Systems, the week of focus, are included. Also included is are the same types of variables on first order circuits, which spans the second half of Week 6 and the first half of Week 7, along with any assessments related to first order circuits in those weeks.

The process for generating these feature from the edX dataset is described in the following chapter, Feature Extraction. This creates an overall feature matrix of 3269 students by 256 features.

variable	module type	selected modules	variables	value
X_{1-61}	lecture videos	Weeks 6, 7, 9	minutes watched	0 - 10
X_{62-92}	homework	Weeks 6,7, 9	was attempted	0-1
			grade	0-1 per subproblem
X_{93-122}	midterm	Questions 1 - 6	was attempted	0-1
			grade	0-1 per subproblem
$X_{122-149}$	final exam	Questions 3 and 4	was attempted	0-1
			grade	0-1 per subproblem
$X_{150-191}$	exercises	Weeks 6,7, 9	was attempted	0-1
			grade	0-1 per subproblem
$X_{192-220}$	tutorial videos	Week 7, 9	minutes watched	0 - 10
$X_{221-256}$	lab problems	Week 7, 9 Labs	was attempted	0-1
			grade	0-1 per subproblem

Figure 3-2: Feature Matrix

3.6 Treatment, Outcome, and Covariate Variables

The **target outcome** for this study is two final exam questions on second order systems. These combine for 8 subproblems, so each student's outcome variable ranges from 0 to 8.

Seven treatments in Week 9 are investigated in this study: all the three homework problems (H9P1, H9P2, H9P3), and the Week 9 Lab which ar the course grade for the week, as well as three exercises (S18E1, S18E2, S18E3).

Covariates used are the variables in the avove table related to the labs, homework, exercises, and tutorial modules that precede the chosen treatment. This range is the set of modules from Week 6 and 7 on first-order circuits and Week 9 second order circuits. All problems on the midterm exam are also included as covariates. Trials for the above treatments each use around 170 of the variables from the feature matrix as covariates.

Chapter 4

Feature Extraction Process

This section details the process for extracting each variable in the feature matrix for for each student. Two processes are used; one for extracting grades on assessments, and another for estimating time spent watching videos.

Each module in the courseware is identified by a module tag, such as H9P1 for a homework problem or S18V1 for a video. These tags are defined in a course production json file associated each course, one per semester. The set of these module tags for the desired features for each student are input into a script that generates a SQL command for each variable. These SQL scripts are executed in the MOOCdb database for 6002x Fall 2012 to generate the feature matrix.

4.1 Feature Extraction Pipeline for Assessments

Each student can have multiple submissions for each subproblem. The number of allowed submissions for each subproblem either can be limited or unlimited, at the discretion of the instructor. This extraction process counts each subproblem as correct if it is ever answered correctly; in practice this is the same value as the last submission per student, as an assessment gives the student immediate feedback of correctness after each submission if it allows multiple submissions.

The following query is generated for all assessments

```
INSERT INTO grades (user_id, problem_name, max_grade)
SELECT submissions.user_id user_id, problems.problem_name,
max(assessments.assessment_grade) max_grade
FROM problems
LEFT JOIN submissions
ON submissions.problem_id=problems.problem_id
LEFT JOIN assessments
ON assessments.submission_id=submissions.submission_id
WHERE problems.problem_name like '%H9P1%'
GROUP BY submissions.user_id, problems.problem_name;
```

The results of the previous created table are then output

```
SELECT CONCAT_WS(',', user_id, problem_name,max_grade)
FROM grades
WHERE user_id is not NULL
INTO outfile '/grades.csv';
```

This data is then converted into covariates. An assessment with n subproblems has *n+1* covariate variables, 1 for if the assessment problem was submitted, and a covariate for each of s subproblems (as all subproblems in this course have a grade of either 0 or 1 for correctness). For example, H9P1 has 8 subproblems, so it will produce 9 covariates.

4.2 Feature Extraction Pipeline for Videos

There are 60 lecture and tutorial videos from Weeks 6, 7 and 9 related to first and second order circuits. Each one has a feature for the estimated number of minutes it was watched, with ten the maximum value.

- All observed events on edX are recorded for each user with an associated timestamp. The video events are PLAY_VIDEO, STOP_VIDEO, and PAUSE_VIDEO.

