From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Economics

2022

Essays on Retail Electricity Markets and Methods for Analyzing Matched Data

Anderson, Rachel S

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01d217qs73p

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

ESSAYS ON RETAIL ELECTRICITY MARKETS AND METHODS FOR ANALYZING IMPERFECTLY MATCHED DATA

Rachel Sara Anderson

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY
THE DEPARTMENT OF
ECONOMICS
Advisers: Janet Currie and Bo Honoré

September 2022

© Copyright by Rachel Sara Anderson, 2022. All rights reserved.

Abstract

My dissertation consists of two chapters in energy economics and two in econometrics.

Chapter 1 studies the relationship between retail electricity prices and household income in Connecticut. Using novel data sets constructed from files published in the Connecticut regulatory authority's online docket system, I show that customers living in the state's poorest zip codes paid significantly more for competitive retail electric service than customers living in above median income zip codes between 2011 and 2018. These differences may be explained by the fact that low-income customers purchased from more expensive retailers, which, coupled with preferences for value-add products such as renewable energy, led them to choose contracts with higher prices.

Chapter 2 examines the role of regulatory policy in deregulated retail electricity markets in the United States using hand-collected data on whether and when a market adopted three pro-competitive policies. While all three policies increased the share of customers choosing to purchase competitive retail electric service, the estimated effect size of policies differ. The policies also had different effects on competitive retailers' entry decisions and prices, such that only one policy plausibly increased the consumer welfare of residential customers.

Chapter 3, which is co-authored with Bo Honoré and Adriana Lleras-Muney, proposes methods for analyzing merged data sets where observations are linked to multiple possible matches. We show that it is straightforward to modify standard econometric models to accommodate such data, and that, when observations are weighted according to the number of matches, the resulting estimator performs well under a variety of data generating processes.

Chapter 4 evaluates the combined performance of methods for merging data sets and estimating linear regression models when the variables of interest are recorded in separate data files. I study four matching methods, that either do or do not allow for multiple matches per observation, and that are either deterministic or probabilistic in nature. Simulation results demonstrate that allowing for multiple matches per observation can produce more

precise estimates, provided the appropriate estimation method is used, and that the choice between using a deterministic or probabilistic method in such cases is less important.

Acknowledgements

First and foremost, I would like to thank my advisors, Bo Honoré and Janet Currie, for their unwavering support through every step of the dissertation process.

To Bo Honoré, thank you for your mentorship from start to end of the PhD. From providing my first introduction to econometrics research, to reviewing results, presentations, and manuscripts for my energy economics papers—I could not have asked for a better academic advisor.

To Janet Currie, thank you for advising me during the challenging COVID-19 circumstances. Our regular meetings were invaluable for helping me integrate my learning about new topics in energy economics. Your feedback and mentorship have helped to refine my thinking and writing, and, ultimately, make me a stronger economist.

To Laura Hedden, thank you for helping me navigate the administrative complexities of graduate school. I appreciate and admire your service to the department. The support that you have shown me is no exception.

I am also grateful to numerous other Princeton faculty and post-docs, whose conversations provided direction and helped me stay motivated through a long research and writing process: Leah Boustan, Swati Bhatt, Smita Brunnermeier, Natalie Cox, Kate Ho, Jesse Jenkins, Jakub Kastl, Ilyana Kuziemko, David Lee, Alessandro Lizzeri, Ulrich Müller, and Esmée Zwiers.

My heartfelt thanks to David Banks and Timur Kuran, who nurtured my love of statistics and economics research during my undergraduate studies at Duke University, and invited me to be their colleague and friend ever since.

I am deeply privileged to have been welcomed into so many social communities while at Princeton. To affiliates of the Princeton Salsa Club, Kopp's Cycle, Big Bangers, Princeton Energy and Climate Scholars, and the unnamed social groups that have included me in celebrations, excursions, and informal hangouts over the years: a tremendous thank you.

Finally, to the family and friends who supported me, I love you so very much. Nothing expresses my gratitude for you better than *ubuntu*: I am because we are. To my brother Ben, thank you for calling and for challenging me to graduate before you retire. Mom and Dad, thank you for being my first teachers and biggest fans. This work is dedicated to the both of you.

To Mom and Dad.

Contents

Abstract . . . 3

Acknowledgements . . . 5

List of Tables . . . 12

List of Figures . . . 14

1 When Competition Hurts: Evidence from Connecticut's Retail Electricity Market **15**

1.1 Introduction . . . 15

1.2 Background . . . 18

1.2.1 Electric Market Restructuring in the United States . . . 18

1.2.2 History of Connecticut's Retail Electricity Market . . . 20

1.2.3 Electric Generation Service Contracts in Connecticut . . . 21

1.3 Data . . . 23

1.3.1 Data Overview . . . 23

1.3.2 Zip Code-Level Data . . . 25

1.3.3 Contract Length Data . . . 28

1.3.4 Green Customer Data . . . 30

1.3.5 Data Limitations . . . 31

1.4 Methods . . . 32

1.4.1 Do low-income customers pay significantly higher prices? . . . 32

1.4.2 Does inertia explain why low-income customers pay more? . . . 35

1.4.3 Do low-income customers choose to pay more? . . . 36
1.4.4 Do low-income customers pay more for price stability? . . . 37
1.4.5 Do low-income customers pay more for renewable energy? . . . 39
1.5 Results . . . 40
1.5.1 Do low-income customers pay significantly higher prices? . . . 40
1.5.2 Does inertia explain why low-income customers pay more? . . . 41
1.5.3 Do low-income customers choose to pay more? . . . 42
1.5.4 Do low-income customers pay more for price stability? . . . 43
1.5.5 Do low-income customers pay more for renewable energy? . . . 44
1.6 Discussion . . . 45
1.7 Conclusion . . . 47
1.8 Figures and Tables . . . 49
1.9 Data Appendix . . . 64

