Algorithms and Applied Econometrics in the Digital Economy

A DISSERTATION PRESENTED
BY
EMILY MOWER
TO
THE DEPARTMENT OF PUBLIC POLICY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
PUBLIC POLICY

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2019

Dissertation Advisor:
Professor Kris Ferreira

Author:
Emily Mower

Algorithms and Applied Econometrics in the Digital Economy

Abstract

This dissertation contains three papers that together study various aspects of the digital economy. Chapter 1 is entitled "The Impact of Financial Assistance on Online Learner Outcomes" and is an evaluation of the financial assistance program at edX, an online learning platform. I use Regression Discontinuity Design to evaluate the financial assistance program's effect on learner completion and certificate rates. I find large effects on both outcomes, even though it was free for learners to complete and pass a course during the period under study. These results indicate that learners value the signal of an edX verified certificate. To quantify the certificate's signaling value for the applicant subpopulation, I estimate a distribution of applicant willingness-to-pay. Lastly, I provide descriptive statistics for the applicant subpopulation, showing that financial assistance applicants are much more engaged than the average platform learner and also disproportionately live in countries with low or medium UN Human Development Index ratings.

Chapter 2 is entitled "Nowcasting Trends in the US Housing Market." In this chapter, I extend a state-of-the-art flu tracking algorithm called Auto-Regressive with GOogle search as exogenous variables (ARGO) to nowcast trends in the US

housing market. I show that ARGO improves nowcasting performance of housing market indicators at the state level and that ARGO with Zillow clickstream measures improves nowcasting at both the state and zip code levels. These results provide evidence that ARGO is a robust model that is applicable to economic domains and that clickstream data is a valuable source of information when used in a penalized model that avoids overfitting and is trained on a sliding window to capture changing usage patterns. To further understand the potential relevance of ARGO to economic nowcasting questions, I present preliminary evidence of ARGO's performance on nowcasting macroeconomic indicators and show that it performs reasonably well during times of economic turbulence by looking at how it would have performed during the Great Recession.

Chapter 3 is entitled "Demand Learning and Dynamic Pricing for Varying Assortments" and is co-authored with my advisor Kris Ferreira of Harvard Business School. In this chapter, we develop a demand learning and dynamic pricing algorithm for a discrete choice setting with frequently varying assortments, where products are characterized by observable attributes and demand can be described by a multinomial logit (MNL) choice model. Our algorithm follows a learn-then-earn approach to deal with the well-known exploration-exploitation tradeoff. We increase the speed of learning during the initial learning phase by introducing methods from Conjoint Analysis to dynamic pricing. We evaluate our algorithm in a 90-day field experiment with an e-commerce company and find that, relative to the company's baseline pricing policies, our algorithm led to a significant increase in revenue over the 90-day period. We measure the treatment effects using

synthetic controls and quantify the probability of observing the treatment effects using randomization inference with Fisher's exact test.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I am very grateful to have encountered so many brilliant and encouraging minds while at Harvard. In particular, I would like to thank my main advisor, Kris Ferreira, who has been an invaluable mentor, as well as the other members of my committee, Josh Goodman and Shane Greenstein, whose enthusiasm greatly enhanced my productivity and whose feedback tremendously strengthened my work. I would also like to thank Jeff Liebman, who taught me many important lessons about both policy and research. Lastly, and most importantly, I would like to thank my husband Jake and my parents for their extraordinary love and support.

Chapter 1

The Impact of Financial Assistance on Online Learner Outcomes

1.1 Introduction

[1]In recent years, the popularity of online courses has grown substantially with the introduction of websites like edX, Coursera, and Udacity. When the sites launched, they boasted laudable goals of expanding access to high-quality education, especially for traditionally underserved populations. Research on online education platforms, however, has found that the sites disproportionately serve learners from affluent neighborhoods (Hansen & Reich, 2015) and countries (Reich & Ruiperez-Valiente, 2019), and that learners of higher socio-economic status

[1]Disclosure: The author was employed by edX at the time the research was conducted.

outperform learners of lower socio-economic status in terms of persistence and certificate rates (Reich & Ruiperez-Valiente, 2019). In this paper, I provide evidence that financial constraints limit the value of online education for disadvantaged learners and show that financial assistance is a highly effective mechanism that significantly improves course engagement and outcomes for the same population.

I study the universe of edX learners who applied for financial assistance between August 2016 and January 2019. During this period, edX learners could take and pass a course for free but were required to pay in order to receive course credit in the form of a verified certificate. Verified certificates offer external proof that a learner passed a course and are obtained once the learner has paid for the course's verified track, completed and passed the course, and verified his/her identity with a webcam and a government-issued ID. During the period under study, all payments on edX had to be done with a credit card, and financial assistance covered 90% of the course's verified certificate cost, meaning learners who received financial assistance had to pay the remaining 10% cost with a credit card in order to be eligible for a verified certificate.

This paper makes three important contributions to the literature on online education platforms. First, as the first paper to study financial assistance in the online platform context, I show that financial assistance is an effective mechanism that significantly increases both completion and certification rates among low-income online learners. Previous research studying financial assistance in traditional education settings has generally found that it increases enrollment rates (e.g. Leslie & Brinkman (1993), Dynarski (2003)) and persistence rates (e.g. Mendoza et al.

(2009), Bettinger (2004)), though the effect on persistence seems to be most significant when the financial assistance is tied to academic outcomes (e.g. Dynarski (2008), Richburg Hayes et al. (2009)). For a more in-depth review of existing research on the effects of financial assistance, I refer the reader to Dynarski & Scott-Clayton (2013) and Deming & Dynarski (2009).

Second, I provide evidence of a significant population of successful, low-income, online learners whose characteristics and behavior differ significantly from previously reported trends. In particular, the majority of financial assistance applicants live in countries with low or medium UN Human Development Index (HDI) ratings, though previous research found that the vast majority of online learners live in countries with high or very high HDI ratings (Reich & Ruiperez-Valiente, 2019) and that young US learners disproportionately live in wealthy neighborhoods (Hansen & Reich, 2015). In addition, applicants in my sample have higher completion and certification rates than previously reported for the general population of online learners (e.g. Reich & Ruiperez-Valiente (2019), Reich (2014)), with no significant differences by country HDI rating.

Third, I measure the signaling value of verified certificates for low-income online learners. Specifically, I trace an empirical distribution of applicant willingness-to-pay for verified certificates. I additionally show that the majority of learners will exert sufficient effort to pass a course only when they will be able to afford a verified certificate, while a sizeable minority will exert effort regardless of whether they will be able to afford the credential. Existing research that reports completion and certification rates sheds some light on learner behavior with respect to interest

in MOOCs as a tool for knowledge accumulation apart from credential building as well as the overall cost of effort, which can be inferred from low completion rates (e.g. Reich & Ruiperez-Valiente (2019), Reich (2014)). However, I am aware of very little research measuring the value of MOOC credentials. Rosendale (2017) evalated hiring manager attitudes toward MOOC credentials and found that managers had a clear preference for traditional educational credentials, while Egloffstein & Ifenthaler (2017) examined employee attitudes toward MOOCs and MOOC credentials for workplace learning. They found that while employees were interested in earning MOOC credentials for job-relevant courses, they perceived low acceptance of such credentials among their superiors. The data I use for this paper covers 3,220 unique course runs and 86 unique prices, which is significantly more variation than has previously been available to MOOC researchers and which affords me a unique opportunity to quantify the value low-income learners place on MOOC credentials.

The rest of the paper is structured as follows. In Section 1.2, I provide an overview of financial assistance at edX, detailing the application process, selection criteria, and relevant restrictions. In Section 1.3, I describe the data used to evaluate edX's financial assistance and summarize demographic information about the applicants in my sample, including their country of residence, highest level of education, and gender. In Section 1.4, I explain the two methods used to measure the causal effect of financial assistance on edX learner outcomes and provide evidence that first-time applicants are not systematically misreporting their income in order to gain eligibility. In Section 1.5, I present the estimated

effects of financial assistance on learner completion rates, certificate rates, and persistence and also provide estimates of the verified certificate's signaling value before concluding in Section 1.6.

1.2 Financial Assistance at edX

EdX offers financial assistance to all low-income learners who submit an application and meet the income eligibility criteria, which varies by country of residence. Learners may receive financial assistance for up to five courses in a 12-month period, and a separate application must be submitted for each course for which the learner would like financial assistance.

To apply, learners are required to identify their country of residence and to complete an application form, which asks four questions. The first question asks applicants to select their annual income in US dollars from a drop down menu with five options: (1) Less Than $5,000; (2) $5,000 - $10,000; (3) $10,000 - $15,000; (4) $15,000 - $20,000; and (5) $20,000 - $25,000. There is no option on the drop down menu for income above $25,000 USD, as learners whose income exceeds $25,000 USD are not eligible for financial assistance. Additionally, applicants are asked: (1) Tell us about your current financial situation. Why do you need assistance? (2) Tell us about your learning or professional goals. How will a Verified Certificate in this course help you achieve these goals? (3) Tell us about your plans for this course. What steps will you take to help you complete the course work and receive a certificate?

All learners who submit valid responses to the application questions and whose income falls below the income eligibility threshold for their country are offered financial assistance. Learners whose income exceeds the income threshold and/or who do not submit valid responses to the application questions are not offered financial assistance. Additionally, since learners are only eligible to receive financial assistance for five courses per 12-month period, once a learner has been granted financial assistance for five courses, all subsequent applications will be denied until the 12-month period expires. In the data, these rules appear to be followed with near-perfect fidelity.

When a learner's application is approved, the learner receives a coupon code that covers 90% of the course certificate cost. The most common course certificate costs observed in the data are \$49 and \$99, which translate to \$4.90 and \$9.90 with financial assistance. The certificate cost must be paid for with a credit card, which can present a challenge for learners in certain low-income countries, where credit cards are less common (Togan-Egrican et al., 2012).

1.3 Data Description and Summary Statistics

Between August 2016 and January 2019, over 35,000 unique learners submitted valid financial assistance applications to edX. For each application, the data contains information about the course and applicant as well as the status of the application (e.g. Approved or Denied).

Course information in the data includes the course title, language, start and

end dates, certificate cost, and pacing type. Courses on edX follow one of two pacing models: self-paced or instructor-paced. All material in self-paced courses is available to learners from the day the course opens, and learners who enroll in the verified track receive a certificate as soon as they achieve a passing grade. In instructor-paced courses, material follows a schedule set by the course instructor, and learners who enroll in the verified track and achieve a passing grade receive a certificate only after the course has ended. Therefore, I drop any applications for instructor-paced courses with end dates later than January 2019, as I cannot observe the outcomes for those courses. For self-paced courses, I drop any applications submitted during the last two months covered by the data, thus allowing applicants a minimum of eight weeks from application date to complete self-paced courses. While this exclusion criterion is imperfect, it avoids selecting on outcomes and thus does not bias results. Additionally, I find the results are highly robust to a wide range of exclusion criteria.

The data also includes learner-specific course information, such as the date the learner enrolled in the course, whether the learner enrolled in the course's verified track, whether the learner passed the course, and for learners who passed the course, the date the learner passed.

Additional applicant information in the data includes the user's self-reported gender, year of birth, level of education, and country of residence. Users were only required to submit their country of residence, so all other demographic variables are missing for between 15% and 20% of applicants. Among applicants who reported their demographic information, most were young adults aged 18 to 34,

Table 1.1: Summary statistics for the applicant population

Covariates	Count	% of Applicants
Gender		
Male	23,068	65.3%
Female	6,996	19.8%
Other/Unknown	5,277	14.9%
Age		
18-24	15,945	45.1%
25-34	9,696	27.4%
35+	3,726	10.5%
Other/Unknown	5,974	16.9%
Highest Level of Education		
High School	8,746	24.7%
Bachelor's	11,687	33.1%
Master's	4,939	14.0%
Other/Unknown	9,969	28.2%

Note that applicants were not required to share this information and so it is missing for 15% to 20% of applicants, depending on the covariate.

with a wide range of educational experience. 25% of applicants reported their highest level of education as high-school, 33% reported a bachelor's degree, and 14% reported a master's degree. Men were three times more likely than women to apply for financial assistance, reflecting a significant gender gap on the platform. See Table 1.1 for more details.

The applicants came from a wide range of countries, with the largest representation from India (36% of applicants), the US (10% of applicants), and Egypt (8% of applicants). Other top countries by applicant volumes included Pakistan, Brazil, Nigeria, Canada, Mexico, Bangladesh, and Colombia. Applicant volumes for each of these countries are presented in Table 1.2. Only half of these countries have high or very high UN Human Development Index (HDI) ratings, a measure of

Table 1.2: Applicant volumes by self-reported country of residence

Country	Count	% of Applicants
India	12,737	36.0%
USA	3,370	9.5%
Egypt	2,727	7.7%
Pakistan	1,104	3.1%
Brazil	899	2.5%
Nigeria	840	2.4%
Canada	505	1.4%
Mexico	482	1.4%
Bangladesh	473	1.3%
Colombia	470	1.3%
Other	11,734	33.2%

The top 10 countries by applicant volumes are named explicitly, while all other countries are grouped together under "Other."

quality of life that accounts for life expectancy, health, educational opportunities, and standard of living[2]. Furthermore, although 51% of the world's population lives in countries with high or very high UN HDI ratings[3], I find that only 37.7% of financial assistance applicants and only 34.7% of financial assistance recipients live in such privileged countries. Therefore, I find evidence that edX's financial assistance disproportionately targets learners living in countries where opportunities for education, health, and a high standard of living are limited.

It is interesting to note that the percentage of financial assistance applicants from countries with high or very high UN HDI ratings falls well below numbers previously reported for the general online learner population. In particular, Reich & Ruiperez-Valiente (2019) estimate that upwards of 80% of online learners live

[2]http://hdr.undp.org/en/content/human-development-index-hdi

[3]http://worldpopulationreview.com/countries/

in such privileged countries, though edX platform numbers for 2018 suggest the number is now roughly 75%.

Most of the analysis in this paper considers each learner's first application only. The rationale for this restriction is that when learners submit their first application, they are similar along all dimensions other than income, and therefore, any analysis that correctly controls for income will yield valid causal effects. By comparison, after having their applications approved or denied, learners may feel more or less supported by edX, which could influence their engagement and course performance, and thus impact the validity of any estimated causal effects.

Overall, I find that financial assistance applicants represent a group of motivated low-income learners who pass courses and obtain certificates at much higher rates than previously reported for the general population of online learners. In particular, I find that applicants whose first application was accepted passed courses at a rate of 45.2% and earned certificates at a rate of 39.4%. For applicants whose first application was denied, the rates were 31.9% and 16.4%, respectively. By comparison, Reich & Ruiperez-Valiente (2019) report that during the 2016-2017 and 2017-2018 academic years, pass rates were only 3.6% for the general population of online learners, 15.9% for learners who stated an intention to complete the course, and 51.3% for learners who paid for the verified track. Among first-time applicants who paid for the verified track, completion rates were 71.7% for accepted applicants (who paid 10% of the certificate cost) and 86.4% for denied applicants (who paid 100% of the certificate cost). The higher pass rates among learners who paid 100% of the certificate price versus only 10% are consistent

with the hypothesis that learners' outcomes improve with their financial commitment, though I leave this topic for future research. Regardless, the observed pass rates are significantly higher than rates previously reported, suggesting that financial assistance applicants represent a highly enagegd and successful group of low-income online learners.

1.4 Methodology

In order to measure the causal effects of financial assistance on learner outcomes, I use two methods: Regression Discontinuity Design (RDD) and Pearson's chi-squared test.

1.4.1 Regression Discontinuity Design

RDD measures causal effects in the presence of a sharp and undisclosed eligibility threshold. In the case of edX financial assistance, country-specific income thresholds fully determine whether a valid financial assistance application is approved or denied. The thresholds are not publicly disclosed, and I do not find evidence of income manipulation for first-time applicants, which suggests learners are unaware of the threshold locations.

In order to protect the confidentiality of edX's financial assistance policies, I do not disclose the threshold location for any country. Instead, I center data from each country on its respective eligibility threshold and report pooled estimates

that include data from all countries using the following regression specification.

$$y_i = \beta_0 + \beta_1 Inc_i + \beta_2 Over_i + \beta_3 Inc_i \times Over_i + \epsilon_i \tag{1.1}$$

Where y_i represents learner i's outcome (e.g. passed), Inc_i represents learner i's self-reported income, centered so that the threshold in learner i's country maps to zero, $Over_i$ is a binary indicator that equals one if learner i's self-reported income falls above the eligibility threshold and equals zero otherwise, and ϵ_i is a random error term. The income data was centered around zero so that the highest income bucket eligible for financial assistance mapped to -0.5 and the lowest income bucket ineligible for financial assistance mapped to 0.5. I used a bandwidth of three income buckets with a rectangular kernel, so my estimates are based off the six central income buckets, which range from -2.5 to +2.5 and which represent \$10K-\$15K below the threshold, \$5K-\$10K below the threshold, <\$5K below the threshold, <\$5K above the threshold, \$5K-\$10K above the threshold, and \$10K-\$15K above the threshold. To ensure robustness in the results, I also ran the RDD specification for the pooled sample including indicators for three age groups (18-24, 25-34, and 35+), gender (male and female), and three education levels (high school, bachelor's degree, and master's degree). Let Z_i be the vector including these 8 covariate indicators. The specification with covariates is thus

$$y_i = \beta_0 + \beta_1 Inc_i + \beta_2 Over_i + \beta_3 Inc_i \times Over_i + \boldsymbol{\beta_4} Z_i \epsilon_i \tag{1.2}$$

In practice, I find the results are unaffected by the inclusion of covariates.

A drawback of the pooled approach is that the only two points guaranteed to include data from all countries are those directly above and below the centered threshold (in every country, at least one income group is eligible for financial assistance and at least one income group is ineligible). Specifically, points further to the left of the centered threshold represent only learners from countries with higher thresholds while points further to the right represent only learners from countries with lower thresholds.

To maintain consistency across the running variable, I therefore also estimate within-country effects for India and the US using the following regression specification

$$y_i = \beta_0 + \beta_1 Inc_i + \beta_2 Over_i + \epsilon_i \tag{1.3}$$

I chose India and the US, because they represent the largest applicant volumes (which collectively account for 46% of all applicants), have different economic conditions, and have different eligibility thresholds. I find very similar treatment effects for both countries, and the effects remain large and highly statistically significant. Unlike the pooled RDD specification, the within-country specification does not allow the slope of the regression line to change across the threshold. This is because with only five income buckets, one side of the threshold will have two or fewer income buckets, so allowing the slope to change would result in over-fitting.

When using RDD to measure causal effects, the principal assumption is that applicants just above and below the threshold are extremely similar; therefore, any major change in outcomes at the threshold can be attributed to the effect of treatment. The running variable in RDD is the variable that determines eligibility.

Normally, this variable is continuous and can realize a fairly wide range of values. With income as the running variable, it would be easy to imagine that someone who made $29,999 and someone who made $30,001 were basically the same and that it was essentially luck that landed the first person below a $30,000 threshold and the second person above the same threshold.

A major challenge with applying RDD to the study of financial assistance at edX is that the running variable only realizes five unique values, and each value includes a range of $5,000 USD for people whose income is at most $25,000 USD. Therefore, in adjacent income buckets, it would be possible to have learners whose income differed by as little as one dollar or by as much as $10,000. It is thus inappropriate to assume that randomness alone is responsible for landing learners just above versus just below an income threshold, though it may still be an appropriate assumption for some learners. Conveniently, however, prior research has shown that learner outcomes improve with income (Hansen & Reich, 2015), so any challenges arising from the discreteness of the running variable should downward bias the estimated effect sizes.

1.4.2 Pearson's Chi-Squared Test

To deal with the challenge presented by the extremely discrete running variable, I also estimate simple differences in pass rates and certificate rates for learners in the income buckets just above versus just below the income eligibility threshold. I use simple asymptotic confidence intervals to estimate statistical significance, which are valid in sample sizes much smaller than those considered in this paper (New-

combe, 1998). The resulting test of statistical significance is known as Pearson's chi-squared test. To estimate the confidence intervals, let p_a be the proportion of learners in Group A with a positive outcome (e.g. learners who passed) and let n_a be the total number of learners in Group A. Define p_b and n_b analogously for Group B. Then, for the difference between p_a and p_b, the two-sided asymptotic confidence interval with confidence level 1 - α is defined as

$$(p_a - p_b) \pm z \times \sqrt{\frac{p_a \times (1 - p_a)}{n_a} + \frac{p_b \times (1 - p_b)}{n_b}} \tag{1.4}$$

Where z is the critical value from the Normal Distribution corresponding to $\alpha/2$. This simple differences approach yields very similar estimates to the RDD approach, and the estimates remain large and highly significant.

1.4.3 Examining Threats to Validity

The primary concern when comparing applicants across the threshold is that if learners manipulate their reported income in order to qualify for financial assistance, then learners on either side of the threshold differ on unobservable qualities in addition to observable income. For example, learners who under-report their income specifically to gain eligibility may be more motivated to gain credit for courses than learners who do not under-report their income. This would upward bias the results and undermine the validity of any estimated effect.

To check for income manipulation, I use the McCrary test and look for covariate balance across the income eligibility threshold. The McCrary test examines

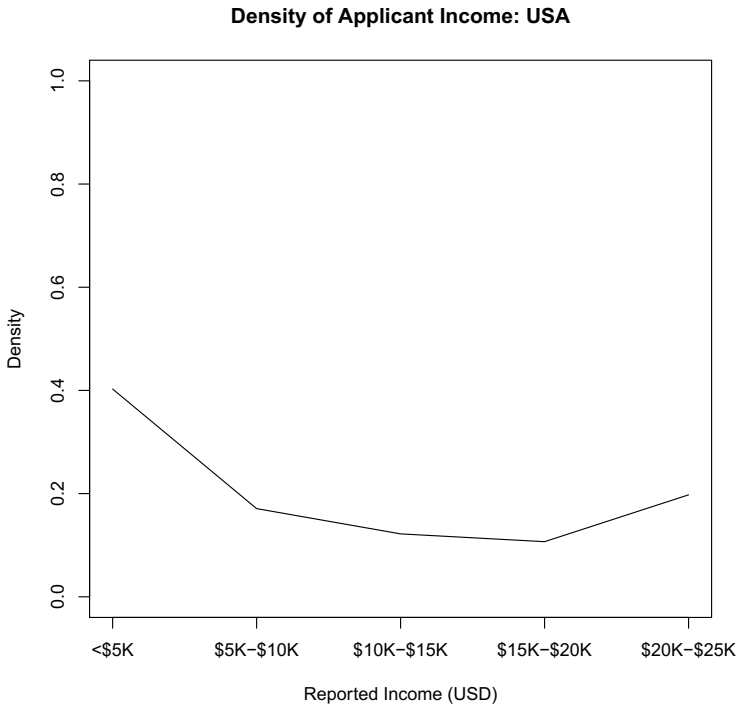
the density of the running variable (e.g. self-reported income) and looks to see whether there is any evidence of bunching below the threshold. Bunching below the threshold would suggest people were aware of the threshold's location and manipulated their self-reported income in order to gain eligibility. Similarly, any discontinuities in covariates across the threshold could also be indicative of income manipulation.

A major challenge in applying these tests to the pooled sample is that the running variable is extremely discrete and the composition of countries changes with the running variable. As a direct result of the discretely changing composition, the density is irregular and it is impossible to distinguish irregularities due to population changes from irregularities due to income manipulation. Therefore, rather than presenting density results for the pooled sample, I present them individually for the two largest countries by applicant volumes, India and the US, which collectively account for nearly half of all first-time applicants. I find no evidence of income manipulation in either country. Furthermore, since income manipulation would only be a concern if it upward biased results and since the estimated effects sizes for these two countries are somewhat larger than for the full-sample, I conclude that applicants are not meaningfully under-reporting their incomes in order to gain eligibility for financial assistance.

The United States

To check for income manipulation among US applicants, I use the McCrary test and plot the density by income in Figure 1.1. To protect the confidentiality of

Figure 1.1: Density of self-reported income for US applicants



The density of self-reported income for US applicants shows no evidence of income manipulation or bunching. To protect the confidentiality of edX's financial assistance policies, I am unable to share the exact threshold location for any country. However, I can reveal that the US threshold is not located at $5,000 USD, which is the only threshold location where one would worry about bunching directly below the threshold.

edX's financial assistance policies, I cannot reveal the location of the US threshold other than to say it does not fall between the <$5,000 bucket and the $5,000 to $10,000 bucket, which is the only point on the distribution where one might be concerned about bunching below the threshold. Therefore, I find no evidence of income manipulation for the general population of US applicants.

I next examine income density by covariates. Figure 1.2 shows the income distribution by gender, Figure 1.3 shows the income distribution by age, and Figure 1.4 shows the income distribution by education. In each of these plots, the income distribution is smoothly changing, and I find no evidence of bunching. One oddity of the plots is that many of them exhibit a slight increase in the highest income bucket. This is likely due to the fact that $25,000 is a relatively low income in the United States, and thus it is likely that some applicants whose true income exceeded $25,000 selected the highest available option.

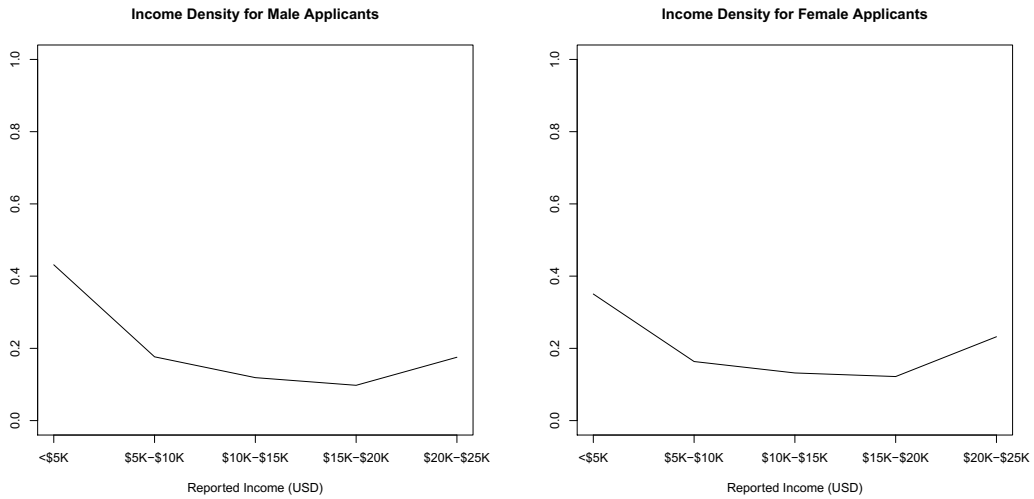
India

In 2017, India's GDP per capita and GNI per capita were both under $2,000 USD[4]. Therefore, it seems likely that most financial assistance applicants from India would have incomes below $5,000 USD and would thus fall under the lowest income bucket. In this case, the density is unlikely to be smooth and the standard McCrary test is inappropriate. Therefore, I instead use data from the Indian Human Development Survey[5] to estimate the income distribution for people in

[4]https://data.worldbank.org/country/india

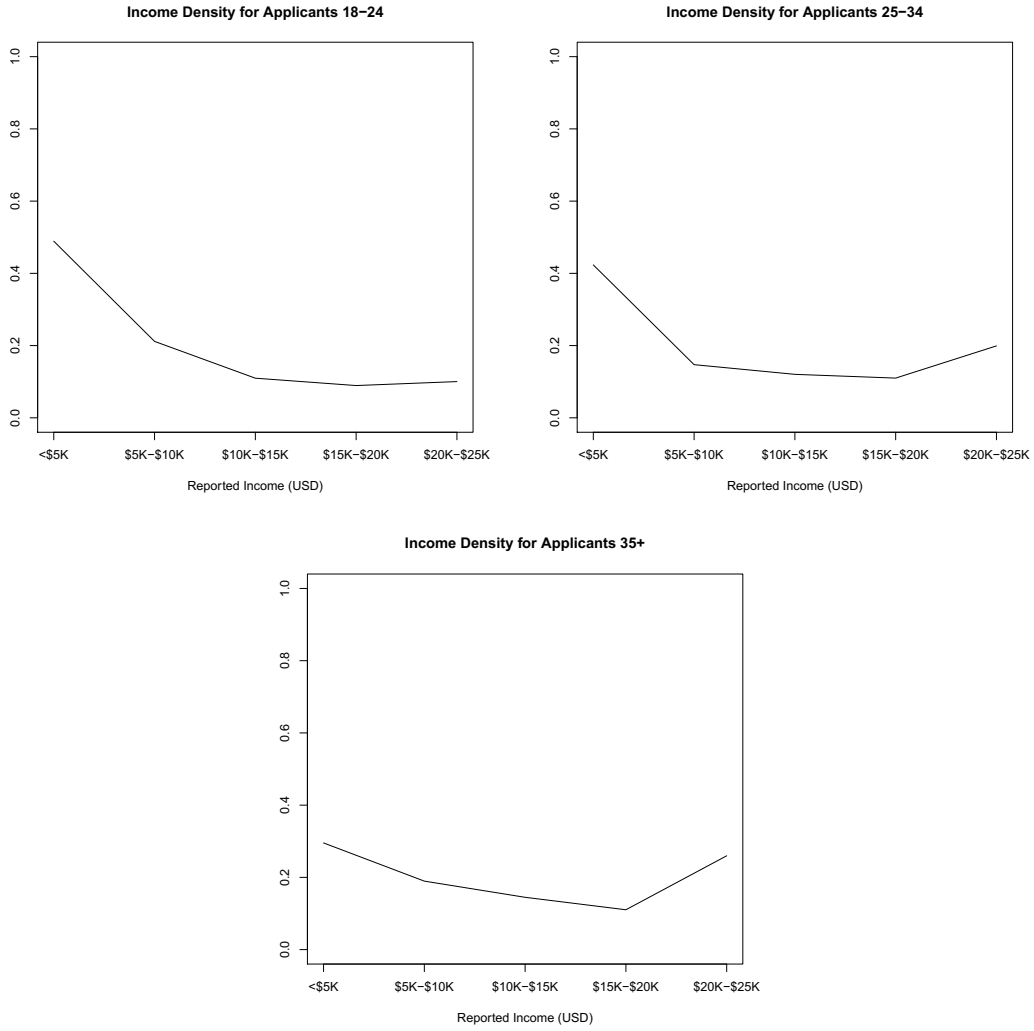
[5]https://ihds.umd.edu/

Figure 1.2: Density of self-reported income for US applicants by gender



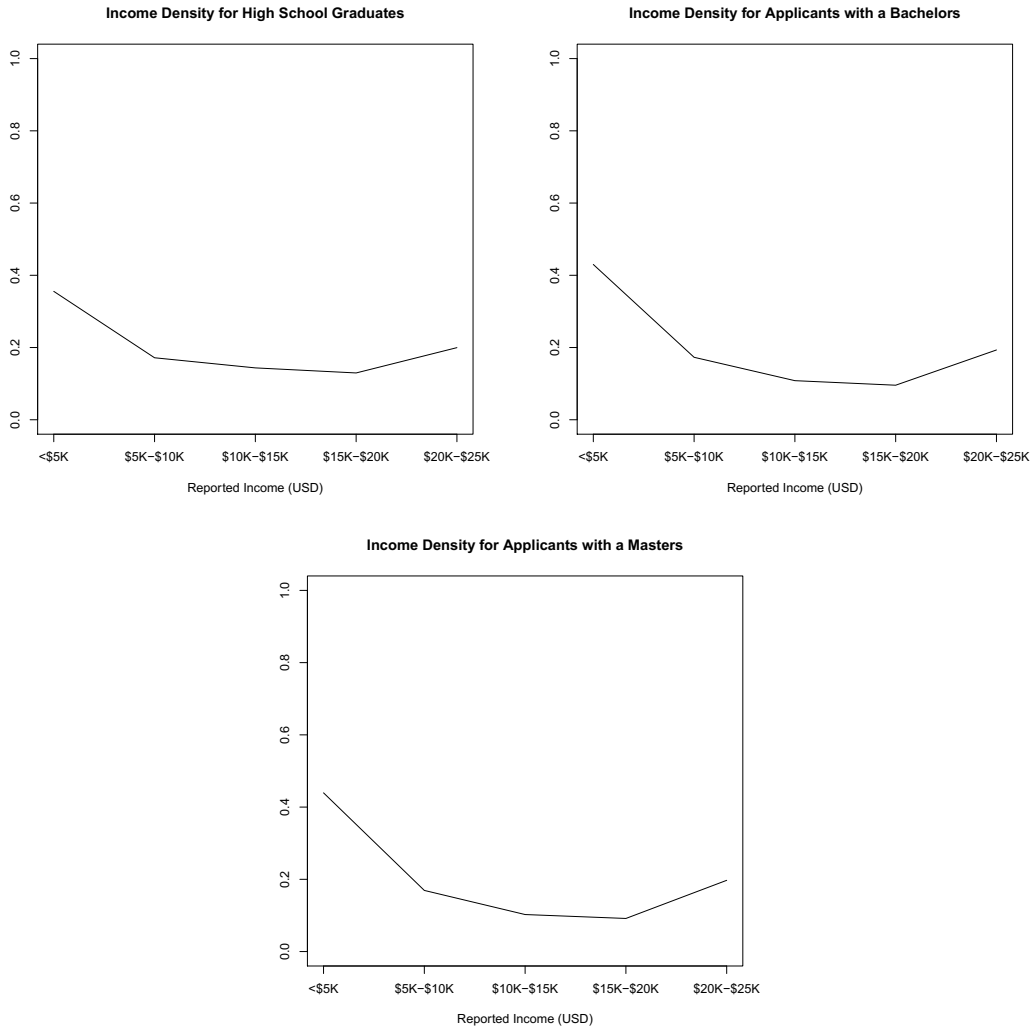
Male (left) and female (right) applicants from the US are fairly evenly distributed throughout the income distribution, with patterns closely resembling that of the overall applicant income density for the US. Therefore, there is no evidence of sharp covariate changes across the threshold, which are generally considered indicative of income manipulation.