All events for users are extracted during the time period of Week 9 (defined as from the end of the midterm to the due date of the homework and labs).

- For each user, each event is processed chronologically to estimate time spent on each video. Every time a PLAY_VIDEO event is recorded, the time from that event until the next event, to a maximum of ten minutes, is added to the total time spent on that video.

- This time spent per video is then converted into covariates. For this study, these covariates are NULL if the user never plays the video, the number of minutes if they spend between 0 and 10 minutes, and 10 if they spend more than 10 minutes. This covariate process is not precise, but determined to be the most effective.

- A preliminary randomized logistic regression was run for video variables predicting final exam problems scores using various cutoff strategies; having a single cutoff (at various cutoff values), and using total time spent in minutes. Using a cutoff time had a much higher correlation coefficient with the outcome variable; however there was no significant difference between cutoff values from 1 to 20 minutes. Therefore, a cutoff of ten minutes was chosen. Estimating time spent is an approximation that can be skewed by various possibilities, such as the user leaving the edX site or the user leaving the video running without watching it. Having a threshold mitigates the erroneous variance resulting from these possibilities.

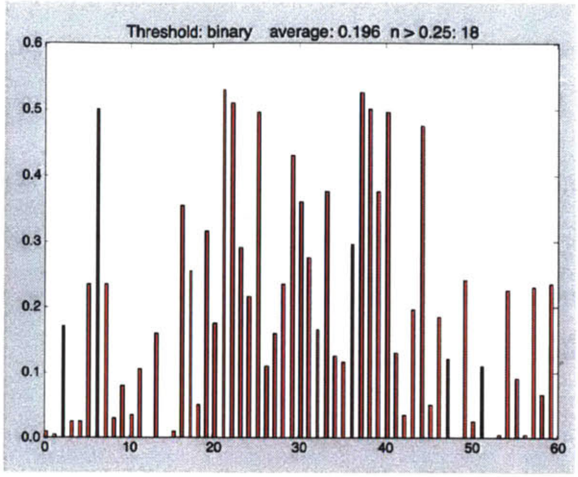


Figure 4-1: Mean correlation coefficent of 0.193 across all videos for binary threshold with one cutoff value at 1 minute in in estimating time spent watching each video

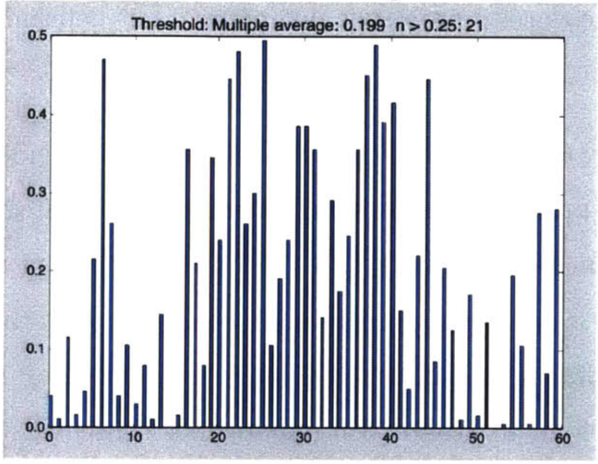


Figure 4-2: Mean correlation coefficent of 0.199 for allowing values up to a maximum of 10 minutes is comparable to the binary threshold

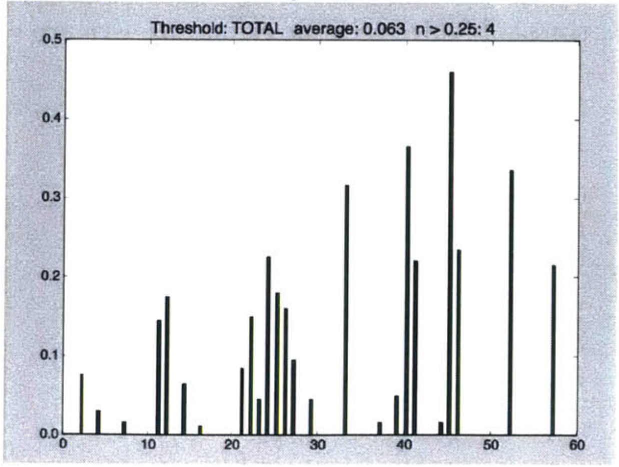


Figure 4-3: Mean correlation coefficent of 0.083 drops off significantly for total time spent watching videos without a threshold

Chapter 5

Procedure

With all variables extracted into the feature matrix, the propensity score method can be executed for a choice of treatment, outcome and covariates. This procedure section details these steps whether a student attempts the first homework problem of week 9 (H9P1) as the treatment variable. The results section in the next chapter shows the result of this method applied to all treatment variables investigated in the study.