2 Do Regulators Make Markets? The Effects of Regulatory Policy on Retail Electricity Markets 70

2.1 Introduction . . . 70
2.2 Background . . . 73
2.2.1 Background on U.S. Electricity Sector Restructuring . . . 73
2.2.2 Market-Based Default Service Rates . . . 76
2.2.3 Purchase of Receivables Rules . . . 77
2.2.4 Opt-Out Aggregation Programs . . . 79
2.3 Data . . . 80
2.3.1 Key Variables . . . 82
2.4 Methods . . . 83
2.4.1 Effects of Policies on Competitive Retail Market Size . . . 83
2.4.2 Effects of Policies on Competitive Retailer Entry and Exit . . . 87

2.4.3 Effects of Policies on Competitive Retailer Prices . . . 89
2.5 Results . . . 90
2.5.1 Effects of Policies on Competitive Retail Market Size . . . 90
2.5.2 Effects of Policies on Retailer Entry and Exit . . . 91
2.5.3 Effects of Policies on Competitive Retailer Prices . . . 92
2.6 Discussion . . . 93
2.7 Conclusion . . . 95
2.8 Figures and Tables . . . 96

3 Estimation and inference using imperfectly matched data 112
3.1 Introduction . . . 112
3.2 Data Generating Process . . . 115
3.3 Linear Instrumental Variables Estimation . . . 116
3.3.1 Identification . . . 117
3.3.2 Estimation . . . 119
3.3.3 Controlling for Estimation of $g(\cdot)$ . . . 124
3.4 Monte Carlo Evidence . . . 126
3.4.1 Data Generating Process . . . 126
3.4.2 Estimation . . . 127
3.5 Possible Generalizations . . . 129
3.5.1 Estimating Nonlinear Models . . . 129
3.5.2 Using Information on Match Quality . . . 131
3.6 Conclusion . . . 132
3.7 Figures and Tables . . . 134

4 Evaluating the joint performance of methods for matching and analyzing linked data 146

4.1 Introduction . . . 146
4.2 Setup . . . 149
4.3 Review of Methods Tested . . . 151
4.3.1 Matching Methods . . . 151
4.3.2 Estimation Methods for Analyzing Linked Data . . . 154
4.4 Monte Carlo . . . 157
4.4.1 Overview of the Design . . . 158
4.4.2 Data Generation Step . . . 159
4.4.3 Matching Step . . . 160
4.4.4 Estimation Step . . . 162
4.5 Results . . . 163
4.5.1 Results from Matching Step . . . 163
4.5.2 Results from Estimation Step . . . 165
4.6 Conclusion . . . 167
4.7 Figures and Tables . . . 169

Bibliography **181**

List of Tables

1.1 Overview of Data Set Periods Covered 51

1.2 Overview of Data Set Variables 51

1.3 Summary of Zip Code-Level Data for Above and Below Median Income Zip Codes 52

1.4 Summary of Zip Code-Level Data for June and September 2018 53

1.5 Summary of Contract Length Data by Customer Type 54

1.6 Summary of Green Customer and Zip Code-Level New Customer Data . . 55

1.7 Relationship between Zip Code Income and Electricity Contract Prices . . 56

1.8 Relationship between Zip Code Income and Electricity Contract Prices Over Time 57

1.9 Difference in Initial and Current Contract Prices by Customer Income . . 58

1.10 Probability of enrolling in a contract that costs more than the default option 59

1.11 Relationship between contract price, length, and customer hardship status 60

1.12 Summary of Estimation Sample for Contract Length Data 61

1.13 Relationship between customer income, contract prices, and renewable energy content 62

1.14 Relationship between a customer's income and whether they choose a green energy contract 63

1.15 Notes on Data Sources Used 64

2.1 State Retail Electric Restructuring Policy Dates 99

2.2 Summary Statistics for Key Variables by Customer Class 100
2.3 Estimated Effects of Policies on Customer Switching Share 101
2.4 Estimated Effects of Policies on Sales Switching Share 102
2.5 Results for Delayed Effects of Policies on Customer Switching Share 103
2.6 OLS Results for Effect of Policies on Retailer Entry and Exit 104
2.7 OLS Results for Effect of Policies on Retailer Prices 105
2.8 Notes on Opt-out Aggregation Program Authorization Dates Used 106

3.1 Baseline Results with Large Sample g Estimates 137
3.2 Baseline Results with Small Sample g Estimates 137
3.3 Results for Modification (a) with Large Sample g Estimates 138
3.4 Results for Modification (a) with Small Sample g Estimates 138
3.5 Results for Modification (b) with Large Sample g Estimates 139
3.6 Results for Modification (b) with Small Sample g Estimates 139
3.7 Results for Modification (c) with Large Sample g Estimates 140
3.8 Results for Modification (c) with Small Sample g Estimates 140

4.1 Overview of Matching Methods 174
4.2 Summary of matching algorithm performance 174
4.3 Performance of multiple match methods by value of L_i 174
4.4 Median Absolute Deviations for Estimators 175

List of Figures

1.1 Competitive Electric Generation Service Contract Prices Over Time 49

1.2 Competitive Electric Generation Service Contract Prices by Zip Code Income 50

2.1 Restructured Wholesale Electricity Markets in the United States 96

2.2 Restructured Retail Electricity Markets in the United States 97

2.3 Competitive Retailer Prices by Customer Class 98

3.1 g Implied by the Baseline Monte Carlo Design 134

3.2 g Implied by Modification (a) to the Baseline Design 135

3.3 g Implied by Modification (b) to the Baseline Design 136

4.1 Creation of Synthetic Datasets 170

4.2 Monte Carlo Distribution of Benchmark OLS Estimators 171

4.3 Match Rates by Linking Procedure 172

4.4 Monte Carlo Distribution of the SW and AHL Estimators 173

Chapter 1

When Competition Hurts: Evidence from Connecticut's Retail Electricity Market[1]

1.1 Introduction

Governments worldwide have deregulated industries to enhance the economic efficiency of markets and lower prices for consumers. Yet, whether consumers benefit from deregulation measures that require consumers to choose a competitive supplier depends on consumers' willingness and ability to search for the lowest prices and avoid low quality products. If search frictions are high or product quality is difficult to observe, then deregulation of such markets may not increase consumer welfare.