Figure 1.3: Density of self-reported income for US applicants by age



Age varies smoothly across the income distribution with no evidence of bunching below any candidate threshold. The density bends slightly upward in the highest income bucket for applicants in the 25-34 (upper right) and 35+ (bottom center) age ranges, likely due to the fact that the highest reportable income was $25,000 USD. Applicants with incomes above $25,000 USD may therefore have selected the highest option available to them rather than deciding not to submit an application at all.

Figure 1.4: Density of self-reported income for US applicants by education



Educational attainment varies smoothly across the income distribution with no evidence of bunching below any candidate threshold. The density bends slightly upward in the highest income bucket, likely due to the fact that the highest reportable income was $25,000 USD. Applicants with incomes above $25,000 USD may therefore have selected the highest option available to them rather than deciding not to submit an application at all.

India with at least a secondary education. I then compare the population income distribution to the applicant income distribution to see if there are any noticeable differences that would suggest applicants are misreporting their incomes in order to qualify for financial assistance. I restrict the population figures to individuals with at least a secondary education, because among Indian applicants who reported their highest level of education, over 98% reported completing secondary school or higher. I then repeat the equivalent analysis for covariates, comparing the income distribution for applicant subpopulations (e.g. women applicants) to their population equivalents (e.g. all women in India with at least a secondary education).

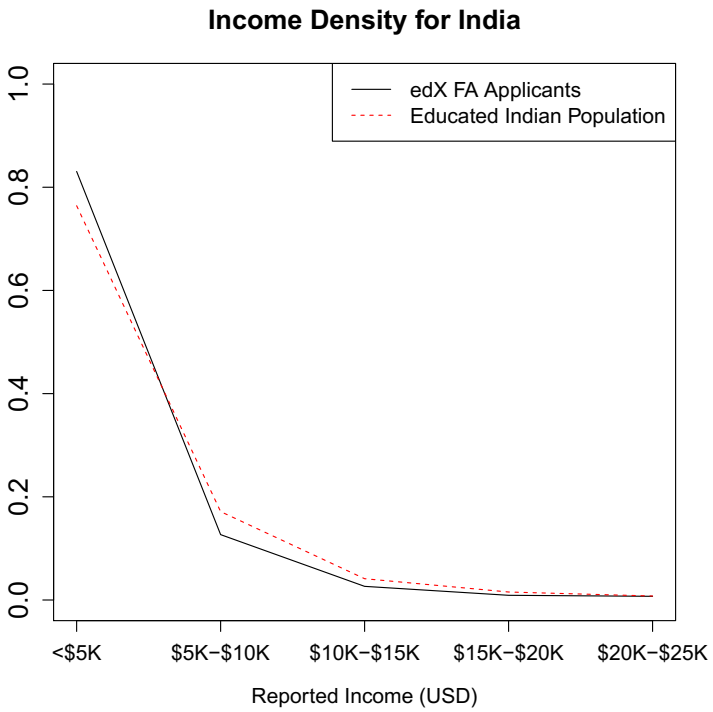
Figure 1.5 shows the distribution of self-reported income for applicants from India compared to the income distribution of everyone in India with at least a secondary education. I inflation adjusted the 2011-2012 Indian Human Development Survey (IHDS) to 2017 values using the IMF's measure of inflation for consumer prices in India[6] and then converted to USD using the INR-USD 2017 average conversion rate[7]. To protect the confidentiality of edX's financial assistance policies, I am unable to reveal the exact location of the income eligibility threshold, but the similarity between the two lines provides evidence that Indian applicants are not systematically manipulating their incomes to gain eligibility for financial assistance.

Figure 1.6 shows the income distribution by gender, Figure 1.7 shows the distri-

[6]https://www.imf.org/external/pubs/ft/weo/2019/01/weodata/weoselgr.aspx

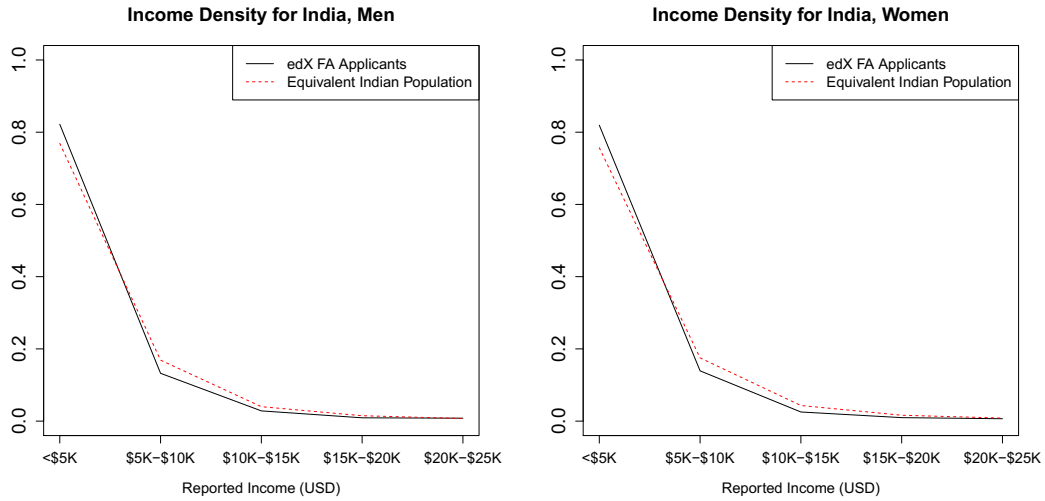
[7]https://www.irs.gov/individuals/international-taxpayers/yearly-average-currency-exchange-rates

Figure 1.5: Income density for India, edX applicants and the general population with at least a secondary education



The black line shows the distribution of self-reported income for applicants from India. The dashed red line shows the distribution of income for everyone in India with at least a secondary education, inflation adjusted from 2012 values to 2017 values. The similarity between the two lines implies applicants did not systematically manipulate their reported incomes.

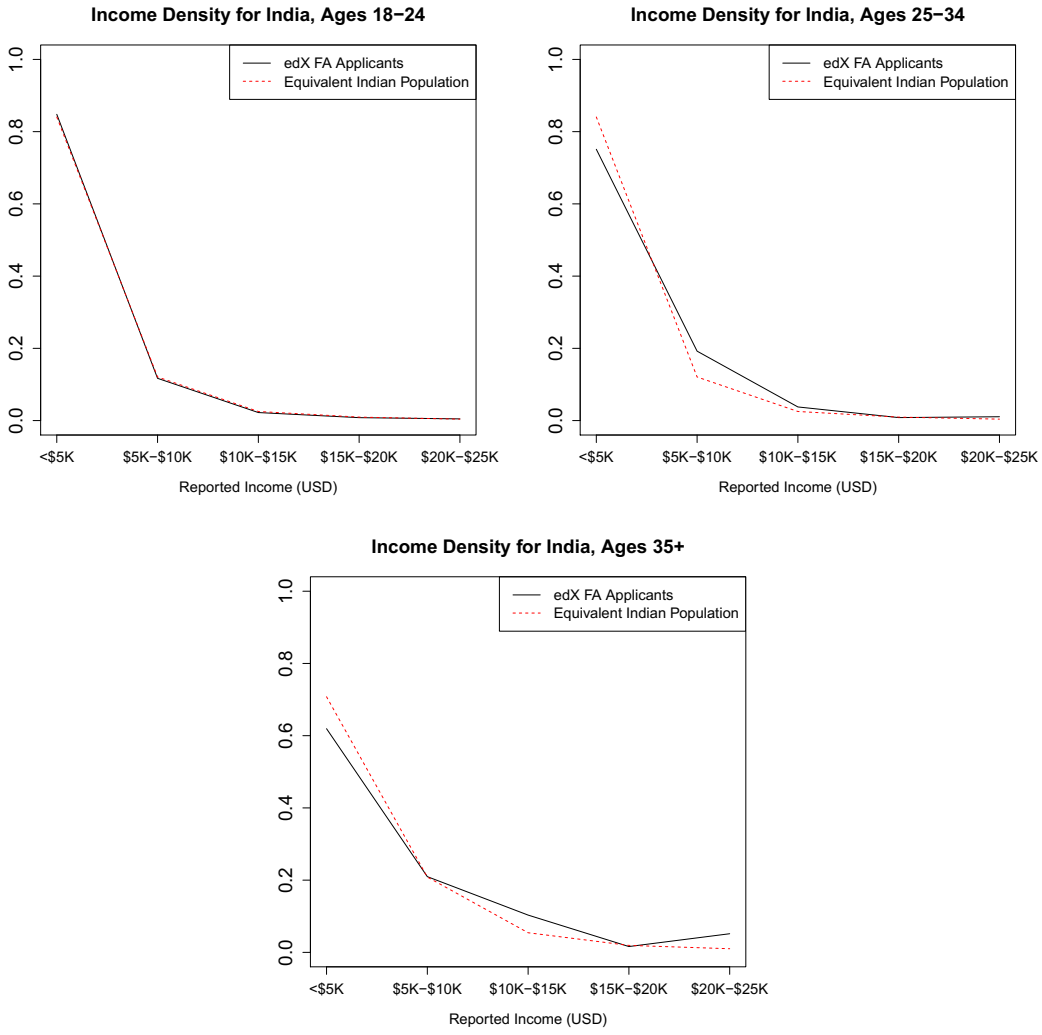
Figure 1.6: Income density for India by gender



For India, the income distribution by gender within the applicant pool closely mirrors that of the corresponding general population (e.g. who completed secondary school or higher).

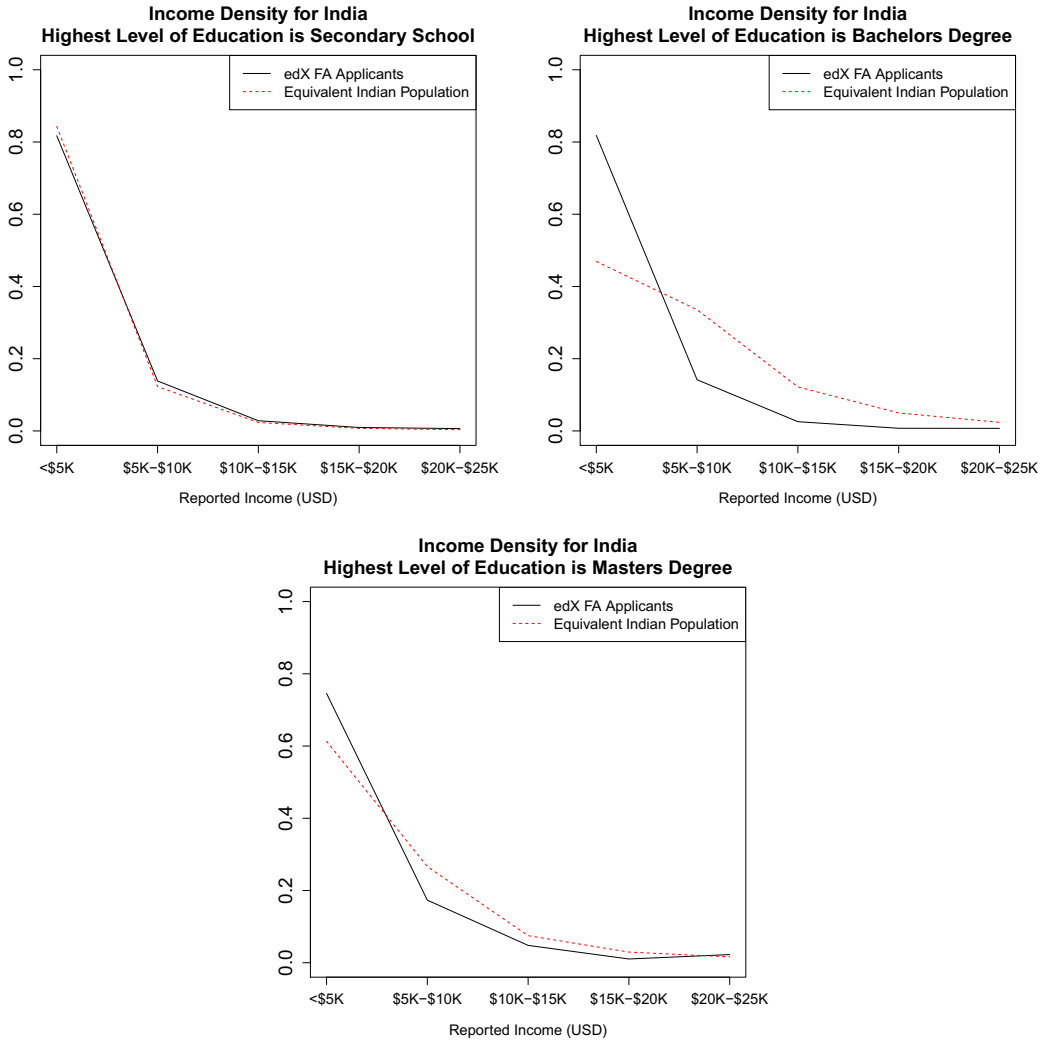
bution by age, and Figure 1.8 shows the distribution by self-reported highest level of education. I find that the applicant income distributions closely align with the population income distributions for both genders, all age groups, and secondary school graduates. However, the income distribution for college graduates differs significantly from the general population and the distribution for applicants with a masters degree also differs noticeably, though the difference is less pronounced than for college graduates. In both cases, the proportion of applicants who fall in the lowest income bucket exceeds the population equivalent. While these differences may represent income manipulation to gain eligibility, it is also possible that highly educated online learners are disproportionately unemployed or underemployed and are thus using edX to improve their employment prospects. For

Figure 1.7: Income density for India by age



For India, the income distribution by age within the applicant pool closely mirrors that of the general Indian population with at least a secondary education.

Figure 1.8: Income density for India by highest level of education



For India, the income distribution for the applicant pool and general population is nearly identical for people with a secondary education but very different for college graduates. A plausible explanation is that college educated applicants are disproportionately unemployed or underemployed and are using online education to improve their employment prospects. However, the difference could represent income manipulation and therefore I estimate effects with and without this population. When estimating the effect without these subgroups, I find a larger effect, suggesting that if any income manipulation is happening in these groups, it is downward biasing estimates rather than inflating them.

robustness, I run the main analyses of the paper with and without these subgroups and find that the effect sizes increase when these two groups are excluded, so income manipulation in these groups is not driving the large and significant estimated effect sizes.

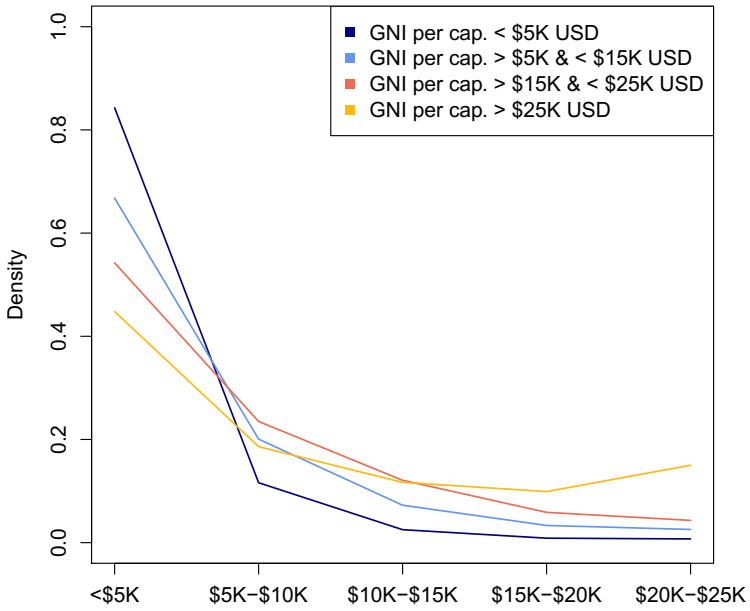
1.4.4 All Countries Other than India and the US

The remaining 190 countries in the sample have a wide range of economic situations. For any countries with economic situations similar to that of India, the standard McCrary test and covariate RD specifications will not be informative about income manipulation, due to the extreme discreteness of the running variable relative to the income distributions in such countries. However, for countries with economic situations more similar to that of the US, the standard McCrary test and covariate RD specifications should be valid. Therefore, I first quantify the extent to which reported income distributions vary smoothly with 2017 GNI per capita[8] and then determine the subset of countries on which to run the standard McCrary and covariate RD tests.

As can be seen from Figure 1.9, I find that applicant income distributions change smoothly as a function of country GNI per capita and that the exponential decline in income persists until country GNI per capita exceeds $15K USD. Given these findings, I run the McCrary test and covariate RD specifications for the subset of countries whose 2017 GNI per capita exceeded $15,000 USD. These countries represent 24% of all applicant countries and 20% of all applicants, and I report the

[8]https://data.worldbank.org/indicator/ny.gnp.pcap.cd?page=1

Figure 1.9: Income density by country GNI per capita



The distribution of applicant incomes is shown for applicants from countries defined by four distinct GNI per capita levels. The applicant income distribution smoothly changes from resembling India's income distribution for low-income countries to resembling the US's income distribution for high-income countries.

main results for this set of countries in Sections 1.5.1 and 1.5.2. Using the base specification in Equation 1.1 and replacing y_i with indicators for gender, age, and education, I find that only one of the covariates exhibits a statistically significant jump at the threshold. Namely, the indicator for whether a learner is over the age of 35 increases a statistically significant 8% across the threshold. The full set of threshold coefficients and corresponding standard errors for each covariate specification are shown in Table 1.3.

Table 1.3: Tests of covariate balance across the threshold for countries with GNI per capita greater than $15,000 USD

	$\hat{\beta}_2$ Coefficient	SE
Gender		
Male	-0.03558	0.02963
Female	0.02286	0.02721
Age		
18-24	-0.03419	0.02764
25-34	-0.03511	0.0325
35+	0.07733	0.02982
Highest Level of Education		
High School	-0.01109	0.02235
Bachelor's	0.03646	0.02737
Master's	0.00512	0.02422

This table shows the coefficient estimate for β_2 from Equation 1.1, where y_i has been replaced by indicators for various covariates to test whether covariate densities change discontinuously across the threshold. The sample is restricted to all applicants with 2017 GNI per capita greater than $15,000 USD, as Figure 1.9 shows that the income distribution is exponentially declining for lower income countries, likely reflecting the true income distribution in those countries, similar to what was shown for India.

1.5 Main Results

In this section, I present the estimated effects of financial assistance on certificate rates, pass rates, and persistence and also estimate the distribution of willingness-to-pay for verified certificates. I find that financial assistance has a large impact on learner outcomes that does not differ much between high-income and low-income countries. Additionally, I find that applicants are most likely to exert effort to pass a course when they are confident they will be able to obtain a verified certificate, and I show that applicants who have already exerted significant effort unsurprisingly value the certificate more than the general applicant population. I also discuss some gender disparities between countries and demonstrate that much work remains to be done in order to achieve gender parity on the platform.

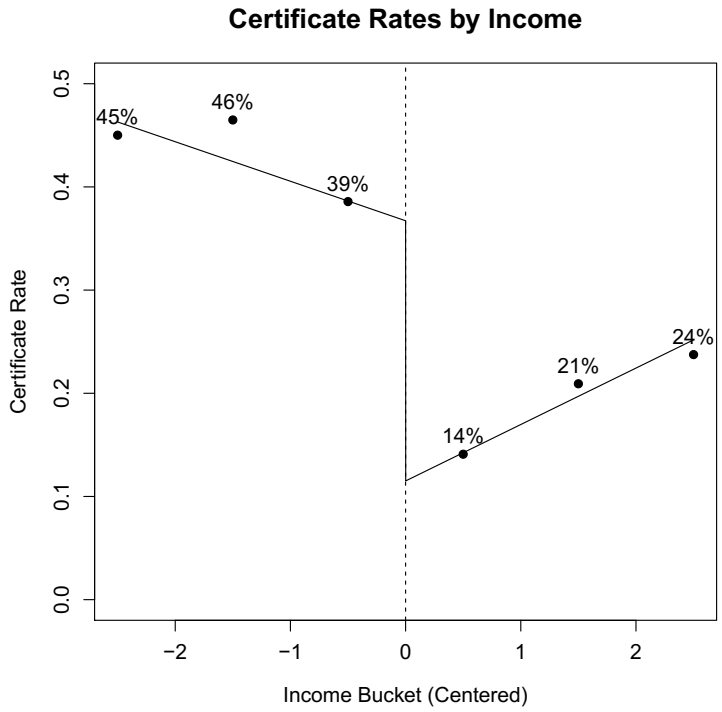
1.5.1 Certificate Rates

Financial assistance had a large positive impact on certificate rates, as is evident from Figure 1.10. In particular, it increased certificate rates by roughly 25 percentage points, from 14% to 39%. The RDD estimate for the pooled sample (all countries) is 25.2pp with a standard error of 1.1pp, while the Pearson's chi-squared approach yields an estimated effect of 24.5pp with a corresponding 95% confidence interval spanning [23.3pp, 25.6pp]. Both estimates are very large and highly statistically significant. Additionally, when I run the RDD specification with covariates from Equation 1.2 for the pooled sample, I find the estimates are very similar to those found from the specification without covariates. The similar-

ity in estimated effects with and without covariates provides additional evidence that results are not driven by income manipulation.

The slopes of the regression lines from Figure 1.10 reflect one of the central challenges of implementing RDD in the current context. In particular, the point just to the left of the threshold represents all countries, but the poorest countries are not reflected in points further to the left. Since financial assistance recipients from the wealthiest countries certify at higher rates than financial assistance recipients from the poorest countries, the downward slope to the left of the threshold can be attributed to changing country composition across the points. For the set of countries with thresholds at or above $15,000 USD (which have at least three points to the left of the threshold), the slope to the left of the threshold is essentially flat. On the right hand side, the upward slope is not affected by changing country composition, since certificate rates just to the right of the threshold are consistently in the neighborhood of 14%. Instead, the upward slope is driven by the fact that only countries with very low GNI per capita have thresholds low enough to observe three points to the right of the threshold, and in those countries, people whose income falls in the upper income buckets are significantly wealthier than people whose income falls in the lowest income buckets. For example, in India, less than 8% of the population with at least a secondary education has an income over $10,000 USD and less than 2% of the same population has an income over $20,000 USD. Therefore, people in the highest several income buckets are significantly wealthier than people in the lowest income bucket, and so it is unsurprising that their outcomes differ substantially. Although these factors

Figure 1.10: Learner certificate rates by income bucket with the RDD fitted line



The RDD for the effect of financial assistance on certificate rates for a pooled sample that includes all countries. The income eligibility threshold is shown as a dashed vertical line at zero.

influence the slopes, and thus the RDD estimated effect sizes, the actual jump in outcomes at the threshold is fairly stable across countries, so I use the Pearson's chi-squared test for robustness. Regardless of the specification or sub-population, I find that financial assistance has a large and statistically significant effect on certificate rates.

The effect sizes for India and the United States (US) are both somewhat larger than the pooled estimates. In particular, the RDD estimate for India is 33.7pp with a standard error of 2.1pp, and the Pearson's chi-squared estimate is 29.8pp

Table 1.4: Estimated effects of financial assistance on learner certificate rates.

	RDD Est.	RDD SE	χ^2 Est	χ^2 95% CI
All Countries	25.2	1.1	24.5	[23.3, 25.6]
All Countries + Covariates	25.6	1.2	-	-
India	33.7	2.1	29.8	[27.9, 31.6]
USA	29.9	3.2	28.5	[22.6, 34.4]
High GNI PC Countries	34.7	2.5	30.0	[26.6, 33.3]

All numbers are measured in percentage points. I run the specification for all countries with and without covariates and find no meaningful difference in the RDD estimate. High GNI PC Countries represent countries with 2017 GNI per capita $\geq$ \$15,000 USD.

with a corresponding 95% confidence interval covering [27.9pp, 31.6pp]. For the US, the RDD estimate is 29.9pp with a standard error of 3.2pp, and the Pearson's chi-squared estimate is 28.5pp with a 95% confidence interval of [22.6pp, 34.4pp]. For all countries with GNI per capita greater than \$15,000 USD, the estimates are somewhat larger than for the US alone. The results are presented in Table 1.4.

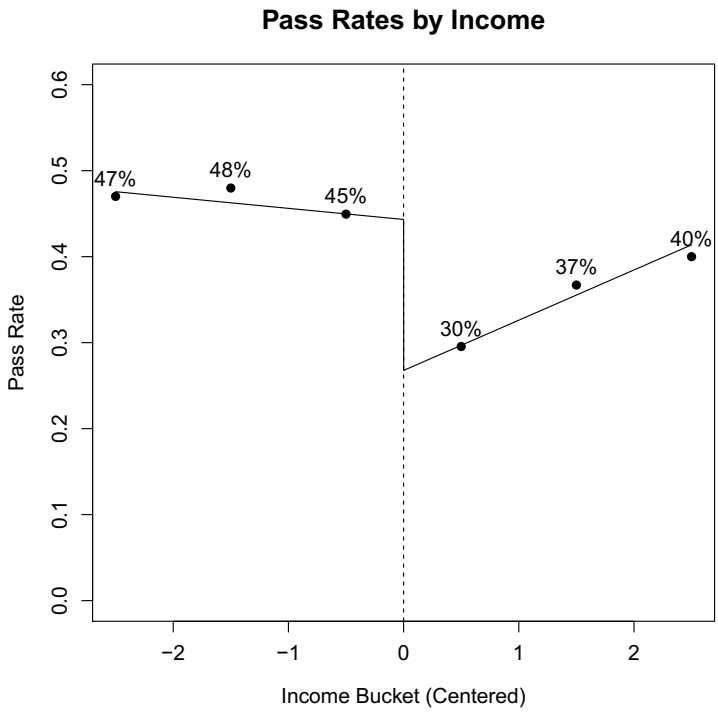
I find no statistically significant differences between India and the US in terms of certificate rates for applicants near the threshold nor in terms of effect sizes, suggesting that applicants from high-income countries do not outperform applicants from low-income countries. However, when I broaden the scope to include more countries, I find that certificate rates are slightly higher for financial assistance recipients in wealthier countries relative to financial assistance recipients in poorer countries, though there is no statistically significant difference for learners above the threshold.