1. **Define the student population of interest.** The population for this trial is all students who have scored at least 1 point on the final. This gives a study population of 3269 students.

2. **Define the treatment, outcome and covariate variables**, and extract them with the SQL extraction scripts from the preceding chapter.
   T = is H9P1 attempted (Homework 9, Problem 1)
   Y = Final Exam Questions 3 and 4. Y's value ranges from 0-8
   C = preceding exercises, labs, exercises, lecture and tutorial videos in Weeks 6,7 and 9, as well as midterm problems, These total to 178 covariates.

3. **Calculate the propensity score for each student.** A logistic regression is fitted with the covariates as features and the treatment as the label. Each student's propensity score is p = $P(T = 1|C_{1.n})$, the probability this model predicts student will choose the treatment based on their covariates.

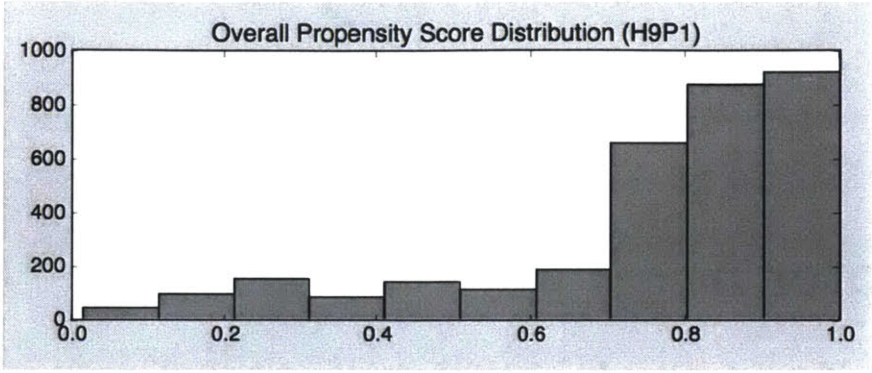


Figure 5-1: Propensity score distribution for all students in the study population. 2530 students received the treatment and 739 did not.

4. **Separate students into treatment groups.**

   All students who attempted H9P1 are placed in the treated group, and those that did not are in the untreated / control group. The propensity score distributions for each subset is shown below, with more propensity scores in the treated group skewed towards 1, and more students in the untreated group with propensity scores closer to 0.

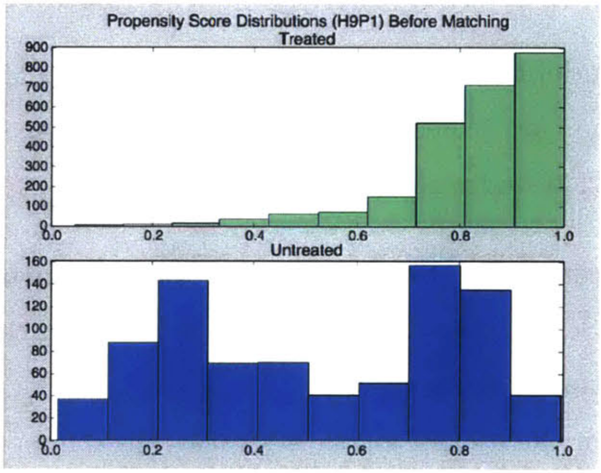


Figure 5-2: Propensity score distributions of the treated (2530) and untreated students (739). The treated student's propensity scores skew much more towards 1.

5. **Check Overlap / Common Support**

   With propensity scores for each student and a matching algorithm chosen for the specified treatment, we check the overlap of the propensity score distribution for the treatment group and the distribution of the untreated group. The students who submit H9P1 skew more towards having a higher propensity to submit it than the untreated group. These distributions show stronger support for the propensity range of 0.5 to 1.