While there is a robust empirical literature to suggest that deregulation of the U.S. power sector has reduced the cost of generating electricity and improved the efficiency of power systems, there is limited evidence that deregulation has lowered retail prices for consumers (Bushnell et al., 2017). To the contrary, recent papers suggest that retail

[1]I thank Andrew Minikowski, John Viglione, Susan Baldwin, and the Connecticut Office of the Consumer Counsel staff whose assistance made this project possible. The conclusions drawn in this work are those of the researcher and do not reflect the views of the State of Connecticut nor any of its agencies.

prices increased as a result of restructuring (Dormady et al., 2019; MacKay and Mercadal, 2021; Upton et al., 2021). A key complication in performing such analyses is the fact that states' deregulation efforts were not uniform—in addition to establishing competitive wholesale electricity markets, a subset of states also deregulated their retail electricity markets. In those states, consumers pay different prices according to their choice of retail electric provider, and so attempts to measure the welfare gains from deregulation must also consider the distributional consequences of decentralized price-setting.

Despite the fact that energy expenditures represent a large share of a household's income, research on price dispersion in competitive U.S. retail electricity markets is lacking, because data on prices paid by individual customers are usually confidential. In recent years, several states have initiated investigations to review such data. Their findings indicate that many customers—and especially low-income customers—who choose their retail electric provider end up paying more than if they had consumed the same amount of electricity under the regulated, default option (Baldwin, 2018).

This paper studies the relationship between retail electricity prices and household income in Connecticut. In Connecticut, residential customers, by default, pay regulated electricity prices for retail electric service from their distribution utility; however, many customers choose to pay market-based prices for service from competitive retailers. Surprisingly, data from this market show that over half of customers who choose their retailer pay higher prices than their utility's regulated default option (Office of Consumer Counsel, 2020).[2] Additionally, utility "hardship" customers—defined as those customers who have demonstrated that they face a significant financial or medical burden—paid over $7.2 mil-

[2]Between 2015-2019, competitive retailer customers collectively paid $240 million more than if they had remained on their utility's regulated default option.

lion more than they would have paid under their distribution utilities' regulated default option between 2016-2018.[3]

Using the publicly available data from this investigation that identify both the price that customers pay and the zip code in which they live, I find that Connecticut households living in below median income zip codes paid higher electricity prices than households living in above median income zip codes between 2011-2018. While similar patterns have been observed in other deregulated retail electricity markets, such trends are usually explained by the fact that low-income customers face higher search costs or greater inertia in those markets (Byrne et al., 2022; Byrne and Martin, 2021; Hortaçsu et al., 2017; Giulietti et al., 2010). These explanations cannot fully explain why low-income customers pay higher prices in Connecticut's electricity market, however, because over half of low-income customers actively chose to pay more for electricity than if they had made no choice at all.

The fact that customers living in some of Connecticut's poorest zip codes chose to pay more for electricity than customers in above median income zip codes may indicate that they have a higher willingness to pay for the non-price attributes of competitive generation service. These attributes include whether a customer's contract offers renewable energy content exceeding the state's renewable portfolio standard, long term price stability, or gift cards upon enrollment. Although my data are insufficient to rule out this explanation entirely, my analysis suggests that it is unlikely. Instead, my findings are consistent with qualitative evidence from state agencies that customers living in low-income areas were targeted with in-person marketing activities that persuaded them to enroll in more expensive contracts.

The findings from this paper have important implications for evaluating the effectiveness of utility payment programs, the Low Income Heating and Energy Assistance Pro-

[3]See the Connecticut Public Utilities Regulatory Authority decision submitted on December 18, 2019, in Docket No. 18-06-02, *Review of Feasibility, Costs, and Benefits of Placing Certain Customers on Standard Service Pursuant to Conn. Gen Stat. § 16-245o(m).*

gram (LIHEAP), and other state energy subsidy programs. For example, if low-income households pay more for electricity without receiving any commensurate benefit, then subsidies to these households amount to a transfer from taxpayers to competitive electric retailers. This argument has led policymakers in Connecticut and New York to limit the ability of energy subsidy recipients to participate in competitive retail electricity markets. To the extent that the results from this paper generalize to other markets—as numerous state investigations suggest that they do—policymakers should consider taking similar actions to protect vulnerable customers.

A brief outline of the paper is as follows. Section 1.2 provides a brief background on electricity market restructuring in the United States and retail electric competition in Connecticut. Section 1.3 describes the data, which are used to document several remarkable facts about the retail electricity prices paid by low-income customers using the methods described in Section 1.4. Section 1.5 presents the results, which are discussed in Section 1.6. Section 1.7 concludes.

1.2 Background

1.2.1 Electric Market Restructuring in the United States

The electric supply industry consists of four main functions: electric generation, transmission, distribution, and retail sales. For most of the 20th century, vertically integrated, monopoly utilities provided bundled electric service to virtually all U.S. households and charged regulated prices set to reflect the cost of providing all four components of electric service.

Inspired by the deregulation of other network industries such as natural gas, airlines, and telecommunications, several U.S. states enacted legislation to restructure their electricity markets in the late 1990s. States' restructuring efforts aimed to introduce market-

based prices and encourage the entry of competitive firms in the electric generation and retail sales functions of the electric industry.[4] Proponents of restructuring contended that competition in electric generation would reduce generation costs and result in lower wholesale electricity prices, and competition in electric retailing would allow those cost savings to be passed through to consumers.

A defining characteristic of restructured retail electricity markets in the United States is that residential and small commercial customers are permitted but not required to choose a competitive electric retailer. In such markets, a customer's monthly electric bill is divided into separate charges for electric generation and distribution services, and the customer may sign a contract with a competitive retailer that determines how much they pay for the electric generation portion of their bill. A customer who does not enter into any contract with a competitive retailer continues to pay regulated prices for a default option that is usually provided by the customer's distribution utility (which additionally provides distribution service to the customer at regulated prices). The procurement and design of the default generation service tariff is overseen by state regulators, such that prices are set to reflect the cost of purchasing electricity in the wholesale market plus a markup to soften competition with entrant, non-utility retailers. All customers, regardless of whether they choose to purchase competitive generation service, continue to pay regulated prices for distribution service from their utility.