1.5.2 Pass Rates

There was no monetary cost to take and pass a course during the period under study. Nevertheless, financial assistance significantly increased applicant pass rates, raising them by roughly 15 percentage points, from 30% to 45%, as can be seen in Figure 1.11. The RDD estimate for the effect of financial assistance on pass rates is 17.6pp with a standard error of 1.2pp, while the Pearson's chi-squared estimate is 15.4pp with a corresponding 95% confidence interval of [14.0pp, 16.8pp]. When I run the RDD specification with covariates according to Equation 1.2 on the pooled sample, the estimates are again very similar to the results obtained from the specification without covariates. The estimated effects of financial assistance on pass rates are presented in Table 1.5.

For India and the US, the estimated effects of financial assistance on pass rates are similar to but slightly larger than the pooled estimate for all countries, mirroring the trend observed for certificates. In particular, the RDD estimate for India is 25.8pp with a standard error of 2.2pp and the Pearson's chi-squared estimate is 21.9pp with a 95% confidence interval of [19.5pp, 24.3pp]. For the US, the RDD estimate is 19.4pp with a standard error of 3.3pp and the Pearson's chi-squared estimate is 18.8pp with a corresponding 95% confidence interval covering [12.5pp, 25.0pp]. For countries with GNI per capita above $15,000, the effect sizes are somewhat larger than for the US alone. Again, I find no statistically significant differences between India and the US in terms of pass rates or effect sizes, though the point estimates suggest applicants from India may pass courses at slightly

Figure 1.11: Learner pass rates by income bucket with the RDD fitted line



The RDD for the effect of financial assistance on pass rates for a pooled sample that includes all countries. The income eligibility threshold is shown as a dashed vertical line at zero.

Table 1.5: Estimated effects of financial assistance on learner pass rates

	RDD Est.	RDD SE	χ^2 Est	χ^2 95% CI
All Countries	17.6	1.2	15.4	[14.0, 16.8]
All Countries + Covariates	18.2	1.3	-	-
India	25.8	2.2	21.9	[19.5, 24.3]
USA	19.4	3.3	18.8	[12.5, 25.0]
High GNI PC Countries	25.2	2.7	19.5	[15.8, 23.2]

The estimated effects of financial assistance on learner pass rates. All numbers are measured in percentage points. High GNI PC Countries represent countries with GNI per capita $\geq$ \$15,000 USD.

higher rates, regardless of whether they receive financial assistance.

1.5.3 Persistence and Retention

Financial assistance has been shown to increase persistence in traditional educational settings (Bettinger, 2004). However, since this is the first paper to study the effects of financial assistance on online learners, it is unknown whether the effects on persistence are similar for online learners. To measure the effects, I restrict my sample to learners who submitted their first application for financial assistance at least one year before the data ends, a restriction that ensures I observe one-year outcomes. I then measure whether each learner enrolled in, passed, or earned a certificate in at least one additional course within one year of first applying for financial assistance. The effects are presented in Table 1.6. Overall, the effects on one-year outcomes are smaller than the main effects presented in Sections 1.5.1 and 1.5.2.

Roughly 75% of first-time applicants enroll in a second course (or more) within one year, and the rate is slightly higher for those who received financial assistance.

Table 1.6: Estimated effects of financial assistance on 1-year outcomes

	RDD Est.	RDD SE	χ^2 Est	χ^2 95% CI
1-year Enrollment Rates	3.8	1.5	3.3	[1.2, 5.4]
1-year Pass Rates	5.0	5.3	9.9	[7.6, 12.1]
1-year Certificate Rates	8.2	3.5	12.8	[10.9, 14.7]
1-year Reapplication Rates	18.7	1.5	21.7	[20.2, 23.2]

The estimated effects of financial assistance on learner certificate rates. All numbers are measured in percentage points.

The RDD estimate for the effect of financial assistance on one year enrollment rates is 3.8pp with a standard error of 1.5pp, and the Pearson's chi-squared estimate is 3.3pp with a corresponding 95% confidence interval spanning [1.2pp, 5.4pp].

One-year pass rates are also higher for first-time applicants who received financial assistance, increasing from 37.4% just above the threshold to 47.3% just below the threshold. The Pearson's chi-squared estimate for the effect of first-time financial assistance on one-year pass rates is thus 9.9pp and the corresponding 95% confidence interval is [7.6pp, 12.1pp]. However, the RDD estimate is 5.0pp with a standard error of 5.3pp, which is not statistically significant. The difference between the two estimates is driven by a steep negative slope to the left of the threshold in the RDD specification.

As with the main results, the effect of first-time financial assistance on one-year certificate rates is larger than its effect on one-year pass rates. In particular, the RDD estimate is 8.2pp with a standard error of 3.5pp, and the Pearson's chi-squared estimate is 12.8pp with a 95% confidence interval spanning [10.9pp, 14.7pp]. Again, the smaller RDD estimate is driven by a steep negative slope to the left of the threshold.

1.5.4 Reapplication Rates

Perhaps unsurprisingly, reapplication rates differ substantially for learners whose initial application was approved versus denied. About 30% of first-time applicants whose incomes fall just below the eligibility threshold re-apply for financial assistance within a year compared to only 8% of applicants whose incomes fall just above the eligibility threshold. The RDD estimate for the effect of first-time financial assistance on one-year reapplication rates is 18.7pp with a standard error of 1.5pp, and the Pearson chi-squared estimate is 21.7pp with a 95% confidence interval of [20.2pp, 23.2pp]. Given the large effect of initial receipt on reapplication rates, it seems likely that at least some of the effect of financial assistance on persistence is driven by edX continuing to financially support learners.

One concern with estimating causal effects of financial assistance on future outcomes is that if applicants whose applications are denied gain information about the location of the threshold, they may reapply with a lower income. This would lead to applicants on both sides of the initial threshold subsequently receiving financial assistance and may downward bias estimated differences.

I find evidence that it is fairly common for applicants who re-apply after being denied to revise their reported income down to a level that qualifies them for financial assistance. 3.6% of all applicants whose first application was denied re-apply with an eligible income within one year, which translates to 44% of the 8.4% of applicants whose first application is denied and who subsequently reapply. Therefore, the smaller effects seen on subsequent course performance

relative to first course performance may be driven in part by applicants who were originally denied financial assistance subsequently receiving it. This evidence provides further support for focusing the primary analysis on first applications only, in order to avoid the bias introduced by applicants learning the threshold's location and manipulating their behavior accordingly. However, it is worth noting that downward changes in reported income are also common among applicants whose first application was accepted and who therefore have nothing to gain from revising their incomes downward. Among first-time applicants whose applications were accepted and whose first reported income was greater than $5,000 (thus making downward revision possible), 12.7% reported a lower income within one year, which accounts for 34.2% of all re-applicants in this group.

1.5.5 The Signaling Value of edX Certificates

During the time period covered by my analysis, learners could access all course features from the audit track, e.g. without paying to enroll in the verified track. Therefore, all learners who passed a course gained the same skills, regardless of whether they purchased a certificate, and thus the value of the certificate was purely as a signal.

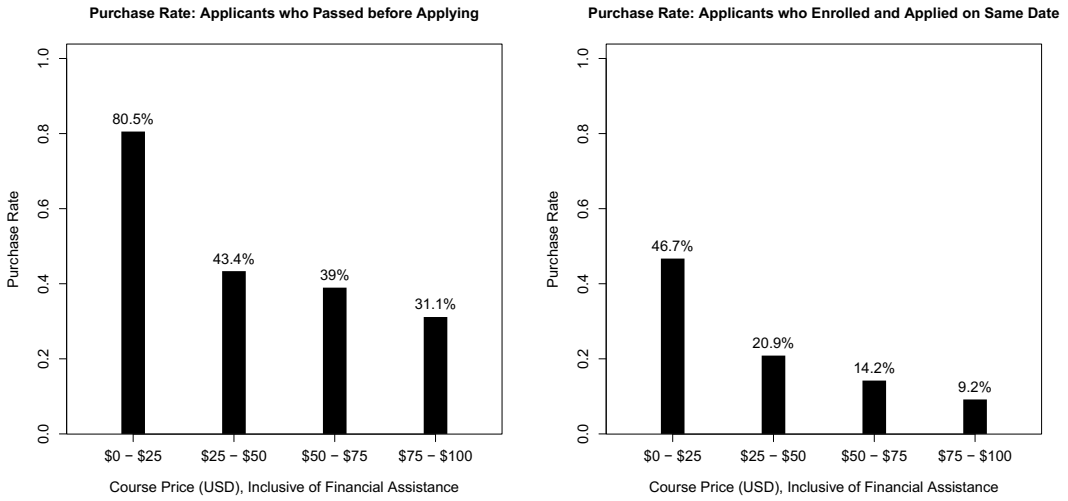
Using observed prices and purchases, I estimate the distribution of applicants' willingness-to-pay for certificates, which is informative about the distribution of the certificate's signaling value for applicants. It is worth noting that the distribution for applicants is not necessarily representative of the distribution for the general population of online learners. To estimate the distribution, I take all

applicants facing a price p, where $\$L \leq p \leq \$U$, and calculate the percent of applicants who purchased a verified enrollment. This yields a curve where each point represents the proportion of applicants willing to pay at least a price between $\$L$ and $\$U$ (or higher) for a verified certificate. However, the analysis is complicated by the fact that verified certificates require not only a monetary cost but also a substantial effort cost. To better understand the cost of effort involved in taking a course, I exploit the fact that applicants fall into one of three groups: "always takers" who purchase a verified certificate whether or not they receive financial assistance, "never takers" who do not purchase a certificate whether or not they receive financial assistance, and "compliers" who purchase a certificate if and only if they receive financial assistance. I assume there are no applicants who would purchase a certificate if they were denied financial assistance but not if they were offered financial assistance, e.g. no "defiers."

Within the group of compliers, some learners pass courses regardless of whether they receive financial assistance, while others are induced to pass by receiving financial assistance. The compliers who only pass when they receive financial assistance account for 15.4% of all applicants, while the compliers who pass irrespective of financial assistance account for 9.1% of all applicants. These subpopulations explain why pass rates change 15.4% across the threshold, while certificate rates change by an additional 9.1% or 24.5% in total. The two complier subgroups presumably differ in their costs of effort and in the value they place on accumulating knowledge for its own sake rather than for signaling.

That the cost of effort differs by learner complicates the analysis of signaling

Figure 1.12: Applicant purchase rates for verified certificates by course price inclusive of financial assistance



(Left) WTP distribution for applicants who had already passed the course at the time of applying for financial assistance. (Right) WTP distribution for applicants who enrolled in a course and applied for financial assistance on the same day. The left-hand plot is informative about the signaling value of the certificate for people who had already exerted sufficient effort, while the right-hand plot is informative about the ex-ante value of the certificate, accounting for the disutility of effort required to pass a course.

value. To ensure clean comparisons, I thus examine the distribution of willingness-to-pay in two very well-defined applicant subgroups: those who passed the course prior to applying for financial assistance and those who applied for financial assistance at the same time as enrolling in the course. Applicants in the first group have no further effort required while applicants in the second group have yet to put in any significant effort. Therefore, the first group sheds light on the signaling value of certificates for applicants who were able to successfully pass a course, while the second group provides information about the certificate's ex ante value. The results are presented in Figure 1.12.

It is clear from Figure 1.12 that applicants who have already exerted significant effort value the certificate more than applicants who have not yet exerted any effort. This reflects the fact that certificates require not only a monetary cost but a significant cost of effort. Further evidence of the effort cost comes from the fact that less than half of all applicants pass the course for which they first applied for financial assistance, regardless of whether their application was approved. Among applicants who enroll in the verified track, the pass rate is much higher but still below 100%, which shows that even learners who value the certificate enough to pay its monetary costs may find the costs of required effort prohibitive.

1.5.6 Within Learner Effects from the 5-Course Threshold

In addition to the main income eligibility threshold, another threshold is encountered by learners who apply for financial assistance for more than five courses in a 12-month period. In particular, learners are only eligible to receive financial assistance for up to five courses per 12-month period, so any applications beyond the limit are denied. Although this restriction is publicly available, it is not posted on the application nor included in the email that informs learners that their application was denied. Therefore, it seems possible that some applicants were unaware of this rule. Only 1,242 out of 35,341 learners in the sample applied for financial assistance for more than five courses within 12 months of their first application. Since it would be reasonable to assume that these learners are fundamentally different from other learners in the sample, I use learner fixed-effects to control for differences in learner behavior.

Table 1.7: Estimated effects of financial assistance on learner outcomes for learners affected by the five course cutoff

	RDD Est.	RDD SE	χ^2 Est	χ^2 95% CI
Certificate Rates	20.0	1.9	21.9	[18.2, 25.6]
Pass Rates	9.5	1.9	12.4	[8.5, 16.3]

The estimated effects of financial assistance on learner certificate rates and pass rates for the subset of learners who applied for assistance for more than five courses in the 12-month period following their first application. All numbers are measured in percentage points.

With this specification, I find similar, though somewhat smaller, effects of financial assistance on course certificate rates and pass rates relative to estimates obtained from my main analyses in Sections 1.5.1 and 1.5.2. The results are presented in Table 1.7. The RDD specification estimates that financial assistance increases certificate rates for this population by 20.0pp with a standard error of 1.9pp, while comparing certificate rates across the threshold yields an estimate of 21.9pp with a 95% confidence interval of [18.3pp, 25.5pp]. It is worth noting that all learners in this sub-sample applied for financial assistance for at least six courses in a 12-month period. Therefore, when comparing outcomes across the threshold, I am comparing outcomes for the same group of learners with and without financial assistance.

As in the other analyses presented thus far, the effect of financial assistance on pass rates for this sub-population is smaller than the effect of financial assistance on certificate rates. In particular, RDD estimates suggest financial assistance increases course pass rates for this population by 9.5pp with a standard error of 1.9pp, and the Pearson's chi-squared approach yields an estimate of 12.4pp with a 95% confidence interval spanning [8.6pp, 16.2pp]. Though slightly smaller, the

results are surprisingly similar to the main results, indicating that effect sizes are robust across different specifications and sub-populations.

1.5.7 Gender Differences

Although I found no significant differences between India and the US with regards to pass rates, certificates rates, or effect sizes, two significant differences between the countries emerge when looking at gender. First, 33% of all US applicants are female compared to only 13% of all Indian applicants, a difference that is highly statistically significant[9]. Second, among applicants in India who receive financial assistance, women pass courses and certify at higher rates than men, differences that are statistically significant at the 95% confidence level. By contrast, in the US, point estimates suggest that women throughout the income distribution pass and certify at lower rates than men, though differences are only statistically significant when considering pass rates for applicants over the income eligibility threshold.

When considering applicants from all countries, men and women show largely similar performance, though men seem to pass and certify at slightly higher rates than women. Additionally, there are more than three times as many male applicants as female applicants. These results are shown in Table 1.8 and suggest that more work is needed to achieve gender parity on the platform, but that women in developing countries may be well served by educational options available online,

[9]When restricting to learners who self-reported their gender as male or female (e.g. excluding those whose gender is unknown and who report genders as other) then the numbers rise to 40% and 16%, still statistically significantly different

Table 1.8: Summary statistics by gender, including share of applicants and course performance measures

	% of Applicants	Cert. Rates with FA	Cert. Rates no FA	Pass Rates with FA	Pass Rates no FA
All Countries					
Women	19.8	38.5	15.3	45.1	28.9
Men	65.3	40.2	17.8	45.9	34.3
Diff. CI	[-46.1, -44.8]	[-3.1, -0.2]	[-4.7, -0.3]	[-2.4, 0.6]	[-8.1, -2.7]
India					
Women	13.0	44.0	12.3	51.8	28.0
Men	68.9	40.4	14.8	46.7	29.4
Diff. CI	[-56.6, -54.7]	[0.7, 6.5]	[-6.6, 1.7]	[2.3, 8.1]	[-7, 4.1]
USA					
Women	32.8	37.1	12.8	38.6	18.7
Men	48.9	43.3	14.5	45.5	30.1
Diff. CI	[-18.4, -13.7]	[-10.4, -2]	[-7.5, 4.1]	[-11.1, -2.7]	[-18.5, -4.3]

Summary statistics by gender show that overall men apply in significantly greater numbers than women and tend to outperform women in terms of pass rates and certificate rates. However, women in India outperform men when they receive financial assistance, suggesting financial assistance may be especially helpful for women in developing countries.
Note: Diff. CI is the 95% confidence interval for the difference between women and men. A negative value indicates the rate for women is less than the rate for men.

especially when they receive financial assistance.

1.6 Conclusion

This paper is the first to examine the impacts of financial assistance on online learner outcomes. Overall, I find that financial assistance is highly effective, significantly increasing both certificate rates and pass rates, even though there was no monetary cost to pass a course during the time period under study. Additionally, I find that learners value the signal provided by verified certificates and experience a significant cost of effort when trying to pass a course. Financial assistance applicants are a highly engaged and successful group of low-income learners, who disproportionately live in countries with low and medium HDI ratings, where opportunities for health, education, and a high standard of living are limited. Although much work still remains before online education platforms will attain their goals of democratizing education, financial assistance appears to be a promising step in the right direction.

Chapter 2

Nowcasting Trends in the US Housing Market

2.1 Introduction

In the United States, the housing market is a fundamental part of the economy. Its central role in the macroeconomy is evident from its significant contribution to the Great Recession of 2008. It also affects most Americans, whether homeowners or renters, on a more personal level. While homeowners have much of their wealth tied up in their home and thus benefit from upward trends in the housing market, renters are subject to the rising rents brought about by such trends. Rapidly rising rents can lead to displacement, especially for low income renters, and are thus a major subject of concern for policy makers in the many US cities where

prices have steadily risen over recent years.

This paper considers the multiple policy questions raised by better understanding current trends in the US housing market at a range of aggregations, from the neighborhood to the national level. The housing market is characterized by significant lags, introduced both by the delay between offer agreement and closing in residential transactions, which averages roughly two months[1], as well as lags in reporting and aggregation of closed transactions. Therefore, the primary focus of this paper is improving nowcast models of the housing market, which I do by leveraging highly successful models from other domains and novel aggregations of clickstream data.

In particular, I adapt a model from epidemiology, which outperformed similar models at nowcasting flu trends. I build on the flu-tracking work, because the flu models were stress-tested and subject to wide scrutiny, resulting in the high-profile discontinuation of Google Flu Trends. Thus, I expect the best-performing model from the flu-tracking literature to be highly robust to a range of conditions. A secondary question of this paper, therefore, is the extent to which a tried and true epidemiological model can be applied to economic domains to address some of the challenges that have recently surfaced in the economic nowcasting literature with regards to mixed success of nowcasting models (Richardson, 2019).

I find that the flu-tracking model, known as Auto-Regressive with GOogle search as exogenous variables (ARGO), extends easily to economic domains, and I provide detailed evidence for the housing market as well as preliminary evidence for

[1]https://www.realtor.com/advice/buy/how-long-does-it-take-to-close-on-a-house/

additional macroeconomic indicators, including unemployment and GDP. Additionally, I find that the model works well using aggregated Zillow clickstream data in place of Google search volumes, suggesting that organizations would be well served to track clickstream activity and incorporate it in their forecasts using a regularized model, such as ARGO.

The rest of the paper is organized as follows. Section 2.2 reviews the relevant literature, Section 2.3 introduces the model, and Section 2.4 describes the data used for the analyses. Section 2.5 presents results at the state and zip code levels, showing that ARGO adds value at both levels of aggregation as well as when used with Zillow clickstream activity in place of Google search data. Section 2.6 discusses possible extensions and provides evidence that ARGO would have performed reasonably well at nowcasting macroeconomic indicators through the last recession. Finally, Section 2.7 concludes.

2.2 Literature Review

The housing market has evolved significantly over the past 40 years in which it has been an active area of study. Before the era of the Internet, information about homes for sale was found primarily through real estate agents and newspaper ads (Smith & Clark, 1982). This led to significant asymmetries in information as well as high search costs (Anglin, 1997). Utility-maximizing buyers facing these high search costs could be described as satisficing, which is to say purchasing an acceptable home because discovering the truly optimal one was intractable (Smith

et al., 1979). With the introduction of websites like `zillow.com`, `realtor.com`, `trulia.com`, etc., where anyone with Internet access can view the full inventory of properties for sale, search costs and information asymmetries have significantly decreased, though not entirely disappeared.

Recent work on housing search proposes a two-stage model, where prospective buyers first narrow the universe of potential choices to create a choice set and then select one choice among the set (Rashidi et al., 2012). Empirical work by Rae (2015) supports this general framework, finding that shoppers restrict their search to small user-defined geographies, which he calls submarkets. Rae uses data from `rightmove.co.uk`, a UK real estate website similar to Zillow and Trulia. Rae's use of Rightmove search data is especially valuable in showing the geographic distribution of demand, which may not be reflected by housing sales transactions, due to factors such as supply-side constraints. Rae addresses the potential issue of noise in Internet browsing data by restricting his sample to logged-in users. Kelly & Teevan (2003) find that dwell time, defined as the amount of time a user spends on a webpage, is an effective measure of user preferences, so dwell time may offer another mechanism by which to effectively filter out noise.

Despite the potential for noise, many sources of Internet data have demonstrated significant value for understanding and predicting housing market dynamics. Glaeser et al. (2018) use Yelp data to nowcast gentrification. They find that changes in neighborhood business composition, and in particular the introduction of Starbucks coffee shops, can indicate gentrification. Bailey et al. (2018) use data from Facebook and Zillow to show that people's housing market outlooks

are influenced by the housing market experiences of their geographically distant Facebook friends.

Google search data in its aggregated form, Google Trends (GT), is by far the most popular Internet search data used to augment nowcast and forecast models. Wu & Brynjolfsson (2015) show that adding Google Trends data to a simple autoregressive model can improve forecast and nowcast models for US housing markets, though they find some heterogeneity in the value of search data across US states. They find that search data improves predictions of housing sales volumes more than housing price indices, a difference they attribute to the fact that both sellers and buyers search for real estate agents and thus the queries they use are not reflective of supply versus demand. Askitas (2016) also uses Google search data but separates queries related to buying versus selling. He constructs an index of relative supply and demand, which he uses to nowcast the S&P / Case-Shiller U.S. National Home Price Index. Beracha & Wintoki (2013) also use Google search data to predict home prices and find explanatory power, especially for "hot" real estate markets and those with low supply land elasticity. The fact that Google search data was especially useful in predicting abnormal upticks and price trends in land-constrained coastal markets is encouraging, since such markets have proven difficult to explain through formal economic models, such as in Glaeser et al. (2014).

In addition to its value in improving housing market models, Google Trends has also proven effective at improving nowcast and forecast models in a wide range of applications, including disease detection (Yang et al., 2017, 2015), macroeco-

nomic indicators (Giannone et al., 2008), consumer sentiment (Choi & Varian, 2012), and more. In particular, Yang et al. (2015) incorporate contemporaneous Google Trends data in a classical time series model to nowcast case counts of the flu. They use a penalty parameter to allow the model to automatically select the most informative collection of lagged flu counts and contemporaneous search volumes in order to provide an accurate and real-time flu prediction model that they nickname ARGO (AutoRegressive model with GOogle search queries as exogenous variables). In this paper, I build upon the ARGO foundation and extend it to incorporate contemporaneous search activity on `zillow.com`.

2.3 Model

In its original implementation, ARGO was used to predict influenza counts in real time (Yang et al., 2015). In that context, the key assumption was that when more people are affected by the flu (either because they have it or because someone they know has it), influenza-related search volumes increase. The authors formalize their assumption using a Hidden Markov Model, where current influenza case counts are dependent on prior influenza counts and correlated with current influenza-related search volumes. In the context of the housing market, I draw parallel assumptions. Specifically, I assume that current market conditions are dependent on prior market conditions and correlated with current housing-related Internet activity, including search volumes and for-sale listings.

To formalize the model, suppose y_t is the seasonally adjusted median sale price

of all homes sold (in a given geographic region) in period t. Note that y_t could equivalently be another outcome of interest, such as the volume of homes sold in period t. Let $\boldsymbol{X}_t$ be an $S \times 1$ vector containing signals about the housing market, such as search volumes by agents and consumers as well as summary information about current for-sale listings. In practice, since there is an average lag of two months between accepted purchase offers and completed transactions, I use search data from 2 months before the sale is recorded. This reflects the formal assumption that search activity and listings are relevant at the time of offer rather than transaction. For quarterly data, I use search data from both the current and prior quarter, since neither one alone sufficiently accounts for the two-month delay.

I use an autoregressive model of lag N, which implies a Markov chain, where each state is a collection of vectors $\{y_{(t-N+1):t}\}_{t \geq N}$. I further assume a Hidden Markov Model, where the Markov chain just described emits signals in each period captured by $\boldsymbol{X}_t$. Intuitively, the model assumes there is predictive value in contemporaneous search and listing data as well as in lagged outcomes up to N lags, but that outcomes older than N lags provide no additional information. The Hidden Markov Model (HMM) is formally represented as follows:

$$\begin{array}{ccccccc} y_{1:N} & \rightarrow & y_{2:(N+1)} & \rightarrow & \cdots & \rightarrow & y_{(T-N+1):T} \\ \downarrow & & \downarrow & & & & \downarrow \\ \boldsymbol{X}_N & & \boldsymbol{X}_{N+1} & & & & \boldsymbol{X}_T \end{array}$$

The formal mathematical assumptions, which map directly to the original ARGO model in Yang et al. (2015), are stated below.

Assumption 1 $y_t = \mu_y + \sum_{j=1}^{N} \tilde{\alpha}_j y_{t-j} + \epsilon_t, \qquad \epsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

Assumption 2 $\boldsymbol{X}_t | y_t \sim \mathcal{N}_S(\boldsymbol{\mu}_x + y_t \tilde{\boldsymbol{\beta}}, \boldsymbol{Q})$

Assumption 3 Conditional on y_t, $\boldsymbol{X}_t$ is independent of $\{y_r, \boldsymbol{X}_r : r \neq t\}$

Where $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, ..., \tilde{\beta}_S)^\top$, $\boldsymbol{X}_t = (x_{1,t}, ..., x_{S,t})$, $\boldsymbol{\mu}_x = (\mu_{x_1}, ..., \mu_{x_S})^\top$, and $\boldsymbol{Q}$ is the covariance matrix. The assumptions imply that the predictive distribution $f(y_t | y_{1:(t-1)}, \boldsymbol{X}_{1:t})$ is Normally distributed with mean linear in $y_{1:(t-1)}$ and X_t and constant covariance (I refer the reader to Yang et al. (2015) for the full derivation). The adapted ARGO model is then:

$$y_t = \mu_y + \sum_{j=1}^{N} \tilde{\alpha}_j y_{t-j} + \sum_{s=1}^{S} \tilde{\beta}_s X_{s,t} + \epsilon_t, \qquad \epsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \tag{2.1}$$

The model is therefore an autoregressive model with exogenous variables (ARX), where the search activity and for-sale listing activity captured by $\boldsymbol{X}_t$ serves as exogenous variables for the time series $\{y_t\}$.

2.3.1 Model Estimation

The model parameters, $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ are regularized using an $\mathbb{L}_1$ penalty and estimated via Maximum Likelihood Estimation as follows. Note that for consistency, I use plain parameter notation to indicate unpenalized parameters trained on a fixed window, parameters with a prime symbol ($'$) to indicate unpenalized parameters trained on a sliding window, and parameters with a tilde ($\tilde{}$) to indicate penalized parameters trained on a sliding window.

$$\{\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}\} = \arg\min_{\{\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}\}} \sum_{t} \left(y_t - \mu_y - \sum_{j=1}^{N} \tilde{\alpha}_j y_{t-j} - \sum_{s=1}^{S} \tilde{\beta}_s x_{s,t} \right)^2 + \lambda ||\tilde{\boldsymbol{\alpha}}|| + \lambda ||\tilde{\boldsymbol{\beta}}|| \quad (2.2)$$

The $\mathbb{L}_1$ penalty imposes a penalty on the sum of the absolute values of the coefficients, which has the effect of shrinking their magnitude, often to zero so that the associated variables are selected out of the model. In this way, the penalty serves as both a shrinkage and a variable selection method. The appeal of penalized regression models lies in their ability to distinguish predictive search terms from noisy ones and to handle settings where there are more predictors than observations, which would cause ordinary least squares regression to fail. Additionally, they produce interpretable models, which may increase their appeal to a broad audience relative to other more complex models. The model is re-trained every period using a 2-year sliding window, which captures changing patterns in search and time series behavior.

2.4 Data

To estimate the model on the housing market, I drew data from a number of different sources, both public and proprietary. I downloaded publicly available data from Zillow[2], FHFA[3], St. Louis Fed[4] and Google Trends[5], all of which were

[2]https://www.zillow.com/research/data/

[3]https://www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index-Datasets.aspx#qexe

[4]https://fred.stlouisfed.org/tags/series?t=population

[5]https://trends.google.com/trends/?geo=US

organized at the US state level, and proprietary data from Zillow's home sales records and clickstream logs that I accessed through an internship with Zillow.

The publicly available Zillow data comprised two separate files, both with state-level outcomes reported monthly. The first contained seasonally adjusted home sales volumes, which started between March 2008 and June 2017, depending on the state, and extended through the end of 2018. The second contained seasonally adjusted median sales prices, which started between March 2008 and May 2015, again depending on the state, and extended through the end of 2018. I transformed the monthly sales counts to quarterly counts by summing across the three months in each quarter. To transform the median sale prices from monthly to quarterly values, I calculated a weighted mean across the three months in each quarter, using weights proportional to the corresponding monthly sales volumes.

The FHFA data contained quarterly measures of the House Price Index (HPI) by US state since 1975, and the Google Trends data contained monthly search volumes by state for ten real estate related Google search terms, dating back to the beginning of 2004. To transform the monthly Google search volumes to quarterly volumes, I took a simple average across the three months in each quarter.

In addition to the publicly available data, I analyzed proprietary Zillow data, both for zip code level outcomes (e.g. median sale price) in several US cities and for state and zip code level search data and listing information. To estimate the zip code level outcomes, I aggregated sales volumes and prices by zip code and month and then applied a seasonal adjustment estimated from a seasonal decomposition of time series by Loess (STL) applied to the time series of all properties in the

city where the zip code is located.