6. **Match Propensity Scores using the nearest neighbor method.** This matches each treated student with an untreated student with the closest propensity score. The students are separated into 2 groups, treatment ($g_{treated}$) for those that submitted H9P1 and control ($g_{control}$) for those that did not. Each matching is done with replacement, where an untreated student can match multiple treated students. A caliper of 0.01 is used. This matching creates two new more balanced populations of students, $m_{treated}$ and $m_{control}$, with very similar propensity score distributions.

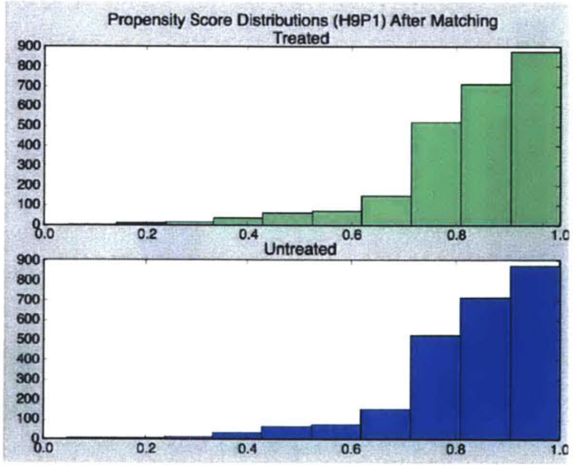


Figure 5-3: Propensity Score distributions of new treatment groups are extremely similar after matching

7. **Calculate the treatment effect on the treated.** $Y(m_{treated}) - Y(g_{control})$ The average outcome for treated students in the matched population, $Y(m_{treated})$, is 6.00 compared to 5.57 for untreated students in the population, $Y(g_{control})$, giving a treatment effect of 0.43.

8. **Calculate the standardization biases for each covariate** This metric is an indication for how biased the two generated populations re-

main. It is the difference present for each covariate between the two generation populations.

This equation, described in Chapter 2: propensity score methods is used,

$$SB = \frac{\hat{b}_{treated} - \hat{b}_{control}}{\sqrt{\frac{\hat{b}_{treated}(1-\hat{b}_{treated}) + \hat{b}_{control}(1-\hat{b}_{control})}{2}}} \tag{5.1}$$

where $\hat{b}_{treated}$ and $\hat{b}_{control}$ are the mean of the binary variable in treated and untreated groups, calculated both before and after matching.

The following graphics show the distributions of standardized biases per covariates before and after matching.

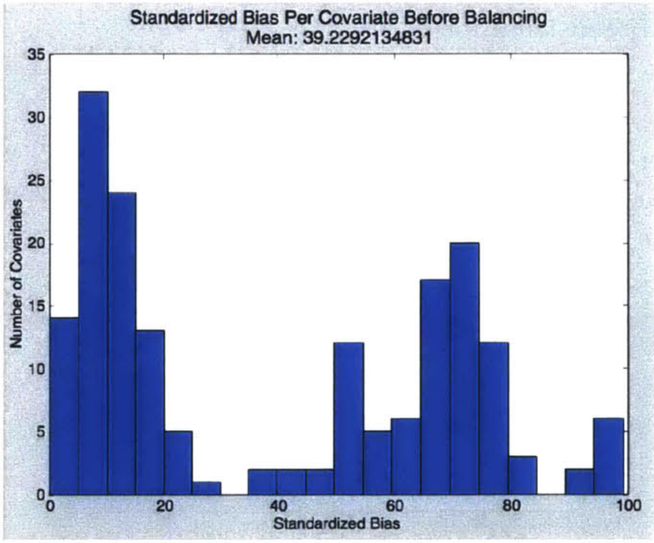


Figure 5-4: Standardized Bias per Covariates *before* matching widely vary and can have high values

Certain sets of modules are much more biased before matching than others. The six covariates with the highest standardized biases, all greater than 90% , were the Week 9 labs. The next highest biases were Week 9 exercises, ranged from 0.7 to 0.9, and the midterm questions, with biases from 50% to 80%. These captured the majority of the self-selecting bias, which makes sense in that students who engaged with the week's exercises and labs were much more