The share of residential customers who choose to purchase competitive electric generation service varies widely across the 13 U.S. states that restructured their retail electricity markets. Connecticut is an interesting state to study because of the size of its competitive retail electricity market—about half of Connecticut households purchased competitive retail electric products in 2014—and the strength of its consumer protection laws, which limit the complexity of contracts that retailers can sell to residential customers. The next

[4]Distribution and transmission continued to be viewed as natural monopolies.

section describes the history and design of Connecticut's retail electricity market in further detail.

1.2.2 History of Connecticut's Retail Electricity Market

In 1998, Connecticut enacted legislation that required the state's two investor-owned electric utilities, Eversource Energy (Connecticut Light and Power d/b/a Eversource Energy) and The United Illuminating Company, to divest their generation assets, and opened the state's retail electricity market to competition. Although the market for residential customers opened to competition in 2000, only one non-utility retailer served customers until 2007. To encourage competitive retailer entry, the state's public utility commission established new rules for setting the utilities' default generation service prices and paying retailers who billed their customers via the utilities' billing system in 2007 and 2008, respectively. Together, these pro-competitive regulatory policies helped to increase the share of customers purchasing competitive electric generation service, from 6 percent in 2007 to nearly 50 percent in 2014.

The Connecticut market underwent another period of regulatory reform during 2014-2015. State legislators enacted two consumer protection laws in response to an increase in customer complaints about competitive retailers' lack of price transparency and deceptive marketing practices during winter 2013-2014.[5] The first law, Public Act 14-75, aimed to increase customers' awareness of their contract terms by displaying information about their contract, along with information about their utilities' default generation service option, on the first page of their monthly electric bills. The law also lowered the maximum fee that competitive retailers could charge their customers for terminating their contracts early, from $150 to $50. The second law, Public Act 15-90, prohibited retailers from sell-

[5]Many customers complained that they were not informed when their fixed-price contracts automatically renewed to variable rate contracts. These customers learned that their contract terms had changed during winter 2013-2014, when wholesale electricity prices skyrocketed and resulted in prices that were 200-300 percent higher than the utilities' default electric generation service prices at the time.

ing or automatically renewing their customers to contracts with prices that were fixed for less than four months at a time. This law only applied to competitive electric generation service contracts sold after October 1, 2015. Customers who entered contracts prior to this date were allowed to remain on their contracts, even if their contract terms did not meet the new standard.

The consumer protection laws also empowered regulators to punish competitive retailers who failed to comply with the state's marketing guidelines and to investigate whether firms were targeting low-income customers with more expensive contracts.[6] Between 2015 and 2018, state regulators imposed civil penalties for marketing violations on more than half a dozen competitive retailers and revoked the license of at least one large retailer. In 2018, state regulators opened an investigation to review whether the utilities' hardship customers—who are protected from electric shut-off during winter months because they face a significant, documented financial or medical burden—should be returned to their utility's default generation service option. The data analyzed in this paper primarily come from files submitted by Eversource Energy to aid with this investigation.[7]

1.2.3 Electric Generation Service Contracts in Connecticut

Residential customers in Connecticut pay some of the highest retail electricity prices in the United States (U.S. Energy Information Administration, 2020).[8] Electric generation service charges account for about half of retail electricity prices, while regulated distribution and transmission charges, public policy costs, and taxes account for the other half. The state's investor-owned distribution utilities, Eversource Energy and The United Illuminating Company, provide a default generation service option for residential and small

[6]Connecticut General Statutes §16-245o(m)

[7]See Public Utilities Regulatory Authority Docket 18-06-02.

[8]In general, electricity prices in New England states are among the highest in the U.S. because of their reliance on natural gas for winter heating and electricity generation. Additionally, the regulated portion of retail electricity prices in Connecticut include tariffs for state environmental policies and a 10-year contract with the Millstone nuclear power plant (Singer, 2021).

commercial customers that is known as the "standard service" rate. The standard service rate consists of a single, marginal price for every kilowatt-hour of electricity that a customer consumes during their billing cycle. Standard service prices are fixed for six months at a time, between January 1-June 30 and July 1-December 31 of each year, and changes are announced 45 days in advance.

To facilitate comparison with the standard service rate, all competitive generation service contracts in Connecticut must consist of a single price per kilowatt-hour of electricity that a customer consumes during their billing cycle. Contracts may differ according to the number of months for which the price is fixed, the renewable energy content of the electricity, and whether they include any cancellation, enrollment, or monthly fixed fees.

Customers can view current standard service prices and shop for competitive electric generation service on *EnergizeCT.com*, the state's official resource for helping households manage their energy use. *EnergizeCT.com* was launched in 2013, replacing the state's previous websites for posting information on competitive retail electric service and other residential energy topics in Connecticut. *EnergizeCT.com* includes a shopping tool that allows customers to compare the cost of different generation service contracts based on their historical electric usage. For each contract that is posted online, the tool calculates the monthly savings or additional cost compared to the cost of the customer's utility's current standard service option, and users can view the number of customer complaints made about the seller in previous years.

In addition to posting offers online, retailers advertise their contracts to customers through door-to-door visitations, telesales, and advertising on social media. In recent years, Connecticut regulators have investigated several retailers for using deceptive marketing practices, such as misrepresenting savings, pretending to be affiliated with the customer's distribution utility, or saying that enrollment is mandatory. Evidence from these investigations suggest that the contracts sold in person and via telesales activities

are more expensive than the contracts posted on *EnergizeCT.com*. Additionally, state consumer advocates have expressed the concern that retailers disproportionately target low-income households with their marketing activities.

Data on the method by which customers enroll in competitive generation service contracts is limited. Interviews with state representatives suggest that only a small fraction of customers enroll in contracts via the *EnergizeCT.com* online shopping tool. Reports on *EnergizeCT.com* website use confirm that the site received around 100,000 monthly visits between 2014-2018, however the majority of those views were for pages other than the shopping tool. There is no information available about the demographics of *EnergizeCT.com* users or whether low-income customers are more or less likely to use the price comparison tool.

1.3 Data

This section describes the three data sets that are used for the analysis. The data sets were constructed using publicly available data files from the Connecticut Public Utilities Regulatory Authority website. Details on the specific files used are reported in the appendix.

1.3.1 Data Overview

This subsection provides an overview of the three data sets and the main variables of interest.