To capture search activity and listing information at both the state and zip code levels, I used Zillow's clickstream data. Specifically, to measure search activity for geography i in time period t, I identified users who viewed at least one home in geography i for at least sixty seconds during time period t, calculated the number of unique sessions in which the user viewed a home in geography i for at least sixty seconds during time period t, and then summarized to obtain the total counts of relevant users and sessions, organized by user type (e.g. agent or consumer). I required users to have viewed a property for at least sixty seconds, since previous research on dwell time found it to be a good proxy for user interest. In the present work, I found better predictive performance when I required users to stay on a home listing page for at least sixty seconds versus when I counted any interaction users had with a home listing page, no matter how short.

I additionally used this data to obtain raw measures of available inventory and median listing price, both of which were then seasonally adjusted using an adjustment factor estimated by a seasonal decomposition of time series by Loess (STL). For the zip code level analysis, the STL was estimated on the equivalent time series for all listings in the relevant city, while for the state-level analysis, the STL was estimated individually for each state.

To calculate the raw inventory and list price measures for geography i in time period t, I first identified each for-sale home in geography i that was viewed at least once during time period t. For homes that experienced any price changes during time period t, I calculated a within-home median list price, resulting in at

most one record per for-sale home. I then calculated the median list price across these records and counted the number of unique properties to yield the list price and for-sale inventory data captured in the vector of exogenous variables $\boldsymbol{X}_t$.

In addition to the housing market, I applied ARGO to key macroeconomic indicators, including unemployment and GDP. For this analysis, I downloaded Google search volumes at the US-level from Google Trends for a number of relevant search terms, including unemployment, Social Security, SNAP, food stamps, etc. Additionally, I downloaded quarterly GDP and monthly unemployment numbers from FRED, the economic data tool for the Federal Reserve Bank of St. Louis.

2.5 Results

To understand the value of ARGO in the housing market context as well as the additional value of incorporating search data from Zillow, I begin by replicating and extending earlier work by Wu & Brynjolfsson (2015). This analysis uses key housing market metrics at the U.S. state level and demonstrates significant value of using ARGO over a more simplistic autoregressive model. In particular, it shows huge returns to using a sliding training window, as the market exhibited strong patterns of non-stationarity following the financial crisis of 2008. I find further model improvement from incorporating a larger number of penalized regressors as well as from using Zillow search data in place of Google search data.

I then apply the same techniques to neighborhood data from lower cost neighborhoods in a number of US cities. I again find that ARGO with Zillow search

data improves over traditional time series models and that Zillow search data adds information even after incorporating a sliding training window and the full set of penalized lagged outcomes.

2.5.1 Nowcasting State-Level Trends

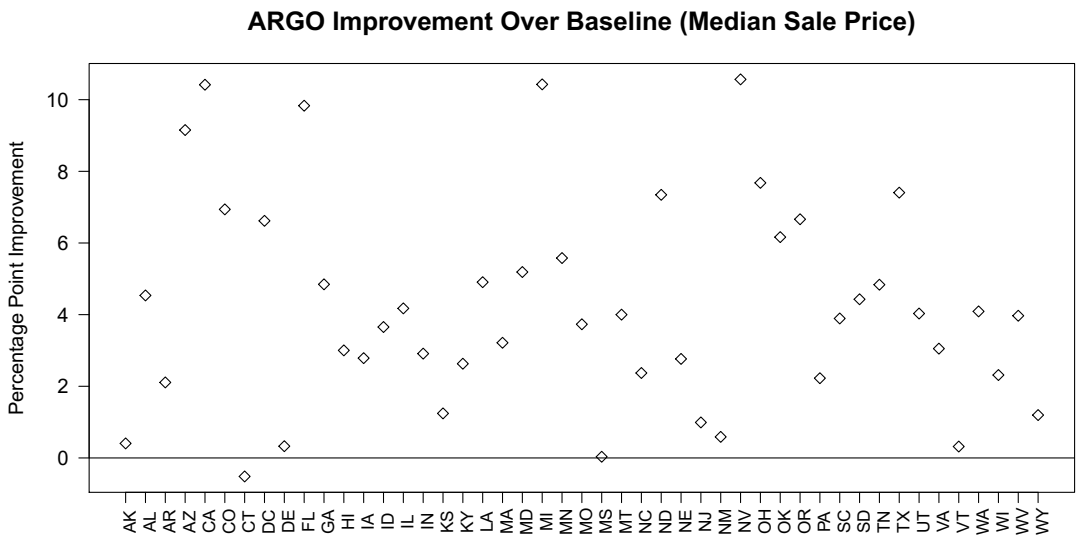
Wu & Brynjolfsson (2015) demonstrate the value of incorporating real estate-related Google search volumes in a simple autoregressive model to nowcast and forecast state-level housing market indicators. They found that when search data was added to an autoregressive model of home sales volumes, the predictive accuracy generally improved, though there was some heterogeneity across states. When search data was added to an autoregressive model of Housing Price Index (HPI), however, no improvement was observed, which the authors attributed to the fact that the search terms they used did not distinguish between buyers and sellers and so gave no measure of relative supply and demand.

In this section, I replicate and extend their work, demonstrating the significant value of using ARGO to predict state-level housing market indicators. I begin with the specification Wu & Brynjolfsson (2015) used as their base specification.

$$\begin{aligned} y_{it} = \mu_y + \alpha y_{i,t-1} + \gamma_1 w_{i,t-1} + \gamma_2 \text{Population}_{it} \\ + \sum_k M_k \times \mathbb{1}(State_i = State_k) + \sum_l R_l \times \mathbb{1}(Region_i = Region_l) \\ + \sum_q T_q \times \mathbb{1}(Quarter_t = Quarter_q) + \epsilon_{it} \end{aligned} \tag{2.3}$$

Where y_{it} is the outcome of interest, such as the volume of home sales in state i during time period t, $w_{i,t-1}$ is another feature of the housing market, such as the

Figure 2.1: Reduction in error from using ARGO relative to the base specification, by state



This figure shows the reduction in error from using ARGO over the period Q2-2012 to Q4-2014 compared to the model from Wu & Brynjolfsson (2015) by state. Dots above the zero line indicate that ARGO performs better than their specification while dots below indicate it performs worse. A value of 6% should be interpreted as a reduction in error of 6 percentage points.

Housing Price Index (HPI), in state i during time period $t-1$, M_k is the fixed effect for state k, R_l is the fixed effect for region l (e.g. Midwest), and T_q is the fixed effect for the quarter of year. I then add Google search volumes to the base specification, similarly to what was done in Wu & Brynjolfsson (2015). Since the specific search terms used by Wu & Brynjolfsson (2015) are no longer tracked by Google Trends, I used similar terms that are currently tracked by Google Trends. Specifically, whereas Wu & Brynjolfsson (2015) included search volumes for the categories "real estate listing" and "real estate agencies," I incorporated search volumes for the category "real estate" and the search terms "real estate agent" and "real estate listings." The specification including Google search terms is the same as the base specification with the addition of contemporaneous and lag-1 search volumes, which are represented respectively by $\boldsymbol{X}_{i,t} = (x_{i,1,t}, x_{i,2,t}, x_{i,3,t})$ and $\boldsymbol{X}_{i,t-1} = (x_{i,1,t-1}, x_{i,2,t-1}, x_{i,3,t-1})$.

$$\begin{aligned} y_{it} = \mu_y + \alpha y_{i,t-1} + \sum_{s=1}^{3} (\beta_s x_{i,s,t} + \delta_s x_{i,s,t-1}) + \gamma_1 w_{i,t-1} \\ + \gamma_2 \text{Population}_{it} + \sum_k M_k \times \mathbb{1}(State_i = State_k) \\ + \sum_l R_l \times \mathbb{1}(Region_i = Region_l) + \sum_q T_q \times \mathbb{1}(Quarter_t = Quarter_q) + \epsilon_{it} \end{aligned} \tag{2.4}$$

Wu & Brynjolfsson (2015) train their model on a static period and then use the fixed model to predict forward on all remaining periods. Their training period covers Q1 2006 through Q4 2008, which was a period of extreme instability in the US housing market and so it is unclear that patterns learned during this period should generalize forward. Therefore, the next model tested trains the same model

as in Equation (2.4) but uses a two-year sliding window to better capture changing dynamics of the housing market. Aside from the sliding window, the specification is the same as in Equation (2.4). As in Section 2.3, I use the prime symbol ($'$) to denote unpenalized coefficients trained on a sliding window, which helps differentiate parameters in Equation (2.5) from those in Equation (2.4).

$$\begin{aligned} y_{it} = \mu'_y + \alpha' y_{i,t-1} + \sum_{s=1}^{3} (\beta'_s x_{i,s,t} + \delta'_s x_{i,s,t-1}) + \gamma'_1 w_{i,t-1} + \gamma'_2 \text{Population}_{it} \\ + \sum_k M'_k \times \mathbb{1}(State_i = State_k) + \sum_l R'_l \times \mathbb{1}(Region_i = Region_l) \\ + \sum_q T'_q \times \mathbb{1}(Quarter_t = Quarter_q) + \epsilon_{it} \end{aligned} \tag{2.5}$$

ARGO uses a penalized regression to select informative lags and search measures. Therefore, the next model extends the set of lagged outcomes to 8 quarters and extends the set of search terms to include "Zillow," "Redfin," "Trulia," "sell home," "home buying," "mortgage," and "realtor," resulting in a total of 10 unique search terms. It also applies an $\mathbb{L}_1$ penalty to all right hand side variables to regularize and avoid overfitting the extended variable set. As before, I use a tilde to indicate penalized parameters trained on a sliding window.

$$\begin{aligned} y_{it} = \mu'_y + \sum_{j=1}^{8} \tilde{\alpha}_j y_{i,t-j} + \sum_{s=1}^{10} \left(\tilde{\beta}_s x_{i,s,t} + \tilde{\delta}_s x_{i,s,t-1} \right) + \tilde{\gamma}_1 w_{i,t-1} \\ + \tilde{\gamma}_2 \text{Population}_{it} + \sum_k \tilde{M}_k \times \mathbb{1}(State_i = State_k) \\ + \sum_l \tilde{R}_l \times \mathbb{1}(Region_i = Region_l) + \sum_q \tilde{T}_q \times \mathbb{1}(Quarter_t = Quarter_q) + \epsilon_{it} \end{aligned} \tag{2.6}$$

The final model specification replaces the Google search terms with measures of search activity and available listings on Zillow, which are represented by the letter z to distinguish them from Google search data, e.g. Zillow search data at time t is the vector $Z_t = (z_{1,t}, ..., z_{S,t})$ while Google search data at time t is the vector $X_t = (x_{1,t}, ..., x_{S,t})$. Specifically, the model uses the number of unique sessions where a home in state i was viewed for more than 60 seconds and the number of unique users who viewed a home in state i for more than 60 seconds, aggregated by user type (e.g. agent, consumer, professional, or unknown / not logged in). It also includes the seasonally adjusted median list price and for-sale inventory for state i during time period t, resulting in a total of 10 measures of Zillow search and market data.

$$\begin{aligned} y_{it} = \mu_y' + \sum_{j=1}^{8} \tilde{\alpha}_j y_{i,t-j} + \sum_{s=1}^{10} \left(\tilde{\beta}_s z_{i,s,t} + \tilde{\delta}_s z_{i,s,t-1} \right) + \tilde{\gamma}_1 w_{i,t-1} \\ + \tilde{\gamma}_2 \text{Population}_{it} + \sum_k \tilde{M}_k \times \mathbb{1}(State_i = State_k) \\ + \sum_l \tilde{R}_l \times \mathbb{1}(Region_i = Region_l) + \sum_q \tilde{T}_q \times \mathbb{1}(Quarter_t = Quarter_q) + \epsilon_{it} \end{aligned} \tag{2.7}$$

The data used in this paper begins later than the data used in Wu & Brynjolfsson (2015), with outcomes data for most states beginning in Q2 2008. In an effort to most closely represent their work, I train the model on the earliest two-year window available to me and then evaluate predictive performance on the subsequent 11 periods, allowing the same length of evaluation as used in their paper. The data begins in Q2 2008 and the model requires 16 pre-periods (8 for the 2-year sliding window and another 8 so that the first training point has a full set of out-

come lags for ARGO). The model is therefore trained from Q2-2010 to Q1-2012 and then evaluated from Q2-2012 to Q4-2014. New York, New Hampshire, Maine and Rhode Island each have missing outcomes during this time period and are thus omitted from the analysis. To evaluate model performance, I calculate the Mean Absolute Percent Error (MAPE) for each state and then average across states, including the District of Columbia. The MAPE is estimated according to the following equation.

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{|y_t - \hat{y}_t|}{y_t} \tag{2.8}$$

Where N indexes over the states and T indexes over the time periods. The models are evaluated on three $\{y_{it}, w_{i,t-1}\}$ pairs. In the first specification, the outcome y_{it} is the seasonally adjusted home sales count (volume of sales) and $w_{i,t-1}$ is the House Price Index (HPI). In the second specification, the outcome y_{it} is the seasonally adjusted median sale price while for the third specification, the outcome y_{it} is the House Price Index (HPI). For both the second and third specifications, $w_{i,t-1}$ is the seasonally adjusted home sales count. Although Wu & Brynjolfsson (2015) used repeat sales of single family homes, I use all sales of single family homes and condominiums in order to avoid skewed results in urban areas. The results are presented in Table 2.1.

The results presented in Table 2.1 demonstrate that there is significant value in using a sliding window to train the model as well as additional incremental value in including a larger set of search terms with an $\mathbb{L}_1$ penalty. Models (2.4) and

Table 2.1: Mean absolute percent error (MAPE) by model at the state level Q2-2012 to Q4-2014

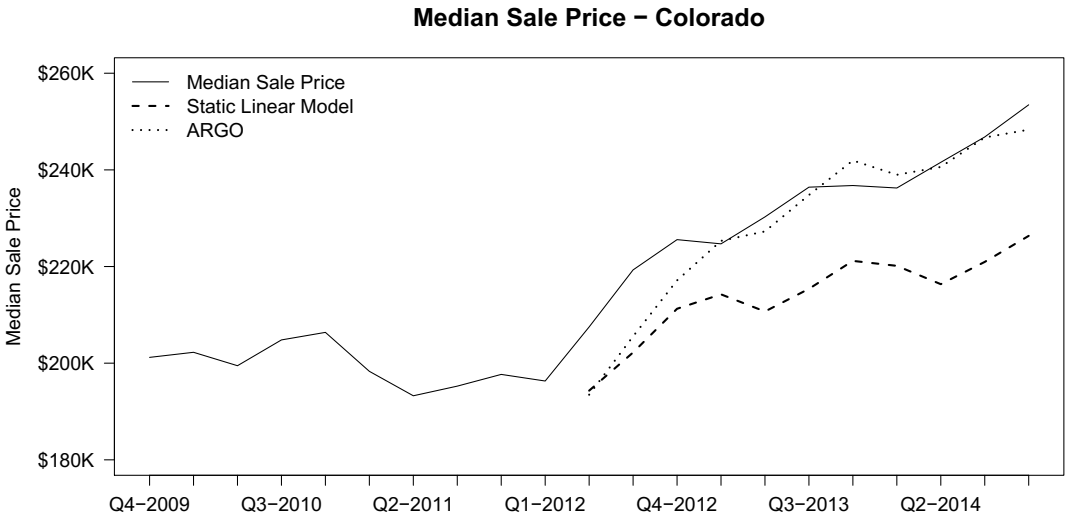
	(2.3)	(2.4)	(2.5)	(2.6)	(2.4) - (2.6)
Home Sales Volumes	0.2042	0.3337	0.1421	0.1370	0.1966
Median Sale Price	0.0537	0.0662	0.0248	0.0238	0.0423
House Price Index (HPI)	0.0148	0.0161	0.0097	0.0086	0.0074

Comparison of model predictive performance between Q2-2012 and Q4-2014 for models trained on 8 quarters of state-level data. Improvement over the specification from Wu & Brynjolfsson (2015) is shown in the rightmost column.

(2.5) share identical specifications but Model (2.4) is trained once on the two-year pre-period Q2-2010 to Q1-2012 while Model (2.5) is re-trained each period on the previous eight quarters of data. This simple difference accounts for about a 50% reduction in error across all outcomes. Adding additional Google search volumes and applying an $\mathbb{L}_1$ penalty to all regressors further reduced model error by up to 10%.

The results do not replicate those found by Wu & Brynjolfsson (2015), as adding unpenalized Google search terms to the static auto-regressive model decreased model performance. The difference is likely due to changing Internet usage patterns, which have both changed the relationship between search volumes and housing market indicators and have made different search terms more predictive over time. This further illustrates the value of allowing the model to both select relevant search terms automatically as well as to continuously update its understanding of the relationship between Internet search volumes and housing market trends. In Figure 2.2, the importance of a sliding window is clearly illustrated for the case of Colorado, whose real estate trends changed course in early 2012. While

Figure 2.2: Median sales prices and predictions for Colorado



The time series of seasonally adjusted median sale price for the state of Colorado is shown as the solid black line. Predictions from a static linear model are represented by the dashed line, and the ARGO predictions are represented by the dotted line. A clear regime change occurred in early 2012 that ARGO was able to quickly adapt to whereas the static model did not.

the models trained on a sliding window were quickly able to recognize the regime change, the static model remained persistently low, due in part to the negative coefficient it learned on population, which continued to rise in Colorado over the period under study.

The next analysis examines whether and to what extent incorporating measures of search activity and for-sale listings from Zillow further improves the ARGO model over its implementation with Google search volumes. The clickstream data from Zillow dates back to mid-2015, which significantly limits the opportunity to evaluate its performance relative to the other models considered thus far. Therefore, I first train the model on the standard 8 quarters of data and evaluate its performance relative to the other models on 4 quarters of outcomes. I then train the model on only 4 quarters, allowing 8 quarters for evaluation, and again compare it to the full set of models. The results are shown in Table 2.2. The model with Zillow clickstream data shows improvement relative to the model with Google search data, though the improvement is very small in some cases. Overall, the difference between the sliding window and static models is no longer notable, likely due to the short evaluation window, which limits the impact of non-stationary trends.

Interestingly, when the models are trained on only four quarters of data, the ARGO model with Zillow search activity does significantly better than the simple autoregressive models and performs roughly as well as when it was trained on eight quarters of data. This suggests that search activity is able to provide sufficient information to compensate for shortened time series without overfitting.

Table 2.2: Mean absolute percent error (MAPE) by model (including ARGO with Zillow data) at the state level

	(2.3)	(2.4)	(2.5)	(2.6)	(2.7)	(2.4) - (2.7)
Trained on 8 Quarters						
Home Sales Volumes	0.0881	0.0817	0.0768	0.0668	0.0668	0.0149
Median Sale Price	0.0195	0.0193	0.0199	0.0190	0.0182	0.0011
House Price Index	0.0053	0.0053	0.0055	0.0085	0.0085	-0.0032
Trained on 4 Quarters						
Home Sales Volumes	0.1375	0.1310	0.1111	0.0769	0.0722	0.0588
Median Sale Price	0.0376	0.0458	0.0291	0.0157	0.0153	0.0305
House Price Index	0.0123	0.0152	0.0106	0.0090	0.0089	0.0064

Comparison of state-level model performance. In order to allow comparison with ARGO using Zillow data, the models are trained and evaluated on later periods than in Table 2.1. The model trained on 8 quarters is evaluated on Q4-2017 through Q3-2018, while the model trained on 4 quarters is evaluated on Q4-2016 through Q3-2018.

2.5.2 Nowcasting Hyper-Local Trends

The Zillow search data affords the new and exciting opportunity of hyper-local nowcasting, which is not possible with Google search data, since Google Trends can only be de-aggregated to the Metropolitan Statistical Area (MSA) level. The results presented here are also more broadly relevant to organizations who wish to improve niche forecasts, for example of a customer segment or of a product line. In particular, I show that clickstream data is likely to improve such forecasts when used in a regularized model, which limits sensitivity to noisy clickstream measures by shrinking coefficients and selecting out the least informative measures.

To understand the value of Zillow clickstream data at the hyper-local level, I study trends at the zip code level in several cities across the US. To further understand the policy relevance of this approach, and in particular to identify

neighborhoods at risk of gentrification (defined here as rising home sales prices), I repeat the analysis by training and evaluating the model only on the subset of lower-cost zip codes in each city. Zip codes are considered lower-cost if the median sale price in 2012 was below the 2012 median sale price for the relevant city. 2012 was used for benchmarking, as data from 2013 forward was used to estimate and evaluate model performance. I use zip codes rather than neighborhoods to ensure well-defined boundaries while maintaining small geographical areas. I drop any zip codes where on average fewer than 2 homes sold per month, zero homes sold in at least one month, or the median sale price exceeded $10M in any month. Overall, these restrictions have a very small effect on my sample while improving the average quality of the time series under study. In particular, these restrictions drop no zip codes in Denver, CO or Washington, DC, one zip code in Atlanta, GA, Boston, MA, and San Francisco, CA, three in Portland, OR, and seven in Chicago, IL. Chicago had 56 zip codes to begin with, while all other cities had between 19 and 28.

To evaluate model performance, I consider several specifications, each trained on 24 months of data, and evaluate them over the period September 2017 through October 2018. The first specification is an extremely simple autoregressive model with one lag, which is trained on the period September 2015 to August 2017.

$$y_{it} = \mu_y + \alpha y_{i,t-1} \tag{2.9}$$

The second model evaluated uses the same specification but is trained on a two-year sliding window. As before, I use the prime symbol ($'$) to distinguish coeffi-

cients trained on a sliding window in Equation (2.10) from coefficients trained on a static window in Equation (2.9).

$$y_{it} = \mu'_y + \alpha' y_{i,t-1} \tag{2.10}$$

The third model includes 24 lags of the outcome variable, with an $\mathbb{L}_1$ penalty applied to all lags. Again, I use the tilde symbol to indicate penalized parameters trained on a sliding window.

$$y_{it} = \mu'_y + \sum_{j=1}^{24} \tilde{\alpha}_j y_{i,t-j} \tag{2.11}$$

The fourth and final model is the full ARGO specification adapted to use Zillow clickstream measures in place of Google Trends, shown below as Model (2.12). This model adds penalized measures of search activity and listings from Zillow to the specification in Model (2.11). The Zillow clickstream measures, which are described in detail in Section 2.4, are represented by $z_{i,s,t}$, where i indexes over the zip codes, s indexes over the Zillow clickstream measures, and t indexes over the time periods.

$$y_{it} = \mu'_y + \sum_{j=1}^{24} \tilde{\alpha}_j y_{i,t-j} + \sum_{s=1}^{10} \tilde{\beta}_s z_{i,s,t} \tag{2.12}$$

At the zip code level, I again observe significant value in using ARGO with Zillow clickstream data to reduce model error when nowcasting median sales prices. In particular, relative to Model (2.9), ARGO leads to an average improvement in Mean Absolute Percent Error (MAPE) of 6% for all zip codes and 11% for lower

cost zip codes. Since outliers can be a concern for the housing market, I also calculate the Median Absolute Percent Error (MedAPE) and find that relative to Model (2.9), ARGO leads to an average improvement of 13% in MedAPE for all zip codes and an improvement of 6% for lower cost zip codes. The average improvement is generally smaller though surprisingly similar when comparing ARGO to Model (2.11), which contains all elements of ARGO except for the clickstream measures. In particular, I find that ARGO reduces average MAPE by 7% for all zip codes and by 6% for lower cost zip codes relative to Model (2.11). It also decreases average MedAPE by 8% for all zip codes and by 1% for lower cost zip codes relative to Model (2.11). Therefore, it appears that time series alone may be overly noisy at very fine levels of aggregation and thus that predictive models at such levels of aggregation would benefit from supplementary measures, such as clickstream or search data, to improve predictive accuracy.

The results for each city are shown in Tables 2.3 through 2.6. Table 2.3 shows the Mean Absolute Percent Error (MAPE) by city for all zip codes, while Table 2.4 shows the MAPE by city for lower cost zip codes only. Similarly, Table 2.5 shows the Median Absolute Percent Error (MedAPE) by city for all zip codes, while Table 2.6 shows the MedAPE by city for lower cost zip codes only.

The city where ARGO shows the worst performance is San Francisco. The San Francisco housing market is extremely expensive and is often featured in the media as a result of its high prices. Therefore, it seems likely that a significant portion of traffic to San Francisco listings may be driven by people who are not active in the San Francisco housing market, but are rather driven to ogle at the price

Table 2.3: Mean absolute percent error (MAPE) for median sale price predictions at the zip code level, by city

	(2.9)	(2.10)	(2.11)	(2.12)	Imp. 1	Imp. 2
Atlanta, GA	0.2056	0.2091	0.1753	0.1747	0.1506	0.0035
Boston, MA	0.1554	0.1576	0.1474	0.1402	0.0979	0.0488
Chicago, IL	0.5427	0.557	0.8262	0.5842	-0.0765	0.2929
Denver, CO	0.1208	0.1214	0.1061	0.1055	0.1266	0.0057
Portland, OR	0.0937	0.0942	0.0920	0.0837	0.1067	0.0908
San Francisco, CA	0.1389	0.1410	0.1268	0.1287	-0.0734	0.0153
Washington, DC	0.1243	0.1251	0.1167	0.1154	0.0717	0.0118

Imp. 1 = The percent improvement of Model (2.12) over Model (2.9). Imp. 2 = The percent improvement of Model (2.12) over Model (2.11). For each city, the model was trained on median sales price data from all zip codes included after restrictions described in Section 2.5.2. All zip-code level outcomes were seasonally adjusted using STL estimated from all sales in the relevant city (e.g. not just for the zip code). The model was evaluated over the period ranging from September 2017 to October 2018.

Table 2.4: Mean absolute percent error (MAPE) for median sale price predictions at the zip code level for lower-cost zip codes, by city

	(2.9)	(2.10)	(2.11)	(2.12)	Imp. 1	Imp. 2
Atlanta, GA	0.2254	0.2299	0.1870	0.1825	0.1904	0.0241
Boston, MA	0.0972	0.0911	0.0877	0.0830	0.1458	0.0534
Chicago, IL	0.2929	0.2924	0.3235	0.2581	0.1191	0.2023
Denver, CO	0.0789	0.0796	0.0755	0.0764	0.0311	-0.0126
Portland, OR	0.0513	0.0519	0.0525	0.0466	0.0905	0.1114
San Francisco, CA	0.1379	0.1400	0.1360	0.1350	0.0217	0.0077
Washington, DC	0.1099	0.1103	0.1039	0.1008	0.0823	0.0301

Imp. 1 = the percent improvement of Model (2.12) over Model (2.9). Imp. 2 = the percent improvement of Model (2.12) over Model (2.11). For each city, the model was trained on median sales price data from all zip codes included after restrictions described in Section 2.5.2. All zip-code level outcomes were seasonally adjusted using STL estimated from all sales in the relevant city (e.g. not just for the zip code). The model was evaluated over the period ranging from September 2017 to October 2018.

Table 2.5: Median absolute percent error (MedAPE) for median sale price predictions at the zip code level, by city

	(2.9)	(2.10)	(2.11)	(2.12)	Imp. 1	Imp. 2
Atlanta, GA	0.1659	0.1775	0.1538	0.1479	0.1087	0.0384
Boston, MA	0.1441	0.1397	0.1049	0.1012	0.2978	0.0351
Chicago, IL	0.2655	0.2683	0.3909	0.2592	0.0237	0.3370
Denver, CO	0.0810	0.0772	0.0812	0.0786	0.0304	0.0327
Portland, OR	0.0693	0.0705	0.0659	0.0596	0.1397	0.0947
San Francisco, CA	0.1225	0.1226	0.1038	0.1060	0.1343	-0.0213
Washington, DC	0.1092	0.1102	0.0986	0.0933	0.1462	0.0541

Imp. 1 = the percent improvement of Model (2.12) over Model (2.9). Imp. 2 = the percent improvement of Model (2.12) over Model (2.11). For each city, the model was trained on median sales price data from all zip codes included after restrictions described in Section 2.5.2. All zip-code level outcomes were seasonally adjusted using STL estimated from all sales in the relevant city (e.g. not just for the zip code). The model was evaluated over the period ranging from September 2017 to October 2018.

Table 2.6: Median absolute percent error (MedAPE) for median sale price predictions at the zip code level for lower-cost zip codes, by city

	(2.9)	(2.10)	(2.11)	(2.12)	Imp. 1	Imp. 2
Atlanta, GA	0.1416	0.1468	0.1567	0.1525	-0.0774	0.0268
Boston, MA	0.0876	0.0728	0.0692	0.0729	0.1683	-0.0532
Chicago, IL	0.2372	0.2349	0.2141	0.2025	0.1461	0.0544
Denver, CO	0.0622	0.0607	0.0589	0.0596	0.0427	-0.0111
Portland, OR	0.0499	0.0507	0.0504	0.0466	0.0672	0.0757
San Francisco, CA	0.0986	0.1033	0.0926	0.0973	0.0124	-0.0507
Washington, DC	0.1057	0.1070	0.1027	0.0999	0.0550	0.0277

Imp. 1 = the percent improvement of Model (2.12) over Model (2.9). Imp. 2 = the percent improvement of Model (2.12) over Model (2.11). For each city, the model was trained on median sales price data from all zip codes included after restrictions described in Section 2.5.2. All zip-code level outcomes were seasonally adjusted using STL estimated from all sales in the relevant city (e.g. not just for the zip code). The model was evaluated over the period ranging from September 2017 to October 2018.

tags by articles or other online media. If true, this would mean the fundamental assumption of ARGO is violated in San Francisco, since activity on Zillow would not be closely related to the number of people active in the San Francisco real estate market. It could additionally be the case that while there are many people who would like to buy, and who may therefore view listings, they are largely priced out and so their activity on the site does not end up being reflected in market transactions.

To understand the conditions under which the models perform well versus not, it is helpful to examine the zip codes with the highest and lowest error rates. In general, ARGO performs best in the same neighborhood where the simple AR model also performs best and similarly the two models tend to perform worst in the same neighborhoods as well. This indicates that the overall error rate is largely driven in abnormalities in the underlying time series, which may be better estimated when using search data but are nonetheless difficult to predict with high accuracy. In particular, the models both do best in zip codes where the median price fluctuates within a relatively narrow range.