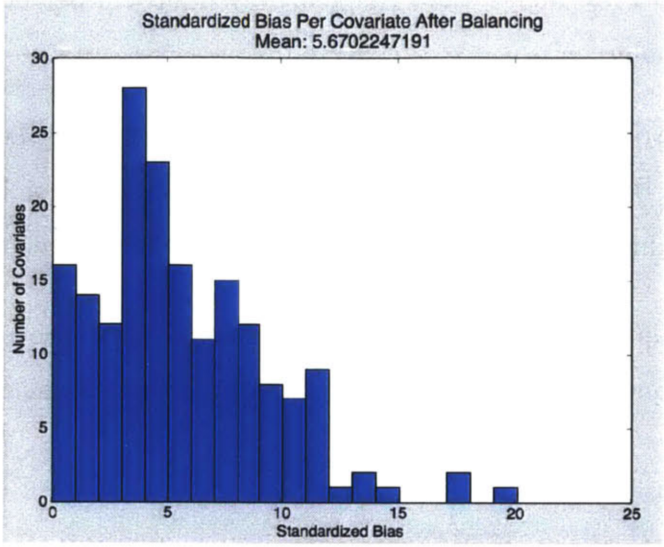


Figure 5-5: Standardized Bias per covariate *after* matching are much reduced, with a mean of 5.67

likely to attempt homework problem in that week. Homework from previous weeks was not nearly as strong of a biased covariate, with most bias levels below 6 percent.

9. **Calculate the variance of the treatment effect, using Lechner's Variance Approximation**

   $\tau_{ATT} = \frac{1}{N_1} Var(Y|T=1) + \frac{(\sum_{j \in I_0}(w_j)^2}{(N_1)^2} * Var(Y|T=0)$

   This formula, as described in Section 2.5 Standard Error of Treatment Effects, accounts for control group students being used multiple times in matching with replacement.

5.1 H9P1 Results

- For H9P1, $g_{treatment}$ has 2530 students and $g_{control}$ has 739 students.
- The average target final exam score is 6.00 for the treatment group and 5.57 for the control group, giving an ATT estimate of 0.43.
- The mean standardized bias across 256 covariates is 5.67. The mean standard-

ized bias is a little higher than the generally desired 5 percent threshold, but not unreasonable.

- The variance of the estimation of the treatment effect (using Lechner's Approximation) for H9P1 is 0.0031. There is a low sampling error in the treatment effect approximation due to a sample size of over 2000 matches.

In the next chapter: Week 9 Treatment Results, the same procedure is repeated for two other homework problems, a lab, and three exercises in the second order circuits module.

Chapter 6

Week 9 Treatment Results

The propensity score procedure run for all three homework problems (H9P1, H9P2, H9P3) and the lab (Lab 9) in the second order circuits unit, as well as 3 exercises (labelled S18E1, S18E2, and S18E3).

The terms each for each row are:

Number Treated: The number of students in group $g_{treated}$ who submitted the assessment in the corresponding column.

Number Untreated: The number of students in group $g_{untreated}$ who did not submit the assessment.

Treated Average Outcome Score: Average outcome for $g_{treated}$, the grade of for the two final exam problems

Untreated Average Outcome Score: Average outcome, the grade of $g_{untreated}$ for the two final exam problems

ATT Effect: treatment effect estimate: average outcome score difference between treated and untreated groups

Mean Standardized Bias Per Covariate: The standardized Bias calculated for each covariate averaged across all covariates, around 170 covariates per trial

Variance Approximation: Variation (also called standard error) that accounts for factors beyond the normal sampling variation such as estimation of the propensity score

	H9P1	**H9P2**	**H9P3**
Number Treated	2530	2410	2440
Number Untreated	739	859	829
Treated Average Outcome Score	6.00	6.03	6.04
Untreated Average Outcome Score	5.57	5.50	5.58
ATT Effect	**0.43**	**0.53**	**0.46**
Mean Standardized Bias Per Covariate	5.67	7.02	6.13
Variance Approximation	0.0031	0.0032	0.0032

Figure 6-1: All three Week 9 homework problems

	Lab 9
Number Treated Students	2503
Number Untreated Students	766
Average Treated Final Exam Score	5.98
Average Untreated Final Exam Score	5.67
ATT Estimate	**0.42**
Mean Standardized Bias Per Covariate	6.59
Variance Approximation	0.0032