The first data set is a repeated cross section of all Eversource Energy residential customers who purchased competitive electric generation service during the months of June and September between 2011 and 2018. The data include the price that each customer pays as well as the zip code in which they live, hence I refer to these data as the zip code-

level data set. I use the zip code-level data to assess whether customers living in poor zip codes pay higher prices than customers living in affluent zip codes.

The second data set is also a repeated cross section of residential customers' competitive generation service contract choices between October 2016 and March 2019. These data are recorded on a monthly basis and include a dummy variable for new customer accounts, as well as other variables that allow me to estimate the length of customers' contracts. I refer to these data as the "contract length" data set.

The third data set contains information on the renewable energy content of contracts sold by two of Connecticut's largest competitive electric retailers between 2017 and 2018. I use these data to measure whether a customer purchases a green energy contract, and so I refer to them as the "green customer" data set.

Tables 1.1 and 1.2 provide an overview of the three data sets, including the time period they cover and which variables they contain. For each data set, the unit of observation is a residential customer i who purchases a competitive generation service contract from a competitive retailer r_i during month t_i. For confidentiality reasons, the data sets do not contain any variables to identify unique contracts nor unique customers. As a result, any linking of the data sets is performed using the name of a customer's competitive retailer, their contract price, and the customer's "type", τ_i.

Customers are divided into four types, depending on whether they use electricity as their main heating source and whether they are a "hardship" customer. Electric heating customers use about twice as much electricity as non-heating customers during winter months, or about 1.5 times more electricity annually. Hardship customers are protected from having their electric service terminated due to nonpayment during winter months because they face a documented financial or medical burden. In total, about 15 percent of Eversource Energy customers used electric heating and 7.5 percent of customers were hardship customers between 2011-2018.

The main outcome of interest is a customer's electric generation service contract price, which is reported in cents per kilowatt-hour (kWh). I convert contract prices from cents per kilowatt-hour into the cost (in dollars) of consuming 750 kWh, which is the default monthly electric usage for comparing offers on *EnergizeCT.com*.[9] For contracts beginning in October 2015 or later, I calculate the cost of consuming 750 kWh usage per month for the first four months of the contract, because all contracts sold after that date had to offer fixed prices for at least four months at a time.

Because all residential customers in Connecticut are eligible to purchase a default generation service product from their distribution utility, I also calculate the difference between a customer's contract price and the contemporaneous price of the default service option. I refer to this value as a customer's monthly markup or savings relative to the current default service option. Positive values are assigned to customers who pay more than they would have paid under their utility's default service option. I obtain data on Eversource Energy's past default service prices from company filings and *EnergizeCT.com*.

1.3.2 Zip Code-Level Data

The first data set is a repeated cross section of all Eversource Energy residential customers who purchased electric generation service from a competitive retailer during the months of June and September between 2011 and 2018. For each customer i observed at time t_i, the data identify the customer's competitive retailer r_i and the price that they pay in cents per kilowatt-hour for all usage billed during the current month. The data set also includes the customer's zip code and whether a customer used electric heating or was a hardship customer (e.g., the customer's type). Additionally, observations from June and

[9]Competitive electric retailers in Connecticut are not allowed to sell contracts that use nonlinear pricing to residential customers, so the cost of consuming 750 kWh equals the customer's contract price multiplied by 750.

September 2018 include a dummy variable equal to one for customers who began their electric generation service contract during the current month.

Table 1.3 presents summary statistics for the zip code-level customer accounts data, computed separately for customers living in above and below median income zip codes. Below median income zip codes include zip codes for which the median household income was below $80,000 in the American Community Survey estimates from 2015-2019. The sample is split into two periods: June 2011-September 2015 and June 2016-September 2018, corresponding with the periods before and after the implementation of the 2014-2015 consumer protection laws.

On average, customers living in below median income zip codes paid higher prices than customers living in above median income zip codes across all years of the sample. The difference between the average price paid by customers living in above and below median income zip codes increased slightly between 2011-2015 and 2016-2018, from $1.79 per month to $3.36 per month. Additionally, the share of customers paying more than their utility's default generation service price increased over time, from 52 percent to 68 percent of customers in below median income zip codes and from 47 percent to 57 percent of customers in above median income zip codes. Customers paid significant markups for competitive electric generation service: between 2015-2018, the median customer in a below median income zip code paid $8.37 more per month than the default service option and over 25 percent of customers paid an additional cost of $17.51 or more per month. For comparison, the average monthly cost of the Eversource Energy default service option was about $70 during the sample period.

Figure 1.1 plots the distribution of competitive generation service contract prices and the cost of the Eversource default service option over time. The average competitive retailer customer paid more than the default generation service price across all months of the sample, except for July 2014-January 2015. During those months, the Eversource de-

fault service price was set at a record high price to allow the utility to recover its generation service costs from winter 2013-2014, when abnormal weather conditions caused wholesale electricity prices in New England to skyrocket.

Figure 1.2 plots contract prices separately for customers living in above and below median income zip codes. The median price was about the same for both zip code groups until 2014, after which the median price for the lower income group was higher. The median price charged to customers in the lower income zip code group was higher than the Eversource default generation service price for most of the sample months. By comparison, the median customer in the higher income group appears less than the contemporaneous default service option for many months of the sample.

The trends in customers' contract prices presented thus far used a customer's current contract price, which may differ from their initial contract price if the customer was automatically enrolled in a new contract after their initial contract expired. It is therefore useful to examine customers' initial contract prices to assess whether the patterns described above can be explained by differences in customers' inertia. The zip code-level data from June and September 2018, which identify which prices were charged to new customers, are summarized in Table 1.4.

Table 1.4 shows that customers living in below median income zip codes chose contracts that cost $10.49 more per month than the contracts chosen by customers living in above median income zip codes. Customers in below median income zip codes were three times more likely than customers in above median income zip codes to enroll in contracts that cost more than the current default service option (57 percent vs. 19 percent), and paid an average markup of $5.55 per month. The average customer in the above median income group chose a contract that offered $5.20 in monthly savings compared to the default option.

1.3.3 Contract Length Data

One way that competitive electric generation service contracts are differentiated is that they offer fixed prices for different lengths of time. Because the zip code-level data do not report the length of customers' contracts, I compile a second data set that I use to estimate the length of customers' contracts.