Given that Chicago and Atlanta consistently showed the highest overall error rates while Boston generally showed the largest improvements from ARGO, I show trends for informative zip codes in each of these three cities. In particular, I show the median sale price true trajectory as well as predictions made from a static AR(1) model, a model with 24 $\mathbb{L}_1$ penalized lags trained on a 2-year sliding window that I refer to as "ARGO lite" (because it contains all elements of ARGO except the search data), and ARGO with Zillow clickstream measures in place

of Google search data. I show the trends and predictions for four zip codes in each city, using MAPE from the ARGO model to identify the zip codes with the highest and lowest error rates. Note that for all three of the cities, the top two and bottom two zip codes were the same whether measured by ARGO MAPE or AR(1) MAPE, though for Chicago, the order within the pairs was not consistent across the two MAPE measures.

Figure 2.3 shows the trends and predictions for Atlanta, GA, Figure 2.4 shows them for Boston, MA, and Figure 2.5 shows them for Chicago, IL. Because it can be difficult to understand the precise performance of each model from those plots, Figures 2.6 through 2.8 show the relative reduction in error from using ARGO versus a simple AR(1) model trained on a 2-year sliding window for the same set of informative zip codes in each of the three cities, and Figures 2.6 through 2.8 show the reduction in error from using ARGO versus ARGO lite (the model with 24 $\mathbb{L}_1$ penalized lags trained on a 2-year sliding window). The trends in Atlanta and Boston confirm that the models do best when prices fluctuate within a small range, while the trends in Chicago highlight the challenge of trying to learn one model for all neighborhoods. The two worst performing neighborhoods in Chicago have median sale prices generally around \$50,000 but the predictions are consistently much higher, around \$175,000. In Chicago, the models learn a very large intercept and a very small coefficient on past sales, because many zip codes exhibit spiky price trends that do not follow a meaningful pattern. Such oscillation highlights the challenges of trying to forecast at such granular levels, where observed trends are subject to significant noise. Additionally, the

poor predictive quality in certain neighborhoods of Chicago highlights a broader question of how best to share information across similar contexts. Although I choose to train one model for all zip codes in a given city, it may instead make sense to train multiple models, using weights to reflect similarities between zip codes. I leave the question of how best to share information across related contexts for future research.

In conclusion, I find that clickstream data improves nowcasting accuracy at the zip code level when used in ARGO, which applies a penalty to the learned parameters in order to limit the risk of overfitting and which trains on a sliding window to capture evolving trends in website usage patterns. The time series at the zip code level are quite volatile, but ARGO nonetheless consistently improves forecasting accuracy across US cities, showing promising policy relevance. In addition, these results demonstrate the value of using penalized measures of clickstream data in improving niche forecasts more broadly, which is likely relevant for organizations who want to improve forecast accuracy for specific product lines or customer segments.

2.6 Extensions

Given ARGO's success at improving nowcast accuracy for trends in the housing market at both the state and zip code level, a natural next question is how well it extends to other economic domains. Additionally, if ARGO is relevant for macroeconomic forecasting, it would be interesting to understand how well it

Figure 2.3: Actual and predicted median sale prices for the Atlanta, GA zip codes with the highest and lowest error rates.

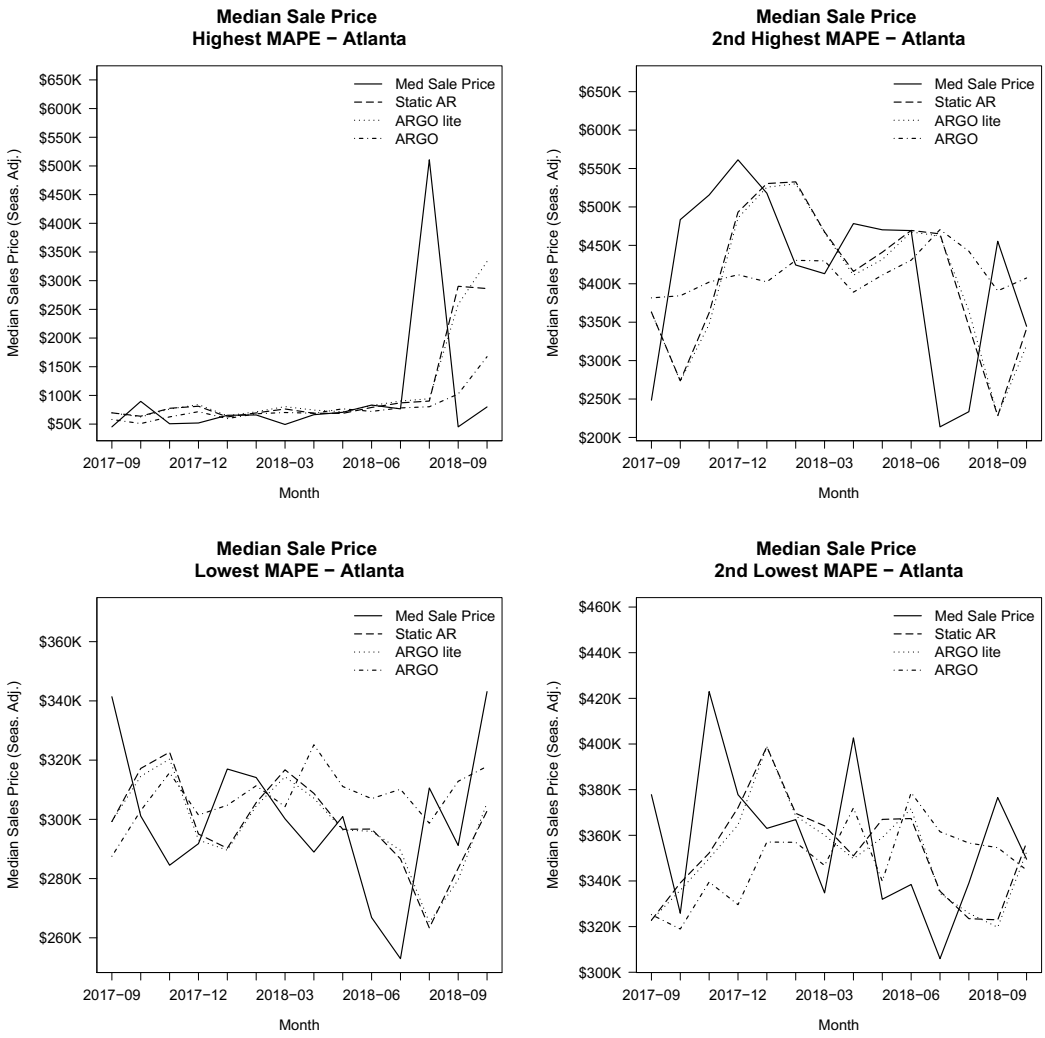


Figure 2.4: Actual and predicted median sale prices for the Boston, MA zip codes with the highest and lowest mean absolute percent error rates.

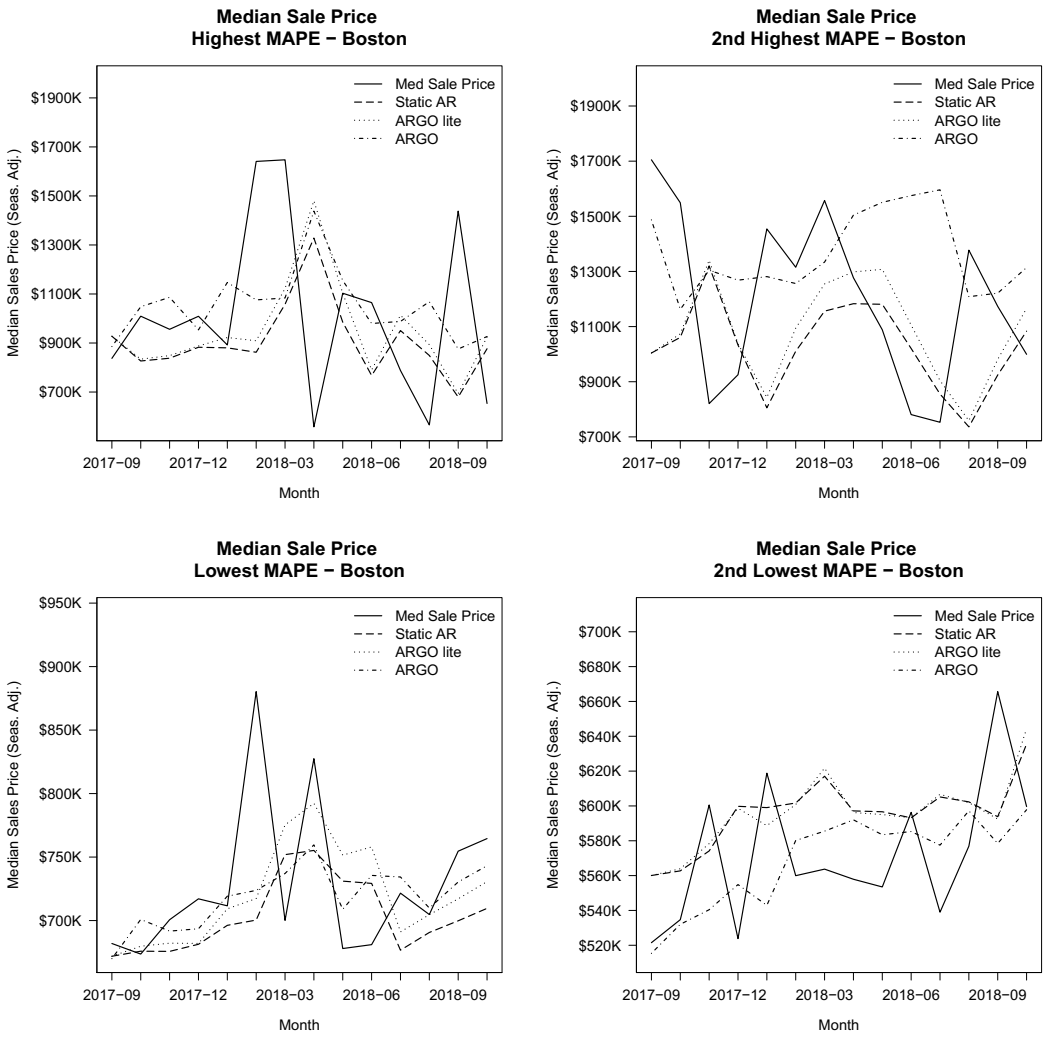


Figure 2.5: Actual and predicted median sale prices for the Chicago, IL zip codes with the highest and lowest mean absolute percent error rates.

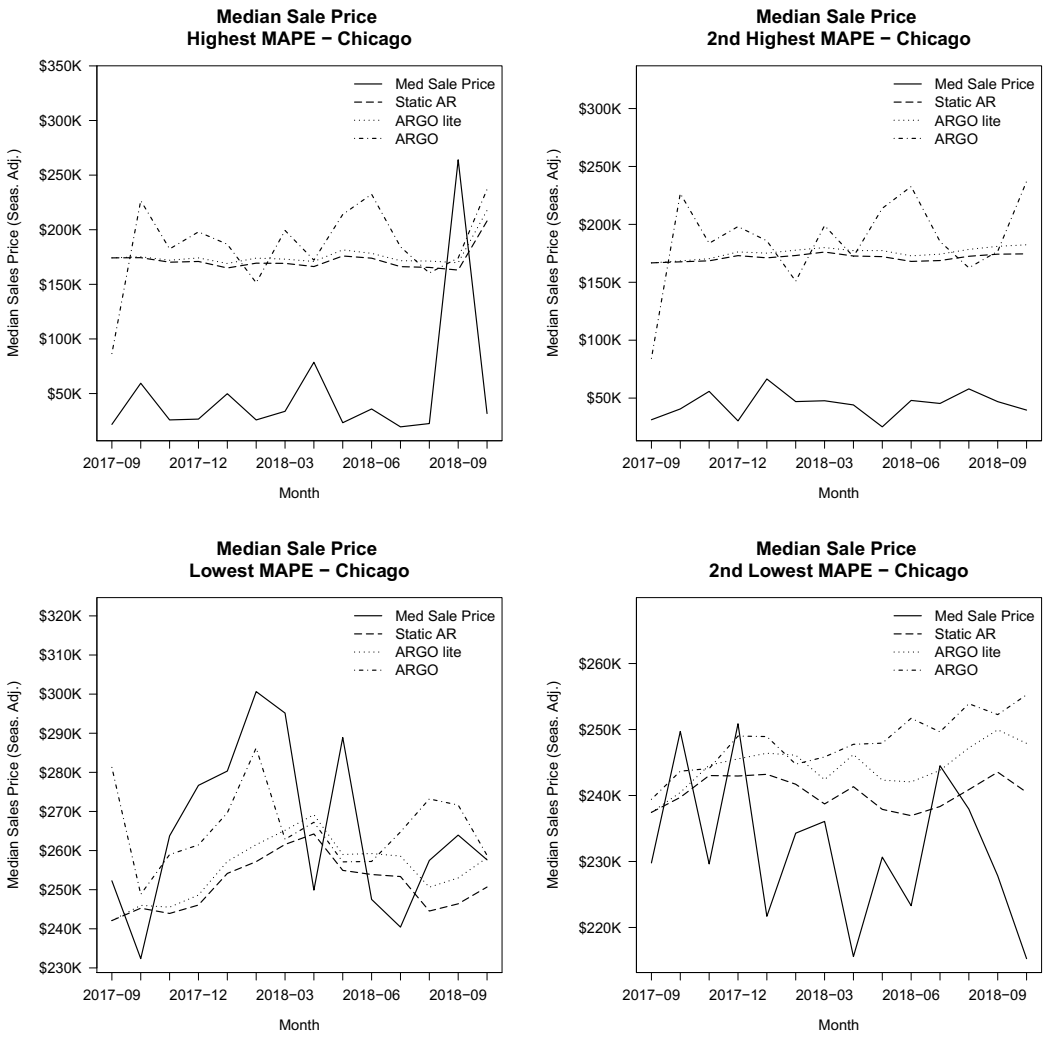
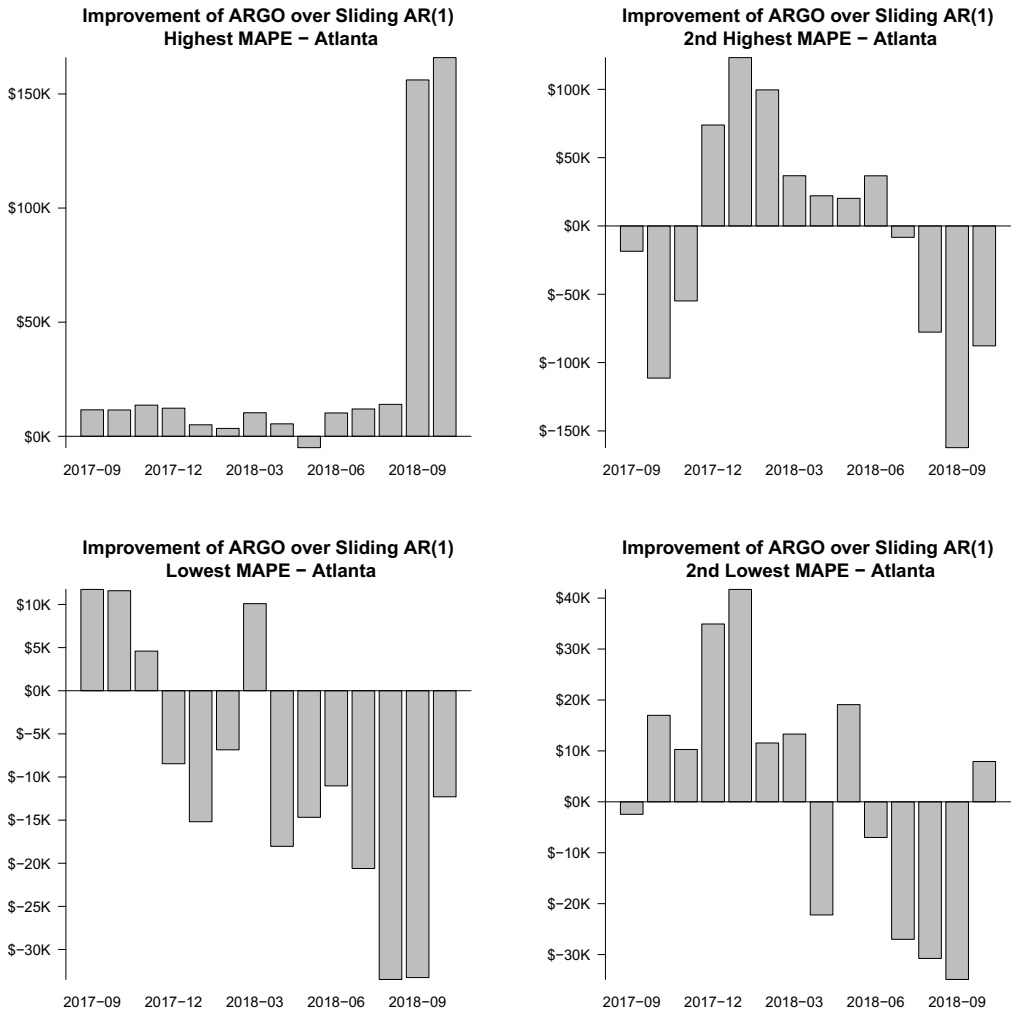
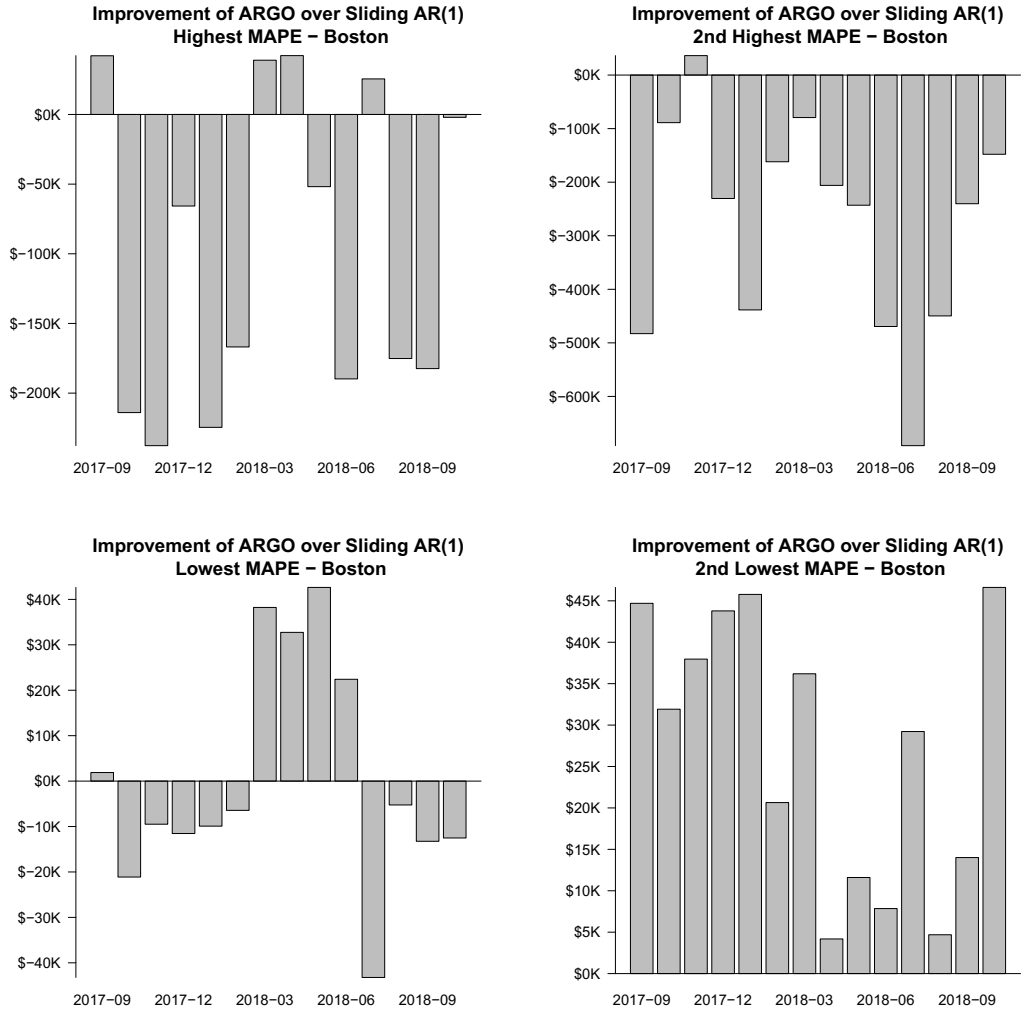


Figure 2.6: Improvement of ARGO over an AR(1) model trained on a 2-year sliding window for the Atlanta, GA zip codes with the highest and lowest mean absolute percent error rates.



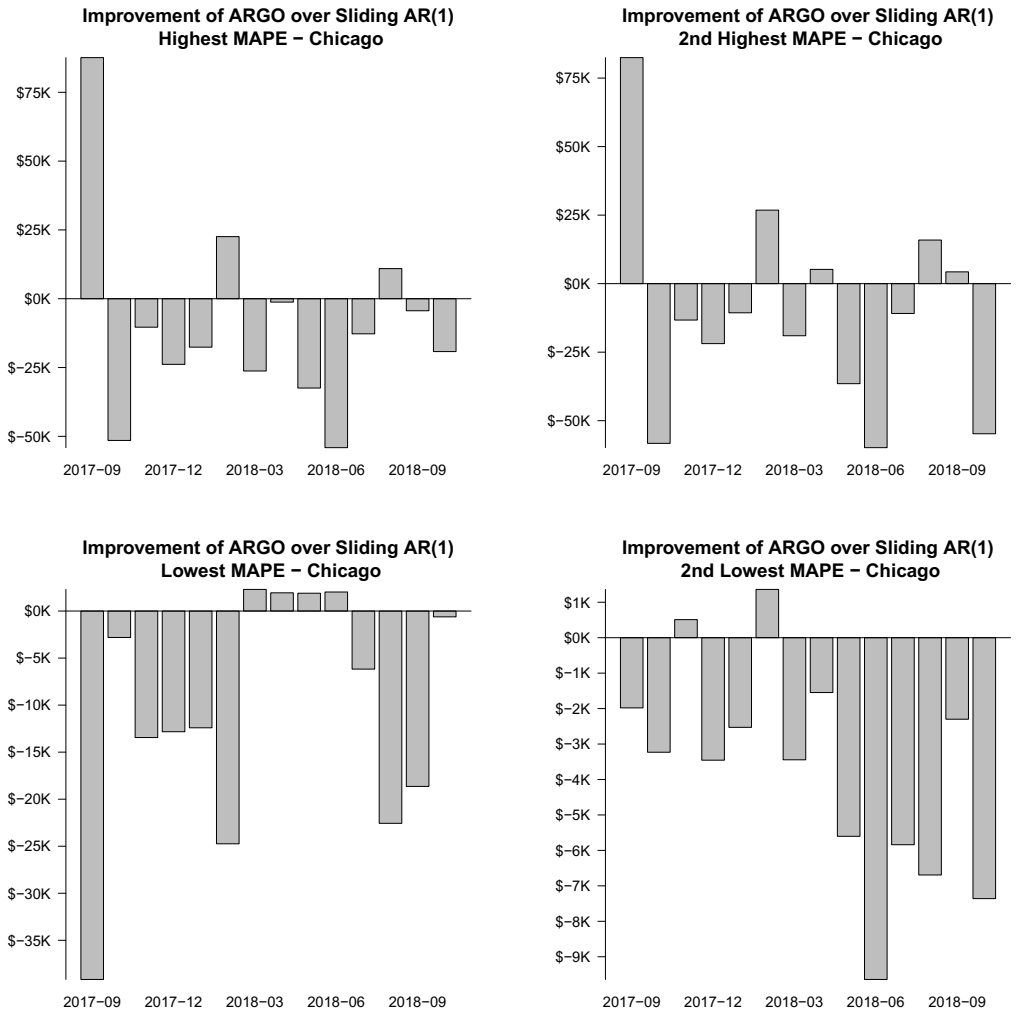
The bars reflect the difference in error rates between the sliding AR(1) model and ARGO in each period. Bars above the horizontal line indicate that ARGO had a lower error (e.g. performed better) than the AR(1) model while bars below the horizontal line indicate that the AR(1) model outperformed ARGO.

Figure 2.7: Improvement of ARGO over an AR(1) model trained on a 2-year sliding window for the Boston, MA zip codes with the highest and lowest mean absolute percent error rates.



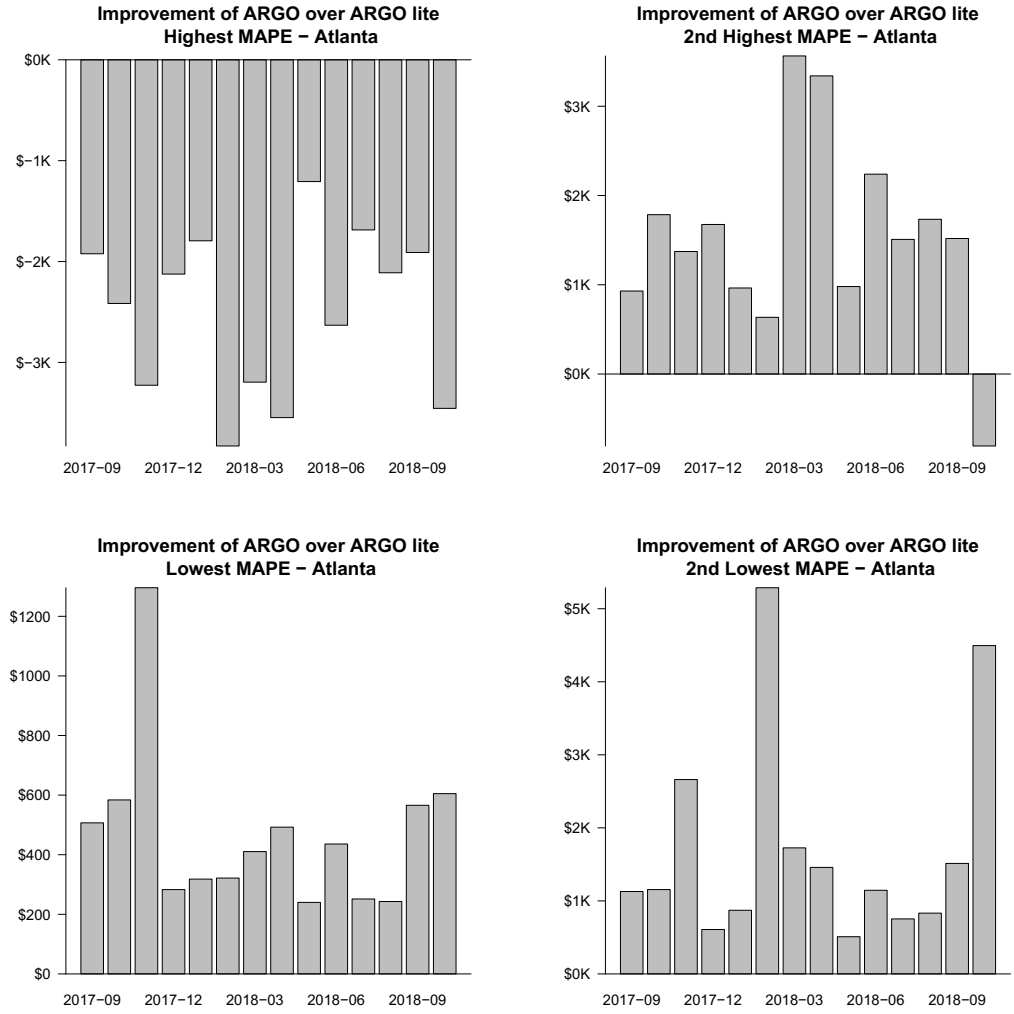
The bars reflect the difference in error rates between the sliding AR(1) model and ARGO in each period. Bars above the horizontal line indicate that ARGO had a lower error (e.g. performed better) than the AR(1) model while bars below the horizontal line indicate that the AR(1) model outperformed ARGO.

Figure 2.8: Improvement of ARGO over an AR(1) model trained on a 2-year sliding window for the Chicago, IL zip codes with the highest and lowest mean absolute percent error rates.



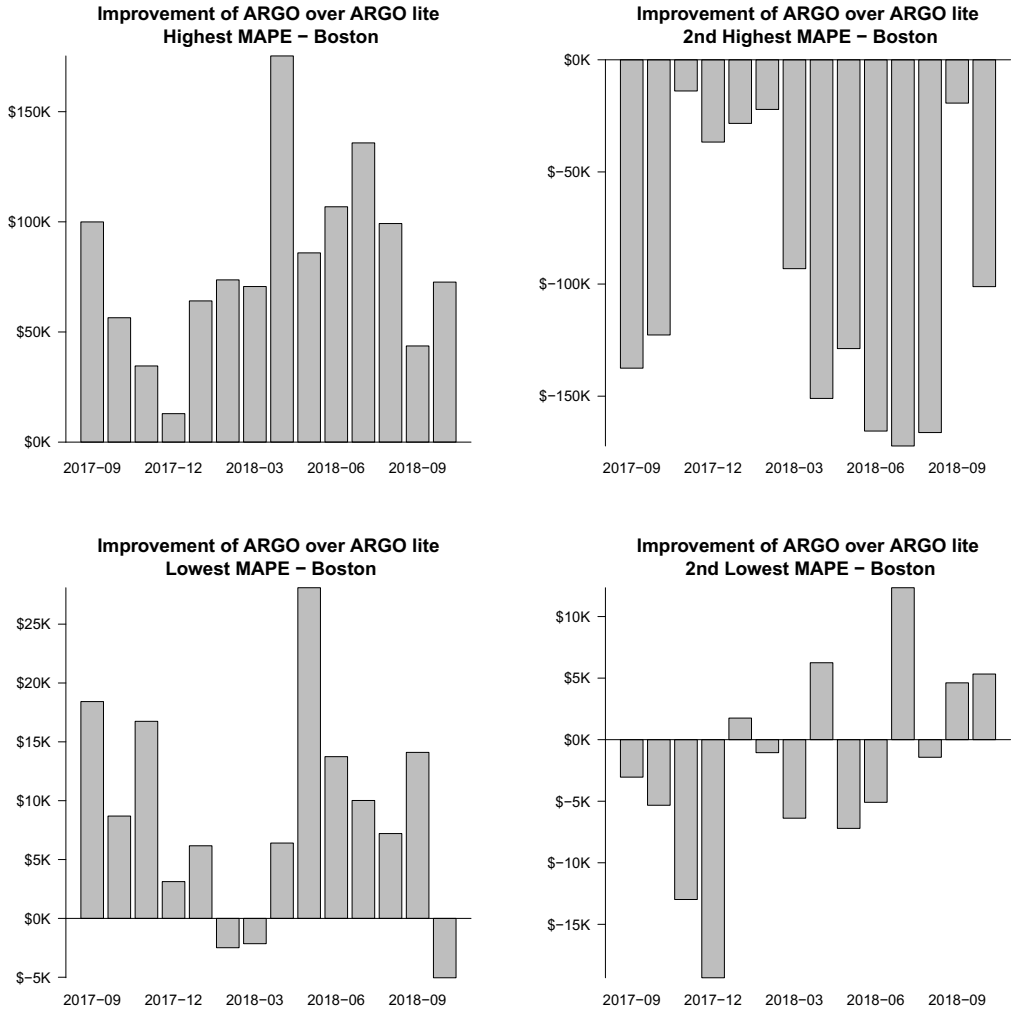
The bars reflect the difference in error rates between the sliding AR(1) model and ARGO in each period. Bars above the horizontal line indicate that ARGO had a lower error (e.g. performed better) than the AR(1) model while bars below the horizontal line indicate that the AR(1) model outperformed ARGO.