Figure 6-2: Week 9 Lab

	S18E1	**S18E2**	**S18E3**
Number Treated Students	967	957	861
Number Untreated Students	2290	2312	2408
Average Treated Final Exam Score	5.66	5.63	5.67
Average Untreated Final Exam Score	5.52	5.45	5.76
ATT Estimate	**0.14**	**0.18**	**-0.09**
Mean Standardized Bias Per Covariate	16.67	23.80	12.89
Variance Approximation	0.0073	0.0075	0.0082

Figure 6-3: First 3 Week 9 Exercises; other exercises had similarly high standardized biases

These results indicate that homework and lab problems seem to have a significant treatment effect in magnitude on the final exam grades, with ATT estimates around 0.4-0.5. Mean Standardized Biases are a little high, ranging from a mean of 6 to 7 percent with a variance of 4.2 across all covariates; each trial has approximately 170 covariates.

The students who submit homework tend more towards having a higher propensity to submit it than the untreated group, which intuitively makes sense. The important element is that there is adequate overlap, or common support, of samples through the distribution for each side. Comparing this to the propensity score distributions of an exercise, the propensity score distributions for S18E1 skew drastically more towards the extremes

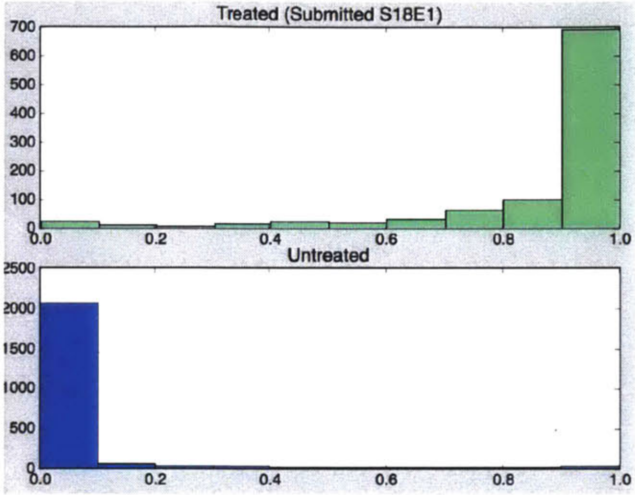


Figure 6-4: Propensity scores for treatment groups of exercise S18E1

.Students are either likely to do either nearly all exercises, or none at all. exercise ATT estimates are much closer to 0; however this imbalance even after matching leads to their Mean Standardized Bias Per Covariate being much higher. This is due to the populations of students in the treatment / control populations being. This leads to fewer comparable samples between the two groups and less of an ability to reach any valid conclusions about the treatment effect using propensity score techniques.

Another indicator is looking at the quantity of matches without replacement and with a caliper. H9P1 yields about 950 matches in this case, while S18E1 yields only around 250.

For all trials, the variance approximation is quite low, with all below 0.01. This is largely due to the high number of matches in this student population, with over 2000 for homework and labs, and over 800 for exercises. This significant sample size is a statistical benefit resulting from the large number of students per MOOC course, even when a study focuses on a fraction of them.

Chapter 7

Conclusion

Online educational courses track detailed data on their participants that has the opportunity to further inform the improvement on of educational courses. For observational studies, different choices of treatment and outcome variables can repeatedly be run retroacively on any subset of treatment modules given the appropriate data was collected. Therefore observational studies are effective for preliminary indicators; findings can inform a more specific study for later course runs, such as experimental A/B tests, which each must be predetermined and set up before students interact with it, but allow for a complete randomization of biases and therefore more confident and conclusive results.

7.1 Research Findings

- The results of this study provide evidence that individual assessments, specifically homework and labs, have an impact on student performance on final exams.

- Identifying good candidates for covariates for course modules is feasible in courses that have an order where content earlier weeks is foundational for learning content in later weeks, and many STEM courses have this format. In this study, the section on first order circuits can be examined when studying students in the week on second order circuits.

- While assessment grades have precise values, other variables related to a student's academic performance can only be indirectly observed an approximated. The amount a student watches videos or reads course material can only be approximated at best from indirect events emitted as they navigate the site. Studies using these must account for the uncertainty inherent with these variables.

- The most important covariates that capture the self-selecting nature of students are other assessments in the same week. For a homework problem, this is the exercises and lab that week.