I obtain data on new and existing competitive retailer customers between October 2016 and March 2019. As before, the data set is a repeated cross section containing the contract choices of residential customers who were enrolled in competitive electric generation service contracts during the sample period. Because the data include all months between October 2016 and March 2019, I use them to estimate the length of a customer's contract by counting the number of consecutive months that I observe the same contract being sold to customers of the same type.

To measure a customer's contract length, I begin by assigning a unique ID to each contract in the data, where a contract is identified by its price, start date, and seller. To avoid assigning contracts that have the same price and seller but different (unobserved) contract lengths to the same ID, I exclude contracts that accepted new customers for three or more months after the first date that the contract appears in the data.[10] I also exclude contracts sold in October 2016, because I am unable to observe whether they were sold in previous months, as well as contracts sold after December 2018, because I cannot observe whether contracts were sold for more than four months. These sample restrictions may disproportionately exclude contracts sold by retailers who do not update their contract menus frequently, as well as retailers who sell multiple contracts with the same price. I do not expect this to introduce non-random sample selection in my analysis because I do not

[10] A customer's contract start date depends on their unique billing cycle, so it is possible that two customers may enroll in the same contract on the same day in June, however one appears in the data as a new customer in June while the other is a new customer in July.

expect customers' demographic characteristics to be correlated with whether they purchase from such retailers.

Next, I assign each new customer a type based on whether they are a hardship customer or use electric heating. This results in four customer types: (i) non-hardship customers without electric heating (78.8 percent of new customers), (ii) non-hardship customers with electric heating (13.5 percent), (iii) hardship customers without electric heating (6.3 percent), and (iv) hardship customers with electric heating (1.3 percent). Beginning with the customer's contract start date, I count the number of consecutive months that I observe a customer of the same type purchasing the same contract. This number becomes the customer's estimated contract length.

My procedure estimates the contract length for 26.6 percent of customers, whose contract terms are summarized in Table 1.5. The distribution of customer types in the sample approximately matches the distribution of customer types in the original data. The proportion of hardship customers in the sample is slightly higher, because I take a conservative approach to estimating a customer's contract length and the fact that there are fewer hardship customers makes it is easier to measure their contract length with confidence. The sample average contract price for each customer type is also similar to the corresponding group average in the original data set: \$63.72 for non-hardship non-heating, \$61.32 for non-hardship heating, \$73.13 for hardship non-heating, and \$71.65 for heating hardship customers. The sample and original data set contain overlapping observations and so it is not possible to use a standard t-test to compare whether these differences are statistically significant.

The average contract length is slightly longer for non-hardship customers. This is likely an artifact of the estimation procedure, which cannot distinguish between contracts that end because they expired or because customers terminated their contracts early. Because there are more non-hardship customers, one customer's early contract termination is less

likely to influence the estimated contract length for all of the customers who purchased that contract. There is no data on whether hardship customers are more or less likely to terminate their contracts early, so there is potential for non-random measurement error in the contract length variable.

1.3.4 Green Customer Data

Another way that competitive electric generation service contracts differ is according to their renewable energy content. Neither of the previous data sets include information on the renewable energy content of contracts, so I compile a third data set on contracts sold by two of Connecticut's largest competitive electric retailers, Town Square Energy and Discount Power, to construct a proxy variable to measure whether a contract includes 100% renewable energy.

Both Town Square Energy and Discount Power underwent investigation for their marketing practices in late 2018, and, as part of those investigations, were forced to publish data on all contracts sold between 2017 and 2018.[11] These data identify the number of green and non-green customers associated with each unique contract price sold during the investigation period. For both retailers, green customers' contracts included 100% renewable energy from wind and hydroelectric sources.

I use the green customer data to estimate whether a customer is enrolled in a 100% renewable energy contract given its price. For each contract price, I compute the share of retailer customers who pay that price and are also green energy customers. The result is a value between 0 and 1 for each retailer, unique contract price, and month between 2017 and 2018.

[11] See PURA Docket 08-0914RE02, *Application of Discount Power, Inc. for an Electric Supplier License – Investigation into Marketing*, and PURA Docket 10-03-11RE03, *Investigation Into Town Square Energy, LLC – Marketing.*

To evaluate whether low-income customers are more likely to purchase more expensive green energy contracts, I merge the green customer data with the zip code-level data on new customers from June and September 2018. I use the zip code-level data from these months because they identify new customers, whose observed contract prices are also their initial contract prices. Town Square Energy and Discount Power collectively served about 30 percent of new customers during June and September 2018.

The merged data set is summarized in Table 1.6. For convenience, I label contracts as "green" if over half of its customers are green customers. Given this definition, I estimate that 12 percent of Town Square Energy customers and 4 percent of Discount Power customers enrolled in green energy contracts. These estimates are consistent with the share of green energy customers calculated using the retailers' raw data.

For Town Square Energy, customers in below median income zip codes appear to have chosen green energy contracts at higher rates, while Discount Power customers in above median income zip codes appear to have chosen green energy contracts at higher rates. Additionally, the average difference between the prices paid by green and non-green customers is about $31 per month for Town Square Energy, compared to just $5.18 per month for Discount Power. Together, these patterns suggest that, at least for Town Square Energy, the gap in prices paid by customers living in above and below median income zip codes is driven in part by green energy customers.

1.3.5 Data Limitations

My methods assume that the unobserved contract terms are the same for customers who enrolled in a competitive electric generation service contract with the same retailer during the same month and paid the same price. In reality, retailers may sell multiple contracts that have the same price but differ in length, renewable energy content, or whether they include any cancellation fees. This introduces the potential for measurement error in

the contract length and green energy proxy variables, which identify a customer's contract length and renewable energy content based on their retailer and contract price. Furthermore, the probability that these variables are measured with error may differ by retailer, depending on the complexity and number of contracts that they offer.

1.4 Methods

The previous section showed that customers living in below median income zip codes paid more for competitive electric generation service than customers living in above median income zip codes in Connecticut. In this section, I describe methods for testing whether this relationship is statistically significant and for testing the validity of various explanations for why such a pattern exists.