Figure 2.9: Improvement of ARGO over ARGO lite for the Atlanta, GA zip codes with the highest and lowest mean absolute percent error rates.



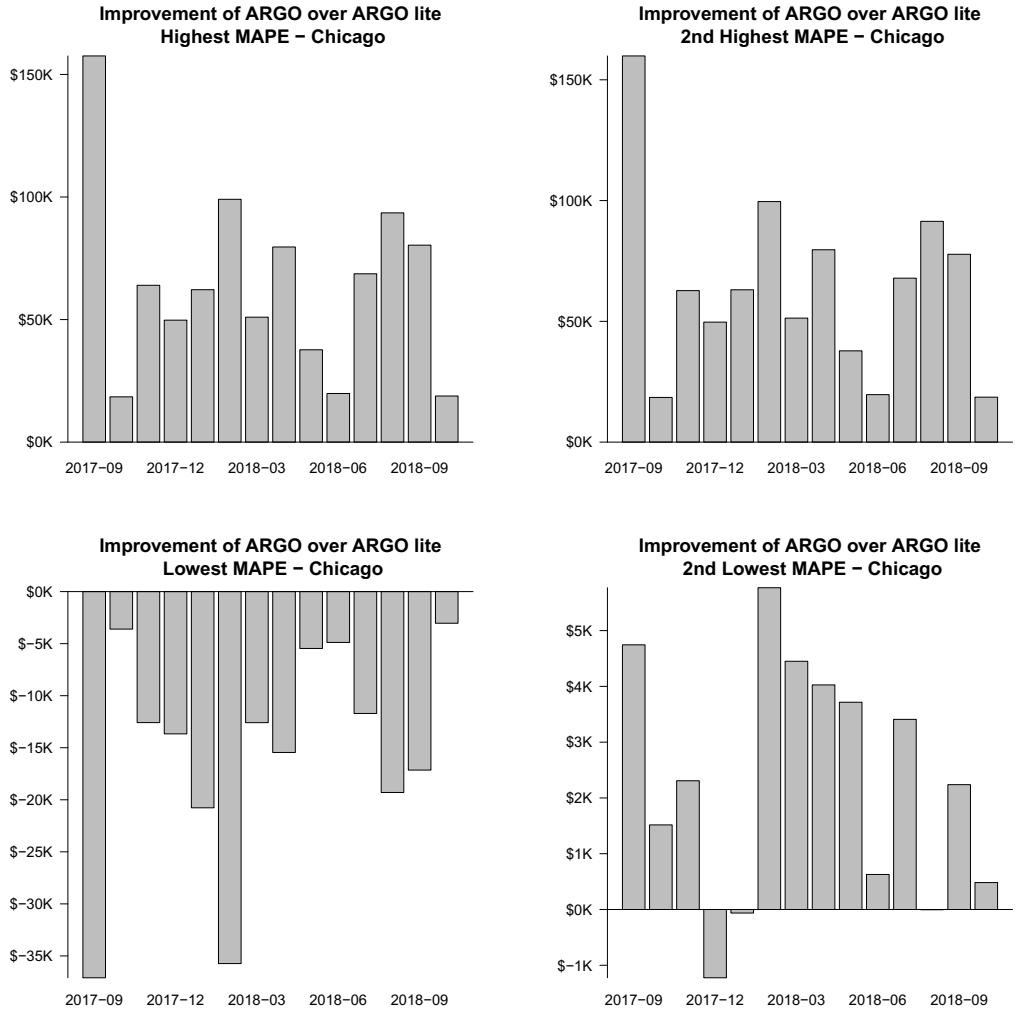
The bars reflect the difference in error rates between the ARGO lite and ARGO in each period. Bars above the horizontal line indicate that ARGO had a lower error (e.g. performed better) than ARGO lite while bars below the horizontal line indicate that ARGO lite outperformed ARGO.

Figure 2.10: Improvement of ARGO over ARGO lite for the Boston, MA zip codes with the highest and lowest mean absolute percent error rates.



The bars reflect the difference in error rates between the ARGO lite and ARGO in each period. Bars above the horizontal line indicate that ARGO had a lower error (e.g. performed better) than ARGO lite while bars below the horizontal line indicate that ARGO lite outperformed ARGO.

Figure 2.11: Improvement of ARGO over ARGO lite for the Chicago, IL zip codes with the highest and lowest mean absolute percent error rates.



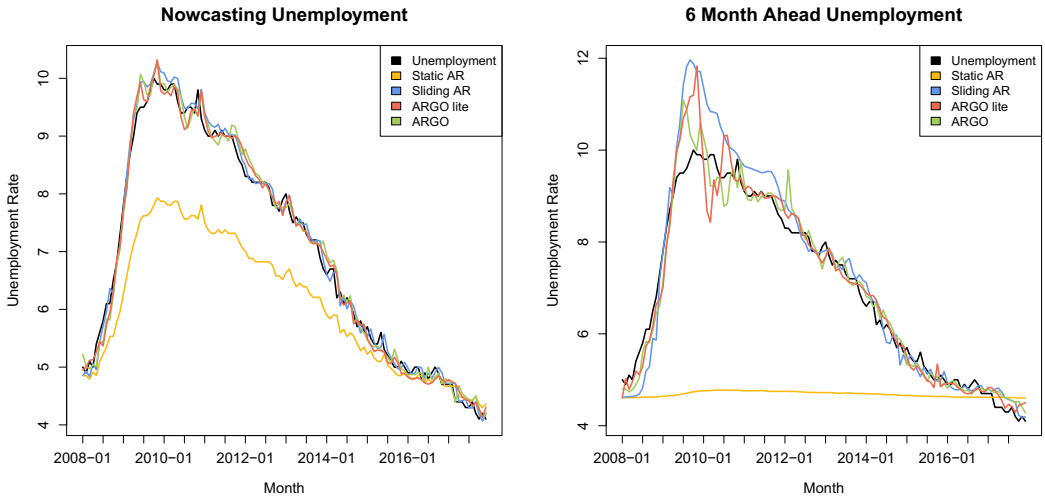
The bars reflect the difference in error rates between the ARGO lite and ARGO in each period. Bars above the horizontal line indicate that ARGO had a lower error (e.g. performed better) than ARGO lite while bars below the horizontal line indicate that ARGO lite outperformed ARGO.

would perform in a downturn and whether search data would provide insight into how deep or long a downturn will be. While I leave a full answer to these questions for future research, I provide some preliminary evidence that suggests ARGO may be a valuable addition to the class of models currently used to forecast key macroeconomic indicators.

In Figure 2.12, I show the actual path unemployment rates took from 2008 through 2014 as well as predictions from a static AR(1) model, a sliding AR(1) model, ARGO lite (ARGO without the search data), and ARGO using Google Trends. I find that for nowcasting, ARGO performs significantly better than the static AR(1) model though roughly on par with both the sliding AR(1) model and ARGO lite. I additionally use the ARGO specification to attempt a 6-month ahead forecast. For this, everything is the same as before, except that instead of predicting an outcome that is contemporaneous with the search data and 1 month ahead of the lagged outcome data, I predict an outcome that is six months ahead of the search data and seven months ahead of the lagged outcome data. In this specification, I find that ARGO outperforms all other methods, notably by overshooting less at the bend where unemployment starts to decrease in early 2010.

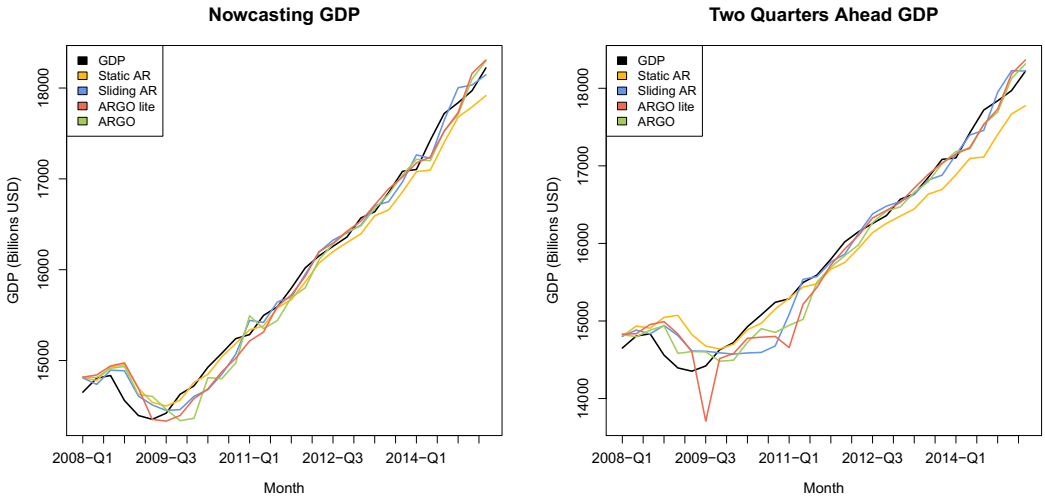
For GDP, the results are similar though slightly less impressive, with ARGO showing little value over non-search models when nowcasting but some gains for six-month-ahead forecasting, especially during the period of transition from 2008 to 2010. Note that because GDP is only reported quarterly, the two-quarter ahead predictions use search volumes and lagged outcomes from the same quarter, which

Figure 2.12: Quarterly unemployment predictions by model



precedes the relevant outcome date by two quarters. The results for GDP are shown in Figure 2.13. It is unsurprising that search data was more relevant for unemployment than for GDP, since unemployment directly affects individuals, who are likely to turn to the Internet to search for new jobs or to understand how to sign up for unemployment benefits. Therefore, ARGO may be most useful for predicting factors such as unemployment and subsequently rolling those improved predictions up to GDP rather than for directly predicting GDP, though further research is needed to understand how best to incorporate ARGO's features in existing state-of-the-art macroeconomic forecast models.

Figure 2.13: Quarterly GDP predictions by model



2.7 Conclusion

In this paper, I extended a state-of-the-art flu tracking model called ARGO to the economic literature by an in-depth application to the housing market. I found that the model performed well at both the state and zip code level and additionally provided some preliminary evidence that the model may add value in improving macroeconomic forecasts, especially during times of economic turbulence. That the ARGO model performed well on an entirely new domain and at various levels of aggregation provides two important lessons. First, it shows that clickstream data holds an important signal whose value is clearest when implemented in a model that both penalizes the clickstream measures to avoid overfitting and trains on a sliding window to understand changing search patterns. Second, it shows that

a powerful yet interpretable model from epidemiology can transfer much of its power to economic domains. Given that the model has already been stress tested and subject to scrutiny in the flu domain, it has proven highly robust and offers important lessons for economics, where interest in nowcasting has grown in recent years.

Nonetheless, significant opportunities for future research on the topic remain. The measures of search used here were collected intuitively with little tuning or revision. It would be valuable to develop a deeper understanding of how best to summarize clickstream data and how such summaries vary by context. Additionally, a question remains of how best to share information across relevant contexts. For the housing market, that may mean how to identify and weight relevant geographies, while for macroeconomic indicators, it may mean understanding where search is most valuable and how to best borrow information across relevant series. For example, it is likely that search better predicts unemployment than GDP, because unemployment is experienced at the individual-level and individuals are likely to turn to the Internet for advice. Therefore, we would likely do best to use search to predict unemployment and then roll the improved unemployment forecast into a GDP forecast rather than attempting to forecast GDP directly with search data.

In this paper, I also showed that ARGO was highly robust to shortening its training window. While traditional time series methods showed significantly worse performance when their training windows were cut in half, ARGO's predictive performance was virtually unaffected. This suggests that search data may com-

pensate for missing outcome lags and may therefore be especially useful in cases where missing outcome data is problematic. Future research should evaluate its effectiveness in addressing the missing data problem and better understand the conditions under which it adds the most value over traditional time series methods.

Chapter 3

Demand Learning and Dynamic Pricing for Varying Assortments

3.1 Introduction

[1]We present our work on a demand learning and dynamic pricing algorithm that efficiently learns customer demand in order to maximize revenue with a small number of price changes. In particular, we study a multi-product, discrete choice setting where products can be fully described by a set of relevant attributes and where assortments change frequently relative to consumers' shopping frequencies. Examples of such settings include retailers who rotate assortments frequently due to a desire to offer new styles or because they are selling perishable or time-

[1]This chapter is co-authored with Kris Ferreira of Harvard Business School

sensitive products.

We describe customer demand using a multinomial logit (MNL) choice model and use dynamic pricing for the purpose of quickly learning consumer demand. Most multi-product demand learning and dynamic pricing algorithms are not contextual and thus cannot be applied to settings like ours where the assortment changes frequently. Furthermore, these algorithms typically either assume a simple demand model with few parameters or require a large number of price changes to learn demand. While such a high volume of price changes may be appropriate for retailers whose prices can be changed rapidly and whose sales volumes per product are high, we are motivated by the setting where retailers are limited either by their sales volume per SKU or by the frequency with which they can change prices. For example, many retailers are unable to change prices frequently due to technological constraints or do not want to change prices frequently because they are concerned about customer perception or strategic consumers.

To address the challenges that arise when retailers are unable or unwilling to frequently change prices, our algorithm changes prices with each assortment and sets prices to learn as quickly as possible. Additionally, we use a contextual attribute-based demand model to share learning across products and assortments, so that our algorithm is effective even in settings where retailers have a limited sales volume for each SKU. Our algorithm follows a learn-then-earn approach, where at first the retailer only prices to learn demand, and then after the retailer is sufficiently confident in the estimated demand model, the retailer prices to earn. Our learning stage is novel in the dynamic pricing literature in that it

borrows methods from conjoint analysis to ensure efficient learning over a short time horizon.

We validate the effectiveness of our algorithm through a three month field experiment with an e-commerce company, where we change the price of products with daily assortment changes. Relative to a control group, our algorithm led to an 8.8% increase in average daily revenue over the three month experiment. The average effect on revenue balances an initial dip in revenue experienced when pricing to learn followed by a significant increase in revenue when pricing to earn.

Our paper makes three main contributions to the literature on dynamic pricing. First, we contribute to the nascent literature for this prevalent retail setting. Namely, we are aware of only one other concurrent paper on dynamic pricing with demand learning for the multi-product discrete choice setting that uses an attribute-based model to account for varying assortments (Javanmard et al., 2019). Second, we develop an algorithm that learns quickly in an environment with varying assortments and limited price changes by adapting the commonly used marketing technique of conjoint analysis to the setting of dynamic pricing. Finally, we estimate the effectiveness of our algorithm in a randomized controlled field experiment. To the best of our knowledge, ours is the only demand learning and dynamic pricing algorithm to be deployed in a field experiment and validated in practice.

3.1.1 Literature Review

Our paper contributes to a vast literature on demand learning and dynamic pricing. For a more in-depth review of the relevant literature, we refer the reader to extensive surveys by Chen & Chen (2015) and den Boer (2015). The tension that motivates demand learning and dynamic pricing research is the classic exploration-exploitation trade-off, which requires the retailer to learn customer demand in order to identify revenue maximizing prices while minimizing revenue lost to pricing to learn rather than pricing to earn. Previous work generally balances this tradeoff in one of two ways. The first approach is learn-then-earn, where at first the retailer only prices to learn demand, and then after the retailer is sufficiently confident in the estimated demand model, the retailer prices to earn; examples include Choi et al. (2012), Witt (1986), Besbes & Zeevi (2009), and Besbes & Zeevi (2012). The second approach is continuous learning and earning, where the retailer alternates between pricing to learn and pricing to earn, balancing the learning and earning potential in each period; examples include Ferreira et al. (2018), den Boer & Zwart (2015), Qiang & Bayati (2016), and den Boer & Zwart (2014). Although the second approach of continuous learning and earning typically provides stronger theoretical guarantees, these algorithms tend to require hundreds or thousands of price changes for reasonable problem sizes to attain strong performance metrics, making them most relevant for settings where prices can be changed with high velocity.

Recent work by Russo & Van Roy (2018) is closely related to our present work

and uses Information-Directed Sampling in a multi-product setting. Similar to our algorithm when it is pricing to learn, their algorithm measures the expected information of each action and uses that to inform which action to take. Unlike our algorithm, however, they use a continuous learning and earning approach that balances the expected information gain and expected regret in each period. Their algorithm also differs from ours in that it does not incorporate context, making it difficult to generalize across similar products and assortments. Another prominent class of algorithms that limits learning across similar products and assortments is the multi-armed bandit, including the contextual bandit. Contextual bandits account for context in the sense of observing the context they are in (e.g. the type of consumer, day of week, etc.) and drawing on their model of what actions to take in that context. However, for a known context, the problem is a simple multi-armed bandit problem classified by discrete actions and no parametric model, which prevents generalization and learning across related products and assortments.

Concurrent work by Javanmard et al. (2019) operates in a nearly identical environment to our paper, though the contributions of our algorithms are distinctly different. In particular, we both assume a multi-product discrete choice setting where price sensitivities vary by product and customer demand is described by a multinomial logit model. One key difference between our settings is that we consider discrete prices with a constrained price set whereas Javanmard et al. (2019) assume continuous and unconstrained prices. The authors propose a novel continuous learning and earning pricing algorithm and show strong T-period regret.

Beyond the key modeling difference, our paper differs from theirs mainly in its focus on learning quickly and providing empirical impact estimation through a field experiment.

Our paper also builds on a vast literature in marketing on conjoint analysis. Conjoint analysis is most commonly used to construct informative surveys or choice experiments that inform product design decisions. A subset of the broader field is devoted to choice-based conjoint analysis, often assuming MNL demand. In choice-based conjoint analysis, products are characterized by a set of attributes and the researcher constructs an informative choice set that represents substitutable products with variation in their attributes. The researcher then asks respondents to select their favorite from among the set of substitutable products. Using observed choices from the informative choice sets, the researcher can then estimate customer utility for each of the product attributes. For example, a researcher interested in cell phone plans might seek to understand how customer utility is affected by the monthly plan cost and the amount of data included. The researcher would thus ask customers to choose between a set of cell phone plans that differed in the monthly cost and the amount of data included in order to estimate the corresponding utility parameters. To design informative choice experiments, like the one just mentioned, choice-based conjoint analysis generally uses principles of optimal experimental design to construct D-optimal choice experiments. D-optimal designs seek to maximize information gain as measured by the determinant of the Fisher Information matrix. A challenge of designing D-optimal choice experiments is that the optimal design is a function of the un-

known parameters. To address this challenge, marketers historically initialized the parameters at zero. Huber & Zwerina (1996) were the first to improve design efficiency by introducing a pre-experiment, which can be used to initialize parameters at more meaningful values. Sndor & Wedel (2001) built upon their work by using a Bayesian approach that incorporates managers' priors and accounts for their uncertainty. Later work by Sndor & Wedel (2005) demonstrated significant value in being able to observe choice behavior across several distinct choice sets. Our algorithm is able to incorporate prior information when it is available and updates estimates after each choice set. In this way, we are able to observe choices across a range of choice sets and are able to quickly update our parameter estimates to reflect observed demand. Our algorithm is sequential in nature, but differs from what are known as "adaptive designs" in the literature on conjoint analysis, which change the designs for a given respondent based on their previous responses (e.g. Louviere et al. (2008), Toubia et al. (2013), Cavagnaro et al. (2013), Saur & Vielma (2018)). In particular, we adapt our designs across assortments and assume consumers in each period are independently and identically distributed. Our paper is the first to integrate conjoint analysis and dynamic pricing, making a novel contribution to the demand learning and dynamic pricing literature by increasing the velocity at which learning can occur.

While our paper mostly builds on prior research in demand learning and dynamic pricing as well as in conjoint analysis, it is also related to existing work in assortment optimization. In particular, many papers in assortment optimization face the exploration-exploitation trade-off in trying to both learn customer

demand and offer optimal assortments. Many of these papers additionally assume that customer demand can be described by a multinomial logit model (see for example Agrawal et al. (2017), Saur & Zeevi (2013), Rusmevichientong et al. (2010)). Recent work by Chen et al. (2018) is closely related to ours, developing a demand learning and assortment optimization algorithm for a setting where customer demand is described by a multinomial logit function of product attributes, so learning can be generalized across similar products and assortments.

3.2 Model

We consider a retailer who sells a varying assortment of substitutable products to customers over a selling season of length T. Specifically, in each time period $t = 1, ..., T$, the retailer offers a set of N_t products; each product can be fully characterized by an observable and exogenous vector of d features, $\boldsymbol{x_i} \in \mathbb{R}^d$ for product $i \in \{1, ..., N_t\}$, where $\boldsymbol{x_i} = \{x_{1i}, ..., x_{di}\}$. The feature vectors of products offered in different periods vary and thus we use "period t" and "assortment t" interchangeably. For each assortment t, the retailer selects price vector $\boldsymbol{p_t} = \{p_{1t}, ..., p_{it}, ..., p_{N_t t}\}$ where the price p_{it} for product i is selected from a discrete set of possible prices $\mathcal{P}_{it}$. Note that prices change from period to period, but not within a single period; this reflects many retailers' interests of changing prices only when assortments change as opposed to dynamically changing prices for a fixed assortment of products. We assume zero cost per product without loss of generality and for ease of exposition; note that this implies revenue maximization

is equivalent to profit maximization. We also assume that demand can be met in each period.

We assume that a customer's expected utility from purchasing product i in period t is a linear function of its features and price:

$$u_{it} = \boldsymbol{x_i}^\intercal \boldsymbol{\beta^f} - \boldsymbol{x_i}^\intercal \boldsymbol{\beta^p} p_{it} + \epsilon_{it} \tag{3.1}$$

Here, $\boldsymbol{\beta^f}$ and $\boldsymbol{\beta^p}$ are parameters that are *unknown* to the retailer at the beginning of the season but can be learned throughout the season via purchase data; $\boldsymbol{\beta^f}$ reflects the impact of features on utility whereas $\boldsymbol{\beta^p}$ incorporates feature-specific price sensitivities. We let ϵ_{it} be a random component of utility and assume that it is drawn independently and identically from a standard Gumbel distribution. We allow for a no purchase (outside) option $i = 0$ for each assortment with $\boldsymbol{x_0} = \vec{0}$ which has utility ϵ_{0t}; we define $p_{0t} = 0$ without loss of generality and for notational convenience. We assume the customer purchases the option (including outside option) that gives her the largest utility.

This model and assumptions give us the well-known multinomial logit (MNL) discrete choice model that has been widely used in academia and in practice (see, e.g., Talluri & van Ryzin (2005) and Elshiewy et al. (2017)) and includes product features. The MNL model yields the following expression for the probability of a customer purchasing product i when offered assortment t:

$$q_{it} = \frac{exp(\boldsymbol{x_i}^\intercal \boldsymbol{\beta^f} - \boldsymbol{x_i}^\intercal \boldsymbol{\beta^p} p_{it})}{\sum_{l=0}^{N_t} exp(\boldsymbol{x_l}^\intercal \boldsymbol{\beta^f} - \boldsymbol{x_l}^\intercal \boldsymbol{\beta^p} p_{lt})} \tag{3.2}$$

We define $\boldsymbol{q_t} = \{q_{1t}, ..., q_{it}, ..., q_{N_t t}\}$.

At the end of every period, the retailer observes the quantity of each item purchased, $\boldsymbol{y_t} = \{y_{0t}, y_{1t}, ..., y_{it}, ..., y_{N_t t}\}$ where y_{it} is the quantity of product i purchased in period t, with y_{0t} representing the number of customers who do not purchase any items. It is also convenient to define $m_t = \sum_i^{N_t} y_{it}$ to be the total number of customers who arrive in period t. We assume that the number of customer arrivals in each period is independent and identically distributed, and that demand is independent across periods.

The problem faced by the retailer is to design a non-anticipatory algorithm that selects a price vector $\boldsymbol{p_t}$ for each assortment (period) $t = 1, ..., T$ in the selling season in order to maximize the total revenue over the season:

$$\max_{\boldsymbol{p_1}, \boldsymbol{p_2}, ..., \boldsymbol{p_T}} \mathrm{E}\left[\sum_{t=1}^{T} \boldsymbol{q_t}^\intercal \boldsymbol{p_t}\right] \tag{3.3}$$

where the expectation is with respect to the choice probabilities defined in Equation (3.2). To be specific, "non-anticipatory" refers to the restriction that the algorithm can use only prior periods' price, feature, and observed sales information ($\boldsymbol{p_{t'}}, \boldsymbol{x_i} \forall i \in \{1, ..., N_{t'}\}, \boldsymbol{y_{t'}} \quad \forall t' < t$) when selecting a price vector for period t. We are motivated by settings where T is relatively small -$\mathcal{O}(10)$ or $\mathcal{O}(100)$- compared to other similar work in the literature which often assumes T is orders of magnitude larger.

To summarize the sequence of events in our model, for each time period (or assortment) t,

1. The retailer selects price vector $\boldsymbol{p_t}$ and offers N_t products with feature vectors $\boldsymbol{x_i}$ for $i = 1, ..., N_t$ with price vector $\boldsymbol{p_t}$ to all customers.

2. Each customer purchases at most one item.

3. The retailer observes $\boldsymbol{y_t}$ and can use this information when choosing the price vectors for subsequent assortments.

3.3 Pricing with Fast Learning

In this section, we propose an algorithm - *Pricing with Fast Learning* - to prescribe price vector $\boldsymbol{p_t}$ in each period t after observing purchase behavior in the prior period. Given the short length of the selling season, it is important for our algorithm to learn the parameters of the utility model, $\boldsymbol{\beta^f}$ and $\boldsymbol{\beta^p}$, as quickly as possible in order to capitalize on that knowledge before the end of the season, and thus our algorithm follows the classic "learning-then-earning" approach.

Algorithm 1 formally outlines our *Pricing with Fast Learning* algorithm, discussed in more detail below. We start by initializing parameters $\hat{\boldsymbol{\beta}}^{\boldsymbol{f}}$ and $\hat{\boldsymbol{\beta}}^{\boldsymbol{p}}$ to $\vec{\mathbf{0}}$. The parameters $\hat{\boldsymbol{\beta}}^{\boldsymbol{f}}$ and $\hat{\boldsymbol{\beta}}^{\boldsymbol{p}}$ will be used as empirical estimates of the true parameters $\boldsymbol{\beta}^f$ and $\boldsymbol{\beta}^p$; in practice, the retailer could include prior information in the initializations if available. Using similar notation, we will let $\hat{q}_{it}$ be the purchase probability as defined in Equation (3.2) and given current parameter estimates $\hat{\boldsymbol{\beta}}^{\boldsymbol{f}}$ and $\hat{\boldsymbol{\beta}}^{\boldsymbol{p}}$. Following the learning-then-earning approach, we initialize our pricing method as *pricing to learn*. As an input to our algorithm, we specify a Boolean switching criteria test that will dictate when the pricing method switches to *pric-*

ing to earn; we will comment on key considerations for the switching criteria test and provide examples in Section 3.3.1.