- Propensity score matching is more applicable to some parts of a course than others. In this 6.002x study, the Week 9 homework problems and labs each had significant treatment effects in magnitude, with lower bias differences between populations after balancing. This technique is ineffective when the propensity distributions are greatly skewed towards the extremes, such as optional exercises where most students either do the vast majority of them or very few. Exercises had a much smaller effect in magnitude of treatment effect, but the comparable populations in exercises were much more unbalanced, leading to less conclusive results.

- Various types of uncertainty arise in this process, and making a distinction between statistical and methodological sources of standard error are critical to evaluating the process. In this study, statistical uncertainties such as sampling error were largely mitigated with the large number of students in the MOOC course, even when focusing on a small fraction of the students. More uncertainty comes from the both the procedure and remaining self-selecting biases. Feature approximations such as time watching videos generate imprecise metrics. The two treatment populations remain unbalanced to some extent per covariate after matching by propensity score. These have much more of an impact on statistically significant meaning than the sampling error.

7.2 Contributions

This thesis describes a methodology design for estimating effects course components have on an outcome, in this case a final exam score. This uses the observational data generated as students interact with the course, using a propensity score method to account for this self-guided interaction. The components of this are:

- Creation of a pipeline to extract features for each student on lecture videos, tutorial videos, homework, labs, exercises, and exams. In this study on 6002x Fall 2012, 256 features were extracted per student.

- With these features as input to the propensity score matching methodology, treatment effect estimates were generated for 3 homework problems, 1 lab, and 3 exercises in Week 9, Second Order Circuits on students' performance on related final exam problems.

- Each trial is evaluated by statistical techniques for amount of remaining self-selection biases and standard error. This gives indicators for the accuracy and usefulness of treatment effect estimation.

- The framework developed in this observational study creates a pipeline that can retroactively estimate the treatment effect and bias levels on various choices of outcome and treatment variables for a given course, and a way to extract relevant covariates.

Bibliography

[1] Marco Caliendo and Sabine Kopeinig. Some Practical Guidance for the Implementation of Propensity Score Matching". 2005

[2] Peter C. Austin. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. 2011

[3] Paul R. Rosenbaum and Donald B. Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. 1983

[4] Susan Lewallen, MD and Paul Courtright. Epidemiology in Practice: Case-Control Studies. 1998

[5] Franck Dernoncourt, Kalyan Veeramachaneni, Colin Taylor, and Una-May OReilly. Moocdb: Developing standards and systems for mooc data science. In Technical Report, MIT, 2013.

[6] Mitros, P.F.; Afridi, K.K.; Sussman, G.J.; Terman, C.J.; White, J.K.; Fischer, L.; Agarwal, A., "Teaching electronic circuits online: Lessons from MITx's 6.002x on edX," Circuits and Systems (ISCAS), 2013 IEEE International Symposium on , vol., no., pp.2763,2766, 19-23 May 2013

Appendix A

Sample Configuration File

```
# defines student population to study.
STUDENT_POPULATION = SCORED_ON_MIDTERM or SCORED_ON_FINAL

# covariate assessment tags fed into sql_generator.py
# tags are defined in an edX courses's production.json file
exercises = ['S12E1' , 'S12E2', ...]
hw = ['H6P3', 'H7P1' ....]
midterm = ['MTQ1','MTQ2' ...]
final = ['Q1Final2012', 'Q2Final2012' ...]
labs = ['First-order_Transients', 'Second-order_Circuits']
tutorials = [...]
ASSESSMENT_TAGS = exercises + hw + midterm + final + labs + tutorials

# video covariates extracted with video_extractor.py
INCLUDE_VIDEOS = True
VIDEO_THRESHOLD_TYPE = BINARY # Can be BINARY or INTEGER
VIDEO_TAGS = ["S18V1", "S18V2" ...]

TREATMENT = H9P1 # can be one or multiple modules
```

```
TREATMENT_DEFINITION = ATTEMPTED # can be ATTEMPTED or CORRECT

# propensity score matching
MATCH_WITH_REPLACEMENT = True
CALIPER_THRESHOLD = 0.01

# histograms
PLOT_OUTCOME_SCORE_DISTRIBUTION = True
PLOT_PROPENSITY_SCORE_DISTRIBUTIONS = True
```