1.4.1 Do low-income customers pay significantly higher prices?

The first part of the analysis uses the zip code-level data on customer accounts. The data include the generation service contract prices for all residential customers served by competitive electric retailers in the months of June and September between 2011 and 2018. Thus, the data set is a repeated cross section with a total of $T = 16$ months. The unit of observation is a customer i living in zip code z_i who purchases competitive electric service from a retailer r_i during month t_i.

To begin, I quantify the difference in prices paid by customers according to the median household income of the zip code in which they live:

$$Price_i = \gamma_{t_i} + \beta_z Zip\ Income_{z_i} + u_i \tag{1.1}$$

where $Price_i$ is the current generation service contract price for customer i during sample month t_i. The categorical variable $Zip\ Income_{z_i}$ is defined according to whether the me-

dian household income of customer's zip code is (i) less than $50,000, (ii) between $50,000-$80,000, (iii) between $80,000-$100,000, and (iv) above $100,000. The coefficients of interest are β_z, and measure the average difference in prices paid by customers in each zip code group compared to customers in an omitted zip code group, which I set as group (iii). The inclusion of sample month fixed effects, γ_{t_i}, improve the precision of the estimates for β_z by controlling for time-varying factors, such as fluctuations in wholesale electricity prices that affect retail electricity prices. Standard errors are clustered by zip code to account for the fact that a customer may appear in multiple sample months.

Because the median household income of a customer's zip code is an imperfect proxy for a customer's income, I estimate another model that uses a customer's hardship status to measure a customer's income:

$$Price_i = \gamma_{t_i} + \beta_z Zip\ Income_{z_i} + \beta_1 Hardship_i + u_i \tag{1.2}$$

which is the same as Equation (1.1), except that it includes the dummy variable *Hardship*, which equals one for utility hardship customers. Most customers qualify for hardship status because of their financial circumstance, and so β_1 measures the difference between the electricity prices paid by Connecticut's most economically vulnerable households and non-hardship customers living in zip codes with median household income between $80,000 and $100,000.

It is not obvious whether β_1 should be positive or negative. Hardship customers are protected from electric service termination during winter months, and, in many cases, receive lump-sum subsidies to offset their total energy costs, which could result in them being less sensitive to the price of electricity than non-hardship customers. However, previous studies have found that low-income customers who receive subsidies toward their natural gas bills tend to be more price elastic than high-income customers, even when they pay subsidized energy prices (Auffhammer and Rubin, 2018).

Another reason that low-income customers may be less price sensitive than high-income customers is that they consume less electricity. To explore whether the price gap between high- and low-income customers differs between high- and low-usage customers, I use data on whether a customer uses electricity as their primary heating source to proxy for electric usage. Electric heating customers use about 1.5 times more electricity than non-heating customers annually, and so I expect that electric heating customers choose cheaper contracts. Specifically, I estimate,

$$\begin{aligned} Price_i =& \gamma_{t_i} + \beta_z Zip\ Income_{z_i} + \beta_1 Hardship_i + \beta_2 Heating_i \\ &+ \beta_3 Hardship_i \times Heating_i + u_i \end{aligned} \tag{1.3}$$

where $Zip\ Income_{z_i}$ is a set of categorical variables defined as before, $Hardship_i$ and $Heating_i$ are dummy variables equal to one if a customer is a hardship customer or receives residential electric heating service, respectively, and γ_{t_i} are month of sample fixed effects.

Equation (1.3) has the form of a differences-in-differences model, where β_1 is the average price difference between hardship and non-hardship customers who do not use electricity as their primary heating source, and $\beta_1 + \beta_3$ is the average price difference between hardship and non-hardship customers who do. If the price gap between high- and low-income customers is related to differences in their electric usage, then I expect the price gap between high- and low-income customers to be smaller among high-usage customers. This prediction corresponds with $\beta_3 < 0$. Additionally, I expect that, conditional on a customer's income, high-usage customers should choose lower prices than low-usage customers, which corresponds with $\beta_2 < 0$ and $\beta_2 + \beta_3 < 0$.

1.4.2 Does inertia explain why low-income customers pay more?

The previous subsection discussed methods for describing the relationship between electricity prices and household income using customers' current electric generation service contract prices. However, a customer's current contract price may not be the same as their initial contract price, because a customer's contract may automatically renew to higher prices after their initial contract expires. If low-income customers are slower to cancel their contracts after their contract terms change compared to high-income customers, then differences in inertia that are correlated with customers' income levels may explain why low-income customers pay more for electricity.

I test whether inertia explains why low-income customers pay more for electricity using the data from June and September 2018, which include an indicator variable for customers whose contracts began during the current month. To compare the gap between high- and low-income customers' initial contract prices—which reflect customers' active contract choices—to the gap between high- and low- income customers' current contract prices, I estimate

$$\begin{aligned} Price_i = \beta_0 + \beta_1 New_{t_i} + \beta_2 Zip\ Income_{z_i} + \beta_3 New_{t_i} \times Zip\ Income_{z_i} \\ + \beta_4 Sept_{t_i} + \beta_5 New_{t_i} \times Sept_{t_i} + u_i \end{aligned} \tag{1.4}$$

where New_{t_i} and $Sept_{t_i}$ are dummy variables equal to one if a customer began their contract during month t_i and whether the customer is observed during September 2018. The coefficients of interest are β_3. If the elements of β_3 are positive for customers living in lower income zip codes, then the gap between high- and low- income customers' new contract prices is larger than the gap between their current contract prices. Such a finding would suggest that low-income customers pay more for electricity than high-income cus-

tomers because they actively choose more expensive contracts, and not because they exhibit greater inertia.

I also estimate a version of (1.4) that include a customer's hardship status to proxy for a customer's income:

$$Price_i = \beta_0 + \beta_1 New_{t_i} + \beta_2 Zip\ Income_{z_i} + \beta_3 New_{t_i} \times Zip\ Income_{z_i} + \beta_4 Sept_{t_i} + \beta_5 New_{t_i} \times Sept_{t_i} + \beta_6 Hardship_i + \beta_7 New_{t_i} \times Hardship_i + u_i$$

where the coefficient β_7 is interpreted in the same way as β_3 in Equation 1.4.