Algorithm 1: Pricing with Fast Learning

Input: switching criteria test, $SWITCH$;
Initialize parameters: $\hat{\boldsymbol{\beta}}^{\boldsymbol{f}} = \hat{\boldsymbol{\beta}}^{\boldsymbol{p}} = \vec{0}$;
Initialize pricing method = *pricing to learn*;
for $t = 1, ..., T$ **do**
  if pricing method = *pricing to learn* **then**
    Let $\boldsymbol{z}_i = (\boldsymbol{x}_i, -\boldsymbol{x_i} p_{it})$. Offer price vector $\boldsymbol{p_t} =$
    $\arg\max_{\boldsymbol{p_t}} \det \left[\sum_{s=1}^{t} m_s \sum_{i=0}^{N_s} \left(\boldsymbol{z_i} - \sum_{l=0}^{N_s} \hat{q}_{ls} \boldsymbol{z_l} \right) \hat{q}_{is} \left(\boldsymbol{z_i} - \sum_{l=0}^{N_s} \hat{q}_{ls} \boldsymbol{z_l} \right)^{\top} \right]$ s.t. $\boldsymbol{p_t} \in \mathcal{P}_t$;
    if $SWITCH$ = TRUE **then**
      pricing method = *pricing to earn*;
    end
  end
  if pricing method = *pricing to earn* **then**
    Offer price vector
    $\boldsymbol{p_t} = \arg\max_{\boldsymbol{p_t}} \sum_{i=0}^{N_t} p_{it} \cdot \frac{exp(\boldsymbol{x_i}^{\intercal} \hat{\boldsymbol{\beta}}^{\boldsymbol{f}} - \boldsymbol{x_i}^{\intercal} \hat{\boldsymbol{\beta}}^{\boldsymbol{p}} p_{it})}{\sum_{l=0}^{N_t} exp(\boldsymbol{x_l}^{\intercal} \hat{\boldsymbol{\beta}}^{\boldsymbol{f}} - \boldsymbol{x_l}^{\intercal} \hat{\boldsymbol{\beta}}^{\boldsymbol{p}} p_{lt})}$ s.t. $\boldsymbol{p_t} \in \mathcal{P}_t$;
  end
  Observe demand $\boldsymbol{y_t}$;
  Update estimates $\hat{\boldsymbol{\beta}}^{\boldsymbol{f}}$ and $\hat{\boldsymbol{\beta}}^{\boldsymbol{p}}$ using data observed through time t
  $(\hat{\boldsymbol{\beta}}^{\boldsymbol{f}}, \hat{\boldsymbol{\beta}}^{\boldsymbol{p}}) = \arg\max_{\boldsymbol{\beta^f}, \boldsymbol{\beta^p}} \sum_{s=1}^{t} \sum_{i=0}^{N_s} y_{is} \frac{exp(\boldsymbol{x_i}^{\intercal} \boldsymbol{\beta}^{\boldsymbol{f}} - \boldsymbol{x_i}^{\intercal} \boldsymbol{\beta}^{\boldsymbol{p}} p_{is})}{\sum_{l=0}^{N_s} exp(\boldsymbol{x_l}^{\intercal} \boldsymbol{\beta}^{\boldsymbol{f}} - \boldsymbol{x_l}^{\intercal} \boldsymbol{\beta}^{\boldsymbol{p}} p_{ls})}$;
end

During the *pricing to learn* phase, the algorithm follows a pricing policy that learns efficiently by maximizing the expected information gain in each period. Specifically, the algorithm sets prices using techniques from conjoint analysis, a method that is common in the marketing literature that uses principles of optimal experimental design to maximize information gain according to a statistical

criterion. The statistical criterion we use to measure information gain is the determinant of the Fisher Information matrix, which is the most common statistical criterion for choice-based conjoint analysis. Conjoint designs that use this criterion are known as D-optimal designs, and have the intuitive appeal of minimizing the volume of the confidence ellipsoid for the parameter estimates. Therefore, in setting prices $\boldsymbol{p_t}$ to maximize the determinant of the Fisher Information matrix, our algorithm minimizes the volume of the confidence ellipsoid for the parameter estimates $\hat{\boldsymbol{\beta}}^{\boldsymbol{f}}$ and $\hat{\boldsymbol{\beta}}^{\boldsymbol{p}}$.

Proposition 1: The Fisher Information matrix for the MNL choice model is

$$I = \sum_{t=1}^{T} m_t \sum_{i=0}^{N_t} \left((\boldsymbol{x_i}, -\boldsymbol{x_i} p_{it}) - \sum_{l=0}^{N_t} q_{lt} (\boldsymbol{x_l}, -\boldsymbol{x_l} p_{lt}) \right) q_{it} \left((\boldsymbol{x_i}, -\boldsymbol{x_i} p_{it}) - \sum_{l=0}^{N_t} q_{lt} (\boldsymbol{x_l}, -\boldsymbol{x_l} p_{lt}) \right)^{\top} \tag{3.4}$$

For a proof of Proposition 1, we refer the reader to Appendix A.

Once parameter estimates are sufficiently stable, the algorithm transitions from pricing to learn to pricing to earn. Here, we price in a greedy fashion by assuming that the current parameter estimates $\hat{\boldsymbol{\beta}}^{\boldsymbol{f}}$ and $\hat{\boldsymbol{\beta}}^{\boldsymbol{p}}$ are the true parameters and maximizing revenue under this assumption. We discuss computational challenges and heuristics for both pricing methods in Section 3.3.2. Finally, regardless of the pricing method used, at the end of each period our algorithm updates the parameter estimates $\hat{\boldsymbol{\beta}}^{\boldsymbol{f}}$ and $\hat{\boldsymbol{\beta}}^{\boldsymbol{p}}$ with the current period's observed purchase data. Thus we note that even when pricing to earn, Algorithm 1 continues to passively learn and improve its parameter estimates.

3.3.1 Switching Criteria Test

Our algorithm requires the retailer to specify a switching criteria test, which is used to determine when the retailer is sufficiently confident in the estimated demand model to switch from pricing to learn to pricing to earn. There are a variety of practical considerations when choosing switching criteria. For example, it may be more effective to focus on stability in recommended prices rather than on stability in parameters, which may continue to change within a small range without affecting price recommendations. Additionally, the switching criteria may depend on the retailer's risk tolerance and how aggressive or conservative they want to be. An additional consideration is the length of the time horizon T. A retailer facing a small time horizon T would likely want a switching criteria test that switches earlier than a retailer facing a much longer horizon.

A simple switching criterion would be to pre-select the switching period τ so that the algorithm prices to learn for periods $t = 1, ..., \tau$ and then prices to earn for periods $t = \tau + 1, ..., T$. This approach is commonly used in other papers that follow a learn-then-earn approach, e.g. Besbes & Zeevi (2009) and Besbes & Zeevi (2012). Other options include requiring a certain level of stability in parameter estimates or recommended prices. For example, a valid criteria would be to check that greedy price vectors for a given assortment are consistent across current and recent parameter estimates. This approach is most similar to the one we use in our empirical application described further in Section 3.4, but we leave it to the retailer to select an appropriate switching test for their setting.

3.3.2 Computation

This subsection discusses computational challenges and solutions when computing the optimal pricing to learn and pricing to earn price vectors in Algorithm 1.

There is no closed-form method to identify the price vector $\boldsymbol{p}_t$ that maximizes the determinant of the Fisher Information matrix. Therefore, when pricing to learn, the algorithm iteratively tries all possible price vectors and selects the price vector that yields the largest determinant. To limit computational cost, it holds in memory only the best-performing price vector and corresponding determinant, replacing the values each time it finds a higher performing price vector. If the space of feasible price vectors is too large to explore, the heuristic methods of swapping, cycling, and re-labeling can be used to identify a near-optimal solution; see, e.g., Sndor & Wedel (2001) for details of these heuristics.

When pricing to earn, we follow existing literature on optimal pricing for the multinomial logit (MNL) model and make the simplifying assumption that the true model parameters are equal to our current parameter estimates. Gallego & Wang (2014) showed that when price sensitivities differ between products, optimal prices are characterized by a constant adjusted markup. In the case of the MNL demand model, they showed that optimal prices are equal to the constant adjusted markup plus the reciprocal of the product's price sensitivity; specifically, $p_i = \frac{1}{\boldsymbol{x_i}^\intercal \hat{\boldsymbol{\beta}}^{\boldsymbol{p}}} + \theta^\star$, where $\theta^\star$ is the constant adjusted markup. Unfortunately, Gallego & Wang (2014) only consider optimal pricing for an unconstrained price set. Therefore, we first find the optimal unconstrained price vector by following the method proposed in

Gallego & Wang (2014). If this price vector is outside the feasible price set, we use a grid search that chooses price vector $\boldsymbol{p_t}$ to satisfy the following optimization problem

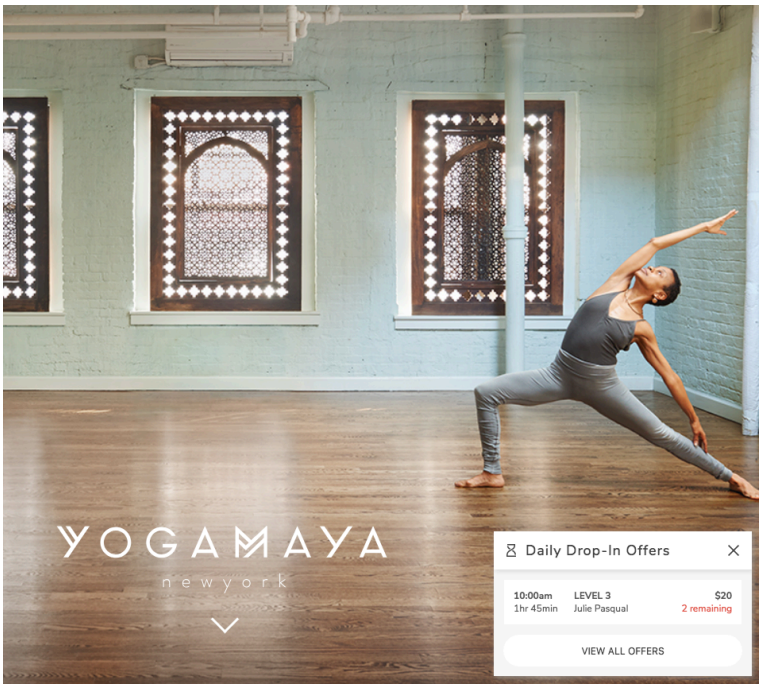
$$\boldsymbol{p_t} = \arg\max_{\boldsymbol{p_t}} \sum_{i=0}^{N_t} p_{it} \cdot \frac{exp(\boldsymbol{x_i}^\intercal \hat{\boldsymbol{\beta}}^{\boldsymbol{f}} - \boldsymbol{x_i}^\intercal \hat{\boldsymbol{\beta}}^{\boldsymbol{p}} p_{it})}{\sum_{l=0}^{N_t} exp(\boldsymbol{x_l}^\intercal \hat{\boldsymbol{\beta}}^{\boldsymbol{f}} - \boldsymbol{x_l}^\intercal \hat{\boldsymbol{\beta}}^{\boldsymbol{p}} p_{lt})} \qquad \text{s.t. } \boldsymbol{p_t} \in \mathcal{P}_t \tag{3.5}$$

3.4 Field Experiment

We collaborated with Zenrez, an e-commerce company that partners with fitness studios across the United States and Canada. Zenrez has two business lines that are relevant to our project. First, they manage customer retention at fitness studios, converting repeat customers to packs and memberships that match the customer's habits. Second, they sell excess capacity of same-day fitness classes at the fitness studios; this is the empirical setting of our project. Because Zenrez helps studios convert their repeat visitors to appropriate packs and memberships, customers who buy last-minute classes directly from Zenrez are predominantly first-time or infrequent customers. Therefore, our assumption that demand is independent over time is likely reasonable in this setting.

To sell excess capacity in fitness classes, Zenrez posts classes every night at 9pm that have remaining capacity for the following day. Zenrez sells these classes through a widget located on partner studios' webpages. For an example of the Zenrez widget, see Figure 3.1. When a user views the widget, they see all classes offered by that fitness studio.

Figure 3.1: The Zenrez widget on a yoga studio's website



The widget is shown in the bottom right hand corner of the website's home page. Our data includes everyone who expanded the widget. Clicking the widget shows the customer the full set of classes sold by Zenrez for the current day and is a necessary first step for anyone who purchases through the widget.

The class assortments change each day, and prices can vary by assortment but are fixed within assortment (e.g. once classes are posted, their prices do not change). Zenrez has flexibility to choose an integer price for each class within a studio-specified interval. The appeal of purchasing last-minute classes from Zenrez rather than directly from the studio is that Zenrez sells the classes at a discount compared to the studio drop-in rate. While a class could feasibly sell out, this happens for less than 3% of classes and thus our model assumption that all demand can be met is reasonable in this setting.

3.4.1 Experimental Design

To evaluate the effectiveness of our algorithm, we implemented a field experiment where prices for treatment studios were set according to our algorithm while prices for control studios were set according to Zenrez's baseline pricing policies. The baseline pricing practice at Zenrez was to price classes proportional to their popularity over the previous four weeks, with the most popular classes being priced at or near the ceiling of the feasible price set and the least popular classes being priced at or near the floor of the feasible price set. In order to identify studios eligible for the experiment, we first restricted to the subset of studios whose average daily revenue over a 6-month pre-period exceeded a minimum threshold. We then further required that the studios had been continuously operating the widget over the full pre-period to ensure pre-period data was reliable and uninterrupted. This process resulted in 52 eligible studios.

Within the group of eligible studios, there was significant heterogeneity in pre-

period revenue and trends, which made a simple difference-in-means or difference-in-differences evaluation unreliable. Therefore, we decided to use synthetic controls to estimate the treatment effect. Synthetic controls accounts for time trends by finding a weighted average of control studios whose trend in the pre-period closely matches the pre-period trend for the treatment studio. The metric that we chose to match trends over time was our primary metric of interest - average daily revenue. Let $\boldsymbol{Y_T}$ be a vector of average daily revenue for the six month pre-period for a treatment studio. Let $\boldsymbol{Y_C}$ be a matrix of average daily revenue for the six month pre-period for the control studios, where rows correspond to studios and columns correspond to months. Synthetic controls find the vector of weights $\boldsymbol{W^\star}$ that solves the following optimization problem

$$\boldsymbol{W^\star} = \arg \min_{||\boldsymbol{W}||=1} (\boldsymbol{Y_T} - \boldsymbol{W Y_C})'(\boldsymbol{Y_T} - \boldsymbol{W Y_C}) \tag{3.6}$$

To identify treatment studios, we calculated the synthetic control using Equation 3.6 and selected treatment studios in a greedy fashion. Specifically, among the full set of eligible studios, the studio with the strongest synthetic control (e.g. minimum error $(\boldsymbol{Y_T} - \boldsymbol{W Y_C})'(\boldsymbol{Y_T} - \boldsymbol{W Y_C})$) was designated to treatment while its control units were designated to control. Synthetic controls were then re-calculated in the same way for all remaining candidate treatment studios, and the studio with the best performing synthetic control was assigned to treatment while its control units were assigned to control. This process continued iteratively until all studios had been assigned to treatment or control, resulting in 23 treatment studios and 29 control studios. We used this technique to specify treatment and

Table 3.1: Summary statistics showing balance in treatment and control groups over a six month pre-period

	Treatment	Control
Avg. Daily Revenue	1.00	1.00
Avg. Daily Purchases	1.00	0.95
Avg. Daily Classes	1.00	0.95
Avg. Price per Class	1.00	1.06
Avg. Widget Traffic	1.00	1.08
Unique Cities*	11	13

All numbers except unique cities are normalized to the treatment studios' pre-treatment levels. The control group is a weighted average using synthetic controls, where Equation 3.6 was applied with $\boldsymbol{Y_T}$ averaged over all treatment studios. *Number of unique cities are reported without normalizing or weighting.

control studios in order to best guarantee that we would be able to evaluate the outcome of the experiment using synthetic controls.

The fitness studios in our sample covered a range of cities, neighborhoods, and types of fitness classes. Given this diversity, we felt it was appropriate to estimate a separate multinomial logit (MNL) demand model for each fitness studio. Demand was estimated as a function of class duration, indicators for the day of week (Monday through Saturday), two time of day indicators (class start time before 8:30am or after 5pm), and a vector of price sensitivities. Price sensitivities were obtained by interacting price with a number of class attributes, including the two aforementioned time of day indicator variables, indicators for the type of exercise class (e.g. yoga, pilates, etc.), and whether the instructor was popular, where popularity was measured in the pre-period using average bookings per class taught by the instructor. Table 3.1 shows balance across treatment and control studios in a variety of summary statistics over the six month pre-period.

We ran the experiment for three months, during which time prices for the treatment studios were set according to our algorithm while prices for the control studios were set according to Zenrez's standard pricing practices. Both treatment and control studios maintained existing price restrictions, namely that prices had to be integer values within a range set by the fitness studios. Studios did not know that they were included in the experiment.

Our algorithm requires a switching criteria test as an input. For the experiment with Zenrez, we define the test as a test of price stability. Specifically, we look for whether greedy prices for two assortments have changed at all over the past three periods. If the greedy prices have not changed, then the parameter estimates are sufficiently stable and the algorithm switches from pricing to learn to pricing to earn. We use two assortments to ensure robustness. The first assortment is the actual assortment to be offered in the next period. However, since that assortment may be small or represent only a small fraction of the full feature set, we generate a second assortment that samples 10 classes from the full set of unique attribute combinations. While the next day's assortment will vary for each period, the second assortment is fixed across all periods. Since we check for stability in greedy prices using estimates from the previous three periods, we require three periods of demand to be observed before checking whether the algorithm should switch from pricing to learn to pricing to earn.

In our experiment with Zenrez, the last demand for assortment t may be observed up until very close to the time that prices for assortment $t + 1$ need to be posted. Since our algorithm takes a non-trivial amount of time to run, we

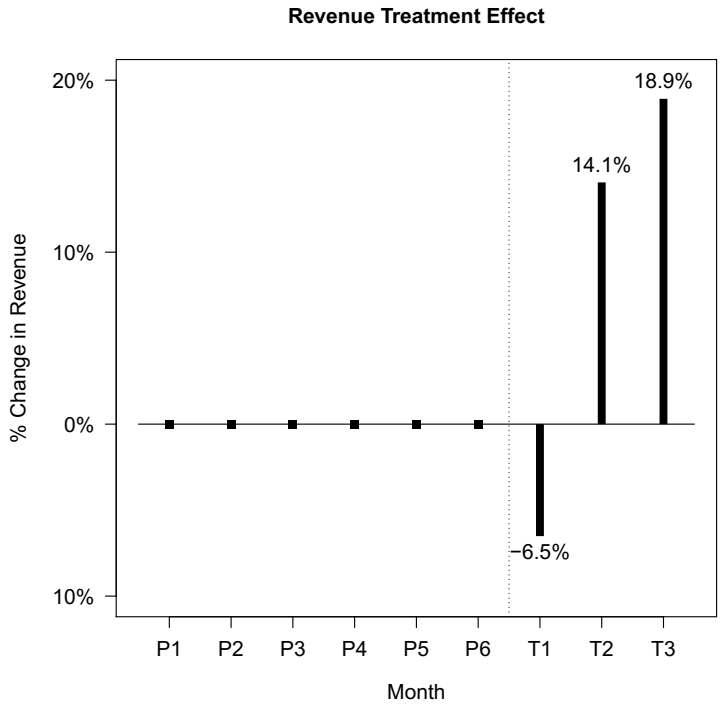
therefore used demand through assortment $t-1$ to set prices for assortment $t+1$, introducing a one period lag. Our switching criteria test requires three periods of observed demand before running, so given the one-period lag in our implementation with Zenrez, assortment $t = 5$ is the first assortment eligible to be priced under pricing to earn.

3.4.2 Results

As described in Section 3.4.1, we evaluated the effect of our algorithm using synthetic controls. When applying synthetic controls to multiple treatment units (studios), it is best practice to average the treatment units together and then generate a synthetic control for the average treatment unit (Abadie et al., 2010). Therefore, we set $\boldsymbol{Y_T}$ as the average daily revenue across all treatment studios for the six month pre-period and used Equation 3.6 to identify a synthetic control whose revenue trajectory closely matched the revenue trajectory of the average treatment studio over the six-month pre-period.

To estimate the treatment effect, we then compared average daily revenue over all treatment studios to its synthetic control for each month in the three-month experiment. The effect sizes we report have been normalized to protect the confidentiality of Zenrez's revenue data. Specifically, all revenue numbers have been divided by the average daily revenue for the treatment group during the six month pre-period, so that they represent percentage effects rather than dollar effects. A percentage effect of 10% should therefore be interpreted as a dollar effect of 10% $\times$ (average daily revenue for the treatment group over the six month pre-period).

Figure 3.2: Revenue difference between the average treated unit and the synthetic control



P1-P6 represent the six pre-period months while T1-T3 represent the three months of our experiment. The start of the experiment is indicated by a vertical dotted line. The treatment effect is normalized by dividing the effect in dollars by average revenue over the six-month pre-period. The algorithm led to a short-term dip in revenue when pricing to learn followed by an increase in revenue when pricing to earn relative to a synthetic control whose prices were set according to baseline practices.

Figure 3.2 illustrates the results of our experiment. First, we can see that we were able to attain a very strong synthetic control; there is a near perfect match in average daily revenue for each of the six pre-period months between treatment and synthetic control studios. Because of this, we attribute the difference between average daily revenue during the three month experiment to the impact of our algorithm. Compared to the synthetic control, the treatment studios in our sample experienced a dip in average daily revenue of 6.5% in the first month of our experiment and an increase in average daily revenue of 14.1% and 18.9% in the second and third months of our experiment, respectively. We note that this dip followed by gain is consistent with our expectations that initial pricing to learn would yield a dip in revenue while subsequent pricing to earn would lead to higher revenue in the long run. Over the three month experiment, treatment studios experienced an 8.8% increase in average daily revenue compared to the synthetic control. Since the revenue gains were persistent at well above 10% across the second and third months, it is reasonable to expect that gains of similar magnitudes would endure over future periods if the algorithm were run for longer.

Randomization Inference with Fisher's Exact Test

In order to quantify the probability of observing our results under the null hypothesis that our algorithm had no effect on revenue, we performed Randomization Inference with Fisher's Exact Test; see Ho & Imai (2006) for details on this test. Let $s = 1, ..., S$ index the studios in our experiment, including both treatment and control studios. Under the potential outcomes framework, each studio has

two potential outcomes. The first potential outcome, r_s^T, is the studio's revenue in the state of the world where that studio is a treated unit, e.g. the studio's prices are set by our algorithm. The second potential outcome, r_s^C, is the studio's revenue in the state of the world where the studio does not receive treatment, e.g. the studio's prices are set according to baseline practices. Treatment assignment is represented by a binary variable D_s, which equals 1 when studio s is assigned to treatment and equals 0 otherwise. Therefore, studio s's realized revenue is $r_s = D_s \times r_s^T + (1 - D_s) \times r_s^C$.

Under the null hypothesis used by Fisher's Exact Test, $r_s^T = r_s^C$, i.e. the treatment effect is zero for all units. This null hypothesis is called the *sharp* null hypothesis, because it applies to the unit-level treatment effect rather than to the average treatment effect across units. Using this framework, the potential outcomes r_s^T and r_s^C for all units are known exactly and the only source of randomness is the treatment assignment vector $\boldsymbol{D}$. In the most common application of Randomization Inference, we would calculate the following test statistic, which corresponds to the difference-in-means estimator for the average treatment effect

$$B(\boldsymbol{D}) = \frac{\sum_{s=1}^{S} D_s r_s}{\sum_{s=1}^{S} D_s} - \frac{\sum_{s=1}^{S} (1 - D_s) r_s}{\sum_{s=1}^{S} (1 - D_s)}$$

However, since we are using synthetic controls to estimate our treatment effect, we replace the difference-in-means estimator with the synthetic controls estimator.

$$C(\boldsymbol{D}) = \frac{\sum_{s=1}^{S} D_s r_s}{\sum_{s=1}^{S} D_s} - \sum_{s=1}^{S} w_s^{\star} (1 - D_s) r_s$$

where $\boldsymbol{W}^\star = (w_1^\star, ..., w_S^\star)$ corresponds to the vector of synthetic control weights with slight abuse of notation. In particular, for studio s such that $D_s = 0$, $w_s^\star$ is from Equation 3.6; for studio s such that $D_s = 1$, we define $w_s^\star = 0$. Note that $\boldsymbol{W}^\star$ varies for each treatment assignment vector $\boldsymbol{D}$.

Since the vector of treatment assignments $\boldsymbol{D}$ is the only source of randomness, the randomization distribution of $\boldsymbol{D}$ completely determines the reference distribution of the test statistic $C(\boldsymbol{D})$. Calculating the test statistic $C(\boldsymbol{D})$ for the full set of treatment assignment permutations yields the exact distribution of C under the sharp null hypothesis, which can then be used to calculate an exact (one-tailed) p-value, according to the following formula

$$p \equiv Pr(C(\boldsymbol{D}) \geq C(\boldsymbol{D}^\star)) \tag{3.7}$$

where $\boldsymbol{D}^\star$ is the actual treatment assignment vector used in the experiment. The null hypothesis of no treatment effect will be rejected if the p-value is less than a pre-determined significance level (e.g. 0.10). This test is *exact* in the sense that it does not depend on large sample approximation and is *distribution-free* because it does not depend on any distributional assumptions.

Given the size of our sample, there are 52 choose 23 or approximately 3.5×10^{14} possible treatment permutations, making it impractical to calculate the test statistic $C(\boldsymbol{D})$ for all possible treatment permutations. Instead, we sampled without replacement from the full set of treatment assignment vectors 50,000 times to approximate the distribution of $C(\boldsymbol{D})$. We chose 50,000 samples, because we found it to be a large enough sample size that results were highly consistent across different

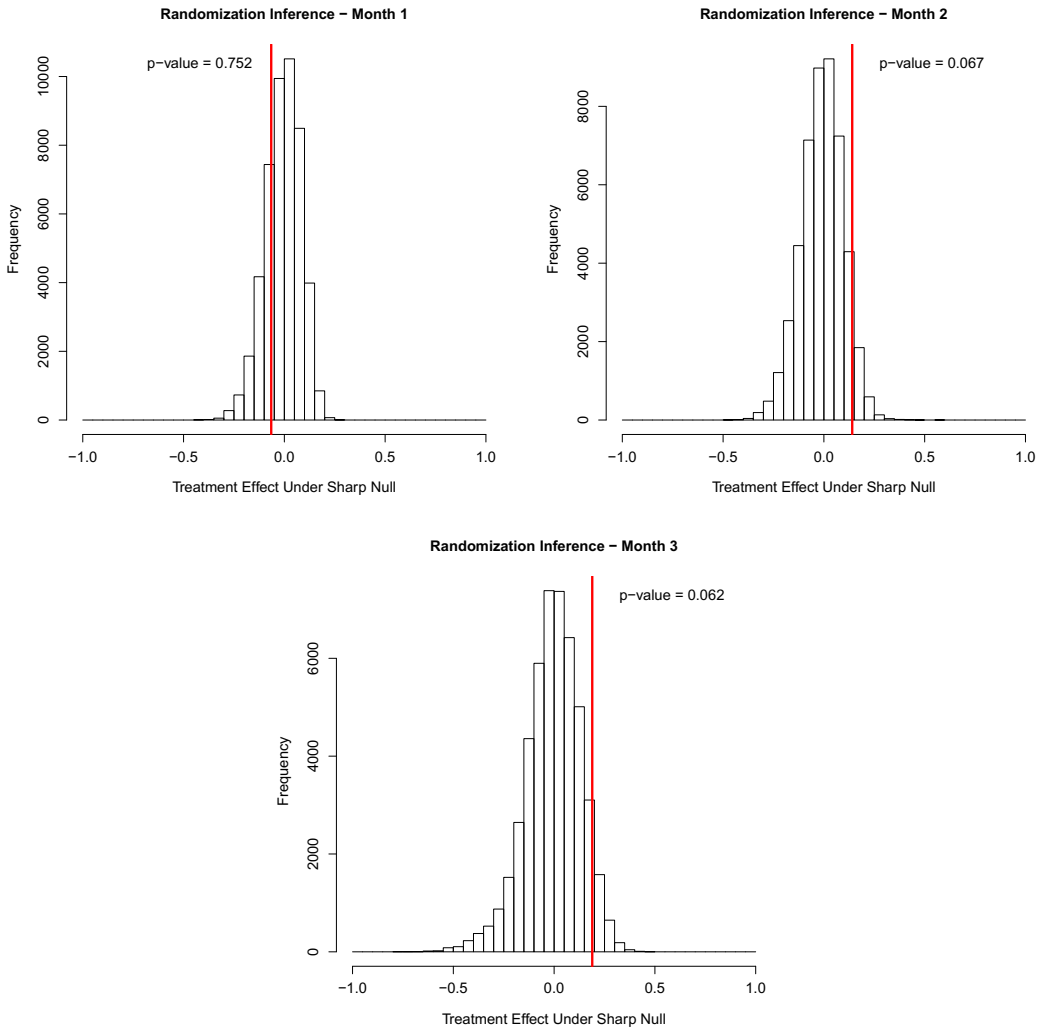
random initializations while computation time was still reasonably fast.

For each sampled treatment assignment vector $\boldsymbol{D}_j$ for $j = 1, ..., 50,000$, we averaged across the treatment group to generate an average treatment unit and then ran synthetic controls, matching on monthly revenue over the six months preceding the experiment. This process yielded the vector $\boldsymbol{W}_j^\star$ of synthetic control weights, which we used to calculate $C(\boldsymbol{D}_j)$ for each of the three months of the experiment. Note that this is the same procedure we used to estimate the monthly treatment effects for the true experiment. In order to be consistent with the true randomization process, we discarded all samples for which the synthetic control match over the pre-period was poor (e.g. because the highest revenue studios were all assigned to treatment and thus no control match was available), resulting in a loss of about 3% of all samples.

This procedure yielded a distribution for $C(\boldsymbol{D})$ for each of the three months of the experiment, allowing us to quantify the likelihood of observing both the initial dip in revenue and the subsequent increases in revenue under the sharp null hypothesis that the algorithm had no effect on revenue relative to baseline pricing practices. The results are displayed in Figure 3.3 along with their accompanying one-tailed p-values, which are defined in Equation 3.7.

From the results in Figure 3.3, we conclude that the initial dip in revenue is unsurprising under the null hypothesis while the subsequent revenue lift in Months 2 and 3 would be quite unlikely under the null hypothesis. Specifically, we find that the 6.5% decline in revenue during Month 1 (relative to the synthetic control) has an accompanying one-tailed p-value of 0.75. Flipping the p-value for easier

Figure 3.3: Randomization inference results for each month of the experiment



Randomization Inference results using Fisher's Exact Test show the distribution of normalized revenue treatment effects under the sharp null, with the actual treatment effect represented by a red vertical line with the corresponding p-value adjacent. Revenue treatment effects are normalized by dividing the effect in dollars by the average revenue for the treatment studios over the six months pre-experiment.

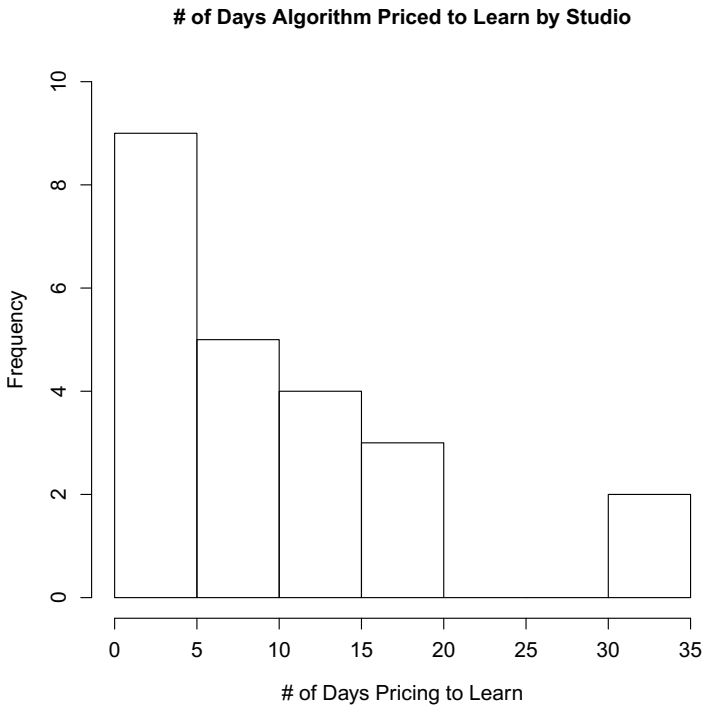
interpretation, we conclude that if the sharp null hypothesis is true, a decline in revenue even larger than 6.5% should be expected 25% of the time. Therefore, we cannot reject the null in Month 1. In Months 2 and 3, treatment revenue exceeded synthetic control revenue by 14.1% and 18.9%, respectively, with corresponding one-tailed p-values of 0.067 and 0.062. Therefore, under the sharp null hypothesis, we would expect to see a greater increase in revenue for each month only around 6-7% of the time. We thus have sufficient evidence to reject the null hypothesis at the 10% significance level and conclude that our algorithm had a strong positive effect on revenue.

Understanding Algorithm Behavior

Figure 3.4 shows the distribution of days spent pricing to learn by each treatment studio. The majority of studios switched from pricing to learn to pricing to earn within 10 days of the algorithm's launch, while two studios took just over 30 days to make the switch. Importantly, we see that our algorithm only required a very short pricing to learn stage and was quickly able to capitalize on that learning in the pricing to earn stage. This is a promising result for many retailers who may not want to or may not be able to frequently change prices: our results show that minimal price experimentation early on can reap huge payoffs as the season progresses.

To better understand the effect of our algorithm, we analyzed prices and sales for the treatment studios during the nine month period that included both the six-month pre-period and the three-month experiment. Figure 3.5 shows the average

Figure 3.4: Distribution of days spent pricing to learn by studio



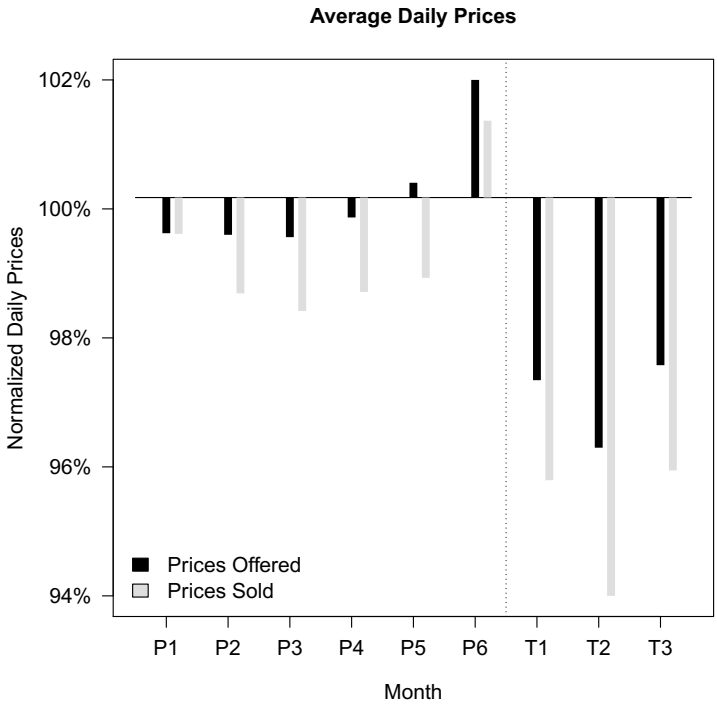
The distribution of days spent pricing to learn by studio. Most studios switched to pricing to earn within 10 days while two studios took more than 30 days to switch.

price of offered classes as well as the average selling price for each of the nine months. To protect confidentiality, we normalized both prices offered and prices sold by the average price offered over the six month pre-period. We can see that the average price of an offered class decreased by approximately 3% during the experiment compared to the six month pre-period. Demand was larger for the cheaper classes, and thus we see a decrease by approximately 4% in the average selling price during the experiment. Figure 3.6 shows that this decrease of 4% in average selling price resulted in an increase of approximately 16% in units sold over the full experiment and an increase of approximately 26% in the second and third months of the experiment, when the algorithm was primarily pricing to earn, leading to an overall positive impact on revenue. Finally, Figure 3.7 shows the average variance in daily prices. Our algorithm increased price variance, especially in the first month when it was predominately pricing to learn. The algorithm's behavior when pricing to learn is unsurprising, as conjoint analysis commonly results in features being set at or near the boundaries and generally leads to an even split between the upper and lower limits, resulting in greatest price variance (Kanninen, 2002).

3.5 Conclusion

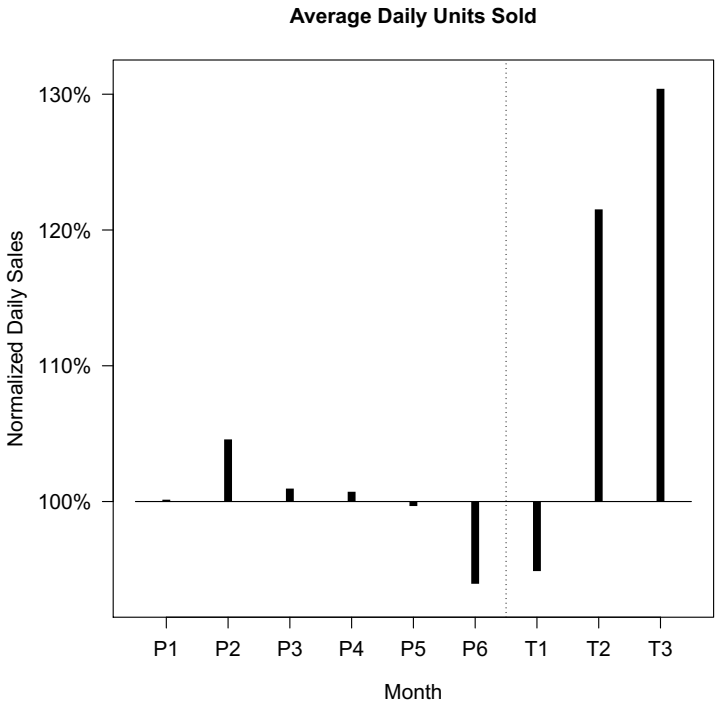
In this paper, we introduced a novel algorithm - *Pricing with Fast Learning* - that efficiently learns customer demand to maximize revenue over a finite horizon T. Our algorithm parameterizes demand for products as a function of their attributes

Figure 3.5: Average prices of all classes offered as well as of just the classes sold



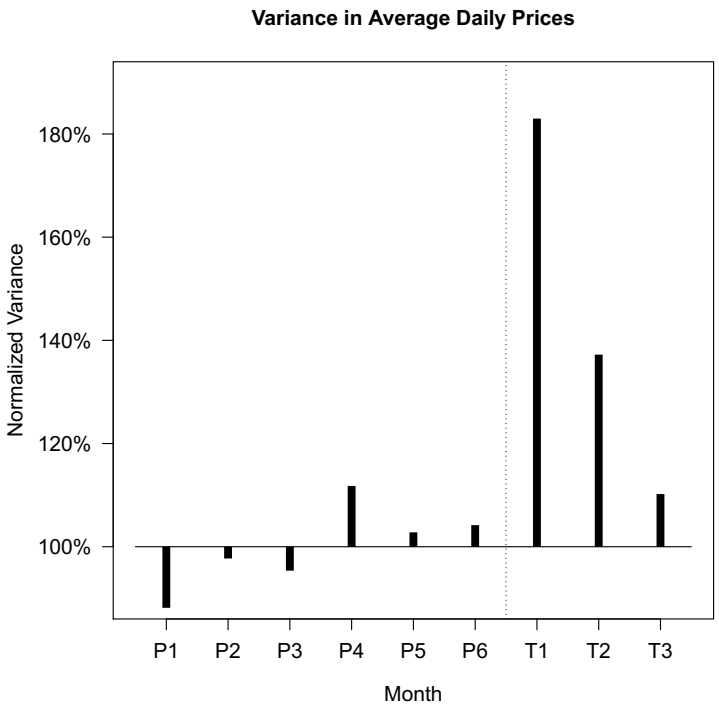
Effect of Algorithm 1 on prices for treatment studios. Black: The average price of all classes offered, normalized by the average price of pre-period classes offered. Gray: The average price of classes sold, normalized by the average price of pre-period classes offered. Time periods P1-P6 represent the six pre-period months and T1-T3 represent the three months of our experiment. The start of the experiment is indicated by a dashed vertical line. It is evident from the figure that our algorithm lowered average prices.

Figure 3.6: Number of average daily purchases, normalized to pre-treatment average



Average daily purchase volumes for the treatment studios, normalized by their value over the six-month pre-period. Time periods P1-P6 represent the six pre-period months and T1-T3 represent the three months of our experiment. The start of the experiment is indicated by a dashed vertical line.

Figure 3.7: Variance in average daily prices, normalized to pre-treatment average



The variance in average daily prices for treatment studios, normalized to the mean variance over the six-month pre-period. Time periods P1-P6 represent the six pre-period months and T1-T3 represent the three months of our experiment. The start of the experiment is indicated by a dashed vertical line.

using the popular multinomial logit demand model, and thus is well-suited for the setting where retailers frequently rotate assortments and need to generalize learning across similar products or assortments. We bridge two distinct literatures on dynamic pricing and conjoint analysis by adapting techniques from conjoint analysis to set prices when pricing to learn. This increases the velocity of learning, making our algorithm perform well for small T. We validate its performance in a field experiment, the first of its kind, where we show that our algorithm quickly leads to net revenue gain over baseline pricing practices despite causing a short initial dip in revenue.

There are several areas of future research that would increase the applicability of our algorithm and improve its performance. First, to improve run-time and enable the algorithm to handle a large number of products and prices, an efficient optimization routine for identifying D-optimal prices is needed. Second, additional theory is needed to understand how the constant adjusted-markup result of Gallego & Wang (2014) extends to settings with constrained prices. Lastly, it would be valuable to develop a better understanding of which switching criteria are optimal for each of the settings commonly encountered in the literature and in practice.

Appendix A

Proof of Proposition 1

From Equation 3.2, the probability that a customer selects choice i from the assortment is defined as

$$q_{it} = \frac{exp(\boldsymbol{x_i}^\intercal \boldsymbol{\beta^f} - \boldsymbol{x_i}^\intercal \boldsymbol{\beta^p} p_{it})}{\sum_{l=0}^{N_t} exp(\boldsymbol{x_l}^\intercal \boldsymbol{\beta^f} - \boldsymbol{x_l}^\intercal \boldsymbol{\beta^p} p_{lt})}$$

Where $i = 0$ represents the outside option with utility equal to zero. Later, we will need the derivative of the choice probabilities with respect to $(\boldsymbol{\beta^f}, \boldsymbol{\beta^p})^\intercal$ (the

transpose of the coefficient vector). The derivative of the choice probabilities is

$$\begin{aligned}
\frac{\partial q_{it}}{\partial(\boldsymbol{\beta^f}, \boldsymbol{\beta^p})^\intercal} &= \frac{(\boldsymbol{x_i}, -\boldsymbol{x_i}p_{it})^\intercal exp(\boldsymbol{x_i}^\intercal\boldsymbol{\beta^f} - \boldsymbol{x_i}^\intercal\boldsymbol{\beta^p}p_{it})}{\sum_{l=0}^{N_t} exp(\boldsymbol{x_l}^\intercal\boldsymbol{\beta^f} - \boldsymbol{x_l}^\intercal\boldsymbol{\beta^p}p_{lt})} - \\
&\quad \frac{exp(\boldsymbol{x_i}^\intercal\boldsymbol{\beta^f} - \boldsymbol{x_i}^\intercal\boldsymbol{\beta^p}p_{it}) \sum_{l=0}^{N_t}(\boldsymbol{x_l}, -\boldsymbol{x_l}p_{lt})^\intercal exp(\boldsymbol{x_l}^\intercal\boldsymbol{\beta^f} - \boldsymbol{x_l}^\intercal\boldsymbol{\beta^p}p_{lt})}{\left(\sum_{l=0}^{N_t} exp(\boldsymbol{x_l}^\intercal\boldsymbol{\beta^f} - \boldsymbol{x_l}^\intercal\boldsymbol{\beta^p}p_{lt})\right)^2} \\
&= (\boldsymbol{x_i}, -\boldsymbol{x_i}p_{it})^\intercal q_{it} - q_{it}\sum_{l=0}^{N_t}(\boldsymbol{x_l}, -\boldsymbol{x_l}p_{lt})^\intercal q_{lt} \\
&= q_{it}\left((\boldsymbol{x_i}, -\boldsymbol{x_i}p_{it})^\intercal - \sum_{l=0}^{N_t}(\boldsymbol{x_l}, -\boldsymbol{x_l}p_{lt})^\intercal q_{lt}\right) \\
&= q_{it}\left((\boldsymbol{x_i}, -\boldsymbol{x_i}p_{it}) - \sum_{l=0}^{N_t}(\boldsymbol{x_l}, -\boldsymbol{x_l}p_{lt}) q_{lt}\right)^\intercal
\end{aligned}$$

The log-likelihood requires finding the log of the choice probabilities q_{it}. Taking the log of the choice probabilities q_{it} yields

$$\ln q_{it} = \boldsymbol{x_i}^\intercal\boldsymbol{\beta^f} - \boldsymbol{x_i}^\intercal\boldsymbol{\beta^p}p_{it} - \ln\left[\sum_{l=0}^{N_t} exp(\boldsymbol{x_l}^\intercal\boldsymbol{\beta^f} - \boldsymbol{x_l}^\intercal\boldsymbol{\beta^p}p_{lt})\right]$$

Let y_t^k be a one-hot vector of length $N_t + 1$ representing customer k's choice. Specifically, suppose the customer purchases option i, then y_t^k has a one in position i (position index starts at 0) and zeros in all other positions. Therefore, for customer k in period t

$$y_{it}^k = \begin{cases} 1 & \text{if customer } k \text{ selects item } i \\ 0 & \text{otherwise} \end{cases}$$

As in Section 3.2, define m_t as the number of customers in period t. Additionally, define m_{it} as the number of customers in period t who select option i, e.g. $m_{it} = \sum_{k=1}^{m_t} y_{it}^k$. The likelihood function is thus

$$\mathcal{L} = \prod_{t=1}^{T} \prod_{k=1}^{m_t} \prod_{i=0}^{N_t} q_{it}^{y_{it}^k} \tag{A.1}$$

And the log-likelihood is accordingly

$$\ln \mathcal{L} = \sum_{t=1}^{T} \sum_{k=1}^{m_t} \sum_{i=0}^{N_t} y_{it}^k \ln q_{it} \tag{A.2}$$

The gradient of the log-likelihood is

$$\begin{aligned}
\frac{\partial \ln \mathcal{L}}{\partial(\boldsymbol{\beta^f}, \boldsymbol{\beta^p})} &= \sum_{t=1}^{T} \sum_{k=1}^{m_t} \sum_{i=0}^{N_t} \left(y_{it}^k - q_{it}\right) \left(\boldsymbol{x_i}^\intercal, -\boldsymbol{x_i}^\intercal p_{it}\right) \\
&= \sum_{t=1}^{T} \sum_{k=1}^{m_t} \sum_{i=0}^{N_t} y_{it}^k (\boldsymbol{x_i}^\intercal, -\boldsymbol{x_i}^\intercal p_{it}) - \sum_{t=1}^{T} \sum_{k=1}^{m_t} \sum_{i=0}^{N_t} q_{it} (\boldsymbol{x_i}^\intercal, -\boldsymbol{x_i}^\intercal p_{it}) \\
&= \sum_{t=1}^{T} \sum_{i=0}^{N_t} m_{it} (\boldsymbol{x_i}^\intercal, -\boldsymbol{x_i}^\intercal p_{it}) - \sum_{t=1}^{T} \sum_{i=0}^{N_t} m_t q_{it} (\boldsymbol{x_i}^\intercal, -\boldsymbol{x_i}^\intercal p_{it}) \\
&= \sum_{t=1}^{T} \sum_{i=0}^{N_t} \left(m_{it} - m_t q_{it}\right) \left(\boldsymbol{x_i}^\intercal, -\boldsymbol{x_i}^\intercal p_{it}\right)
\end{aligned}$$

The Hessian is the matrix of second derivatives of the log-likelihood and can be

found by differentiating the gradient with respect to $(\boldsymbol{\beta^f}, \boldsymbol{\beta^p})^\intercal$.

$$\frac{\partial \ln \mathcal{L}}{\partial(\boldsymbol{\beta^f}, \boldsymbol{\beta^p})(\boldsymbol{\beta^f}, \boldsymbol{\beta^p})^\intercal} = -\sum_{t=1}^{T} m_t \sum_{i=0}^{N_t} \left((\boldsymbol{x_i}, -\boldsymbol{x_i} p_{it}) - \sum_{l=0}^{N_t} q_{lt} (\boldsymbol{x_l}, -\boldsymbol{x_l} p_{lt}) \right) q_{it} \left((\boldsymbol{x_i}, -\boldsymbol{x_i} p_{it}) - \sum_{l=0}^{N_t} q_{lt} (\boldsymbol{x_l}, -\boldsymbol{x_l} p_{lt}) \right)^\top$$

For the multinomial logit model, the Fisher Information matrix is defined as

$$I = -E\left[\frac{\partial \ln \mathcal{L}}{\partial(\boldsymbol{\beta^f}, \boldsymbol{\beta^p})(\boldsymbol{\beta^f}, \boldsymbol{\beta^p})^\intercal} \right] \tag{A.3}$$

Therefore, the Fisher Information matrix is

$$I = \sum_{t=1}^{T} m_t \sum_{i=0}^{N_t} \left((\boldsymbol{x_i}, -\boldsymbol{x_i} p_{it}) - \sum_{l=0}^{N_t} q_{lt} (\boldsymbol{x_l}, -\boldsymbol{x_l} p_{lt}) \right) q_{it} \left((\boldsymbol{x_i}, -\boldsymbol{x_i} p_{it}) - \sum_{l=0}^{N_t} q_{lt} (\boldsymbol{x_l}, -\boldsymbol{x_l} p_{lt}) \right)^\top \tag{A.4}$$

The asymptotic covariance matrix of the MNL parameter estimates is exactly equal to the Fisher Information matrix. These asymptotic estimates are appropriate even in relatively small samples. A D-optimal design maximizes the determinant of the Fisher Information matrix and results in shrinking the confidence ellipsoid of the parameter estimates. Therefore, when choosing D-optimal prices, we want to choose prices $\boldsymbol{p}$ to maximize

$$\left| \sum_{t=1}^{T} m_t \sum_{i=0}^{N_t} \left((\boldsymbol{x_i}, -\boldsymbol{x_i} p_{it}) - \sum_{l=0}^{N_t} q_{lt} (\boldsymbol{x_l}, -\boldsymbol{x_l} p_{lt}) \right) q_{it} \left((\boldsymbol{x_i}, -\boldsymbol{x_i} p_{it}) - \sum_{l=0}^{N_t} q_{lt} (\boldsymbol{x_l}, -\boldsymbol{x_l} p_{lt}) \right)^\top \right|$$

References

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of Californias Tobacco Control Program. *Journal of the American Statistical Association*, 105(490), 493–505.

Agrawal, S., Avadhanula, V., Goyal, V., & Zeevi, A. (2017). MNL-Bandit: A Dynamic Learning Approach to Assortment Selection. *arXiv:1706.03880 [cs]*. arXiv: 1706.03880.

Anglin, P. M. (1997). Determinants of Buyer Search in a Housing Market. *Real Estate Economics*, 25(4), 567–589.

Askitas, N. (2016). Trend-Spotting in the Housing Market. *Cityscape*, 18(2), 165–178.

Bailey, M., Cao, R., Kuchler, T., & Stroebel, J. (2018). The Economic Effects of Social Networks: Evidence from the Housing Market. *Journal of Political Economy*, 126(6), 2224–2276.

Beracha, E. & Wintoki, N. B. (2013). Forecasting Residential Real Estate Price Changes from Online Search Activity. *Journal of Real Estate Research*, 35(3), 283–312.

Besbes, O. & Zeevi, A. (2009). Dynamic Pricing Without Knowing the Demand Function: Risk Bounds and Near-Optimal Algorithms. *Operations Research*, 57(6), 1407–1420.

Besbes, O. & Zeevi, A. (2012). Blind Network Revenue Management. *Operations Research*, 60(6), 1537–1550.

Bettinger, E. (2004). How Financial Aid Affects Peristence. In *College Choices: The Economics of Where to Go, When to Go, and How to Pay For It* (pp. 207–237). University of Chicago Press.

Cavagnaro, D. R., Gonzalez, R., Myung, J. I., & Pitt, M. A. (2013). Optimal Decision Stimuli for Risky Choice Experiments: An Adaptive Approach. *Management Science*, 59(2), 358–375.

Chen, M. & Chen, Z.-L. (2015). Recent Developments in Dynamic Pricing Research: Multiple Products, Competition, and Limited Demand Information. *Production and Operations Management*, 24(5), 704–731.

Chen, X., Wang, Y., & Zhou, Y. (2018). Dynamic Assortment Optimization with Changing Contextual Information. *arXiv:1810.13069 [cs, econ, stat]*. arXiv: 1810.13069.

Choi, H. & Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, 88, 2–9.

Choi, T.-M., Chow, P.-S., & Xiao, T. (2012). Electronic price-testing scheme for fashion retailing with information updating. *International Journal of Production Economics*, 140(1), 396–406.

Deming, D. & Dynarski, S. (2009). *Into College, Out of Poverty? Policies to Increase the Postsecondary Attainment of the Poor*. Technical Report w15387, National Bureau of Economic Research, Cambridge, MA.

den Boer, A. V. (2015). Dynamic pricing and learning: Historical origins, current research, and new directions. *Surveys in Operations Research and Management Science*, 20(1), 1–18.

den Boer, A. V. & Zwart, B. (2014). Simultaneously Learning and Optimizing Using Controlled Variance Pricing. *Management Science*, 60(3), 770–783.

den Boer, A. V. & Zwart, B. (2015). Dynamic Pricing and Learning with Finite Inventories. *Operations Research*, 63(4), 965–978.

Dynarski, S. (2008). Building the Stock of College-Educated Labor. *Journal of Human Resources*, 43(3), 576–610.

Dynarski, S. & Scott-Clayton, J. (2013). *Financial Aid Policy: Lessons from Research*. Technical Report Working Paper 18710, NBER.

Dynarski, S. M. (2003). Does Aid Matter? Measuring the Effect of Student Aid on College Attendance and Completion. *American Economic Review*, 93(1), 279–288.

Egloffstein, M. & Ifenthaler, D. (2017). Employee Perspectives on MOOCs for Workplace Learning. *TechTrends*, 61(1), 65–70.

Elshiewy, O., Guhl, D., & Boztug, Y. (2017). Multinomial Logit Models in Marketing - From Fundamentals to State-of-the-Art. *Marketing ZFP*, 39(3), 32–49.

Ferreira, K. J., Simchi-Levi, D., & Wang, H. (2018). Online Network Revenue Management Using Thompson Sampling. *Operations Research*, 66(6), 1586–1602.

Gallego, G. & Wang, R. (2014). Multiproduct Price Optimization and Competition Under the Nested Logit Model with Product-Differentiated Price Sensitivities. *Operations Research*, 62(2), 450–461.

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4), 665–676.

Glaeser, E. L., Gyourko, J., Morales, E., & Nathanson, C. G. (2014). Housing dynamics: An urban approach. *Journal of Urban Economics*, 81, 45–56.

Glaeser, E. L., Kim, H., & Luca, M. (2018). Nowcasting Gentrification: Using Yelp Data to Quantify Neighborhood Change. *SSRN Electronic Journal*.

Hansen, J. D. & Reich, J. (2015). Democratizing education? Examining access and usage patterns in massive open online courses. *Science*, 350(6265), 1245–1248.

Ho, D. E. & Imai, K. (2006). Randomization Inference With Natural Experiments: An Analysis of Ballot Effects in the 2003 California Recall Election. *Journal of the American Statistical Association*, 101(475), 888–900.

Huber, J. & Zwerina, K. (1996). The Importance of Utility Balance in Efficient Choice Designs. *Journal of Marketing Research*, 33(3), 307.

Javanmard, A., Nazerzadeh, H., & Shao, S. (2019). Multi-Product Dynamic Pricing in High-Dimensions with Heterogenous Price Sensitivity. *arXiv:1901.01030 [cs, stat]*. arXiv: 1901.01030.

Kanninen, B. J. (2002). Optimal Design for Multinomial Choice Experiments. *Journal of Marketing Research*, 39(2), 214–227.

Kelly, D. & Teevan, J. (2003). Implicit feedback for inferring user preference: a bibliography. *ACM SIGIR Forum*, 37(2), 18–28.

Leslie, L. L. & Brinkman, P. (1993). *The economic value of higher education: Larry L. Leslie and Paul T. Brinkman*. American Council on Education/Oryx Press series on higher education. [Washington, D.C.] : Phoenix, AZ: American Council on Education ; Oryx Press.

Louviere, J. J., Street, D., Burgess, L., Wasi, N., Islam, T., & Marley, A. A. (2008). Modeling the choices of individual decision-makers by combining efficient choice experiment designs with extra preference information. *Journal of Choice Modelling*, 1(1), 128–164.

Mendoza, P., Mendez, J. P., & Malcolm, Z. (2009). Financial Aid and Persistence in Community Colleges: Assessing the Effectiveness of Federal and State Financial Aid Programs in Oklahoma. *Community College Review*, 37(2), 112–135.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17(8), 857–872.

Qiang, S. & Bayati, M. (2016). Dynamic Pricing with Demand Covariates. *SSRN Electronic Journal*.

Rae, A. (2015). Online Housing Search and the Geography of Submarkets. *Housing Studies*, 30(3), 453–472.

Rashidi, T. H., Auld, J., & Mohammadian, A. K. (2012). A behavioral housing search model: Two-stage hazard-based and multinomial logit approach to choice-set formation and location selection. *Transportation Research Part A: Policy and Practice*, 46(7), 1097–1107.

Reich, J. (2014). *MOOC Completion and Retention in the Context of Student Intent*. Technical report, Educause Review.

Reich, J. & Ruiperez-Valiente, J. A. (2019). The MOOC Pivot. *Science*, 363(6423), 130–131.

Richardson, P. (2019). Nowcasting and the Use of Big Data in Short-Term Macroeconomic Forecasting: A Critical Review. *Economie et Statistique / Economics and Statistics*, (505d), 65–87.

Richburg Hayes, L., Brock, T., LeBlanc, A., Paxson, C. H., Rouse, C. E., & Barrow, L. (2009). Rewarding Persistence: Effects of a Performance-Based Scholarship Program for Low-Income Parents. *SSRN Electronic Journal*.

Rosendale, J. A. (2017). Gauging the value of MOOCs: An examination of American employers perceptions toward higher education change. *Higher Education, Skills and Work-Based Learning*, 7(2), 141–154.

Rusmevichientong, P., Shen, Z.-J. M., & Shmoys, D. B. (2010). Dynamic Assortment Optimization with a Multinomial Logit Choice Model and Capacity Constraint. *Operations Research*, 58(6), 1666–1680.

Russo, D. & Van Roy, B. (2018). Learning to Optimize via Information-Directed Sampling. *Operations Research*, 66(1), 230–252.

Saur, D. & Vielma, J. P. (2018). Ellipsoidal methods for adaptive choice-based conjoint analysis. *Operations Research*.

Saur, D. & Zeevi, A. (2013). Optimal Dynamic Assortment Planning with Demand Learning. *Manufacturing & Service Operations Management*, 15(3), 387–404.

Smith, T. R. & Clark, W. A. V. (1982). Housing Market Search Behavior and Expected Utility Theory: 1. Measuring Preferences for Housing. *Environment and Planning A*, 14(5), 681–698.

Smith, T. R., Clark, W. A. V., Huff, J. O., & Shapiro, P. (1979). A Decision-Making and Search Model for Intraurban Migration. *Geographical Analysis*, 11(1), 1–22.

Sndor, Z. & Wedel, M. (2001). Designing Conjoint Choice Experiments Using Managers Prior Beliefs. *Journal of Marketing Research*, 38(4), 430–444.

Sndor, Z. & Wedel, M. (2005). Heterogeneous Conjoint Choice Designs. *Journal of Marketing Research*, 42(2), 210–218.

Talluri, K. T. & van Ryzin, G. J. (2005). *Theory and Practice of Revenue Management*. Springer-Verlag.

Togan-Egrican, A., English, C., & Klapper, L. (2012). *Credit Cards and Formal Loans Rare in Developing Countries*. Technical report, Gallup.

Toubia, O., Johnson, E., Evgeniou, T., & Delqui, P. (2013). Dynamic Experiments for Estimating Preferences: An Adaptive Method of Eliciting Time and Risk Parameters. *Management Science*, 59(3), 613–640.

Witt, U. (1986). How can complex economic behavior be investigated? The example of the ignorant monopolist revisited. *Behavioral Science*, 31(3), 173–188.

Wu, L. & Brynjolfsson, E. (2015). The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. In *Economic Analysis of the Digital Economy* (pp. 89–118). University of Chicago Press.

Yang, S., Kou, S. C., Lu, F., Brownstein, J. S., Brooke, N., & Santillana, M. (2017). Advances in using Internet searches to track dengue. *PLOS Computational Biology*, 13(7), e1005607.

Yang, S., Santillana, M., & Kou, S. C. (2015). Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences*, 112(47), 14473–14478.