1.4.3 Do low-income customers choose to pay more?

The previous section proposed a test for determining whether low-income customers pay more for electricity than high-income customers because they exhibit greater inertia or because they actively choose more expensive contracts. Two reasons that low-income customers may choose more expensive contracts are that (a) they face higher search costs that limit their ability to choose the cheapest contracts available, and (b) they prefer more expensive competitive electric supply contracts with non-price attributes, such as contracts that offer renewable energy or price stability.

Because residential customers in Connecticut, by default, pay regulated prices for their utility's standard service option, customers should only be willing to pay more for a competitive generation service contract if it increases their utility relative to the standard service option. If low-income customers are more likely than high-income customers to choose contracts that are more expensive than the standard service option, then search costs alone cannot explain why low-income customers choose more expensive generation service contracts. Instead, such a finding would support the idea that low-income customers have a higher willingness to pay for the non-price attributes of contracts.

Using the zip code-level data on new customers from June and September 2018, I estimate versions of Equations 1.1-1.3, where the dependent variable is a binary variable equal to one if a customer enrolls in a new contract that costs more than the current standard service price. The coefficients corresponding with customers living in below median income zip codes and hardship customers measure how much more likely they are to choose a contract that costs more than their utility's standard service option, compared to non-hardship customers and customers in above median income zip codes.

1.4.4 Do low-income customers pay more for price stability?

Because the standard service rate changes every six months and updates are announced only 45 days prior, risk averse customers may be willing to pay more than the current standard service rate to avoid facing uncertainty over the future price of electricity. Low-income customers, in particular, may have a higher willingness to pay more for low-usage fixed-price contracts than high-income customers, because they spend a larger share of their income on monthly energy expenditures.[12]

To evaluate whether low-income customers' preference for low-usage fixed-price contracts explains why they pay more for electricity, I estimate the following equation using the contract length data,

$$\begin{aligned} Price_i = \gamma_{t_i} + \alpha_{r_i} + \beta_1 Hardship_i + \beta_2 Heating_i + \beta_3 Hardship_i \times Heating_i \\ + \beta_4 Long\ Term_i + \beta_5 Hardship_i \times Long\ Term_i + u_i \end{aligned} \tag{1.5}$$

[12] Also, Abito and Salant (2019) find that lower income individuals are more likely to buy extended warranties for TVs and other products, and that strong demand and high profits for extended warranties are driven by consumers distorting the failure probability of the insured product, rather than standard risk aversion or sellers' market power. Thus, it is also possible that low-income customers are willing to pay more for contracts that offer long term price stability because they mispredict the probability that the default service rate will increase.

where $Long\ Term_i$ is a dummy variable equal to one if a customer's contract term is longer than 18 months.[13] The explanatory variables are defined as in Equation 1.3, except that (1.5) includes retailer fixed effects α_{r_i} to control for potential measurement error that may be correlated with the identity of a customer's chosen retailer for the reasons discussed in Section 1.3.3. The coefficient of interest is β_5, which measures how much more hardship customers pay for long term contracts. If $\beta_5 > 0$, then this is evidence that hardship customers pay more for long term contracts, which may be the case if they face higher search costs, or choose long term contracts that also include renewable energy. I use hardship customers to proxy for low-income customers because the contract length data do not identify the zip codes for where customer lives.

There are other reasons to include retailer fixed effects in (1.5), that also apply to the regression models discussed above. For example, retailer fixed effects control for unobserved differences in a retailer's quality or branding that may influence the prices they charge for competitive electric generation service. Should these unobserved factors be uncorrelated with whether customers are hardship customers or use electric heating, then including the retailer fixed effects improves the precision of the other coefficient estimates. However, the use of retailer fixed effects changes the interpretation of the estimates, such that they measure the gap between high- and low-income customers of the same retailer. Such comparisons are not useful for understanding why low-income customers choose more expensive contracts than high-income customers, because customers of the same retailer do not necessarily choose from the same set of contracts. This is because retailers may update their prices on a daily basis, set prices via negotiation, or advertise different prices online than they sell through in-person marketing activities. As a result, the primary methods in this paper do not include retailer fixed effects.

[13] I choose 18 months as the cut off because it maximizes statistical power after estimating versions of Equation 1.5 using multiple categorical variables to describe a customer's contract length.

1.4.5 Do low-income customers pay more for renewable energy?

To evaluate whether low-income customers pay more for electricity because they enroll in more expensive green energy contracts, I use the data set on green energy contract prices for Town Square Energy and Discount Power that are linked to the zip code-level data on new and existing customer accounts from June and September 2018.

Because I use zip code-level data, I proxy for a customer's income using the median household income of their zip code. Specifically, I estimate

$$\begin{aligned} Price_i = {} & \beta_{0,r_i} + \beta_{1,r_i} LowSES_{z_i} + \beta_{2,r_i} Hardship_i + \beta_{3,r_i} Heating_i \\ & + \beta_{4,r_i} Hardship_i \times Heating_i + \beta_{5,r_i} Green_i + \beta_{6,r_i} LowSES_{z_i} \times Green_i \\ & + \beta_{7,r_i} Sept_{t_i} + u_i \end{aligned} \tag{1.6}$$

where $Green_i$ equals one if a customer is enrolled in a green energy contract, and $Low\ SES_{z_i}$ equals one if a customer lives in a below median income zip code. The coefficients of interest are β_{6,r_i}, which measure the additional cost that low-income customers pay more for green energy contracts compared to high-income customers. The coefficients are allowed to vary by retailer to account for the fact that differences in how a retailer brands its products may attract customers with different preferences for green energy.

Additionally, I estimate a linear probability model where the dependent variable is $Green_i$, which equals one if a customer enrolls in a green energy contract,

$$\begin{aligned} Green_i = {} & \beta_{0,r_i} + \beta_{1,r_i} LowSES_{z_i} + \beta_{2,r_i} Hardship_i + \beta_{3,r_i} Heating_i \\ & + \beta_{4,r_i} Hardship_i \times Heating_i + \beta_{5,r_i} Sept_{t_i} + \beta_{6,r_i} New_i + \beta_{7,r_i} New_i \times Sept_{t_i} + u_i \end{aligned} \tag{1.7}$$

I include the variables New and $New \times September$, which equal one for new customers in June and September 2018, respectively, to control for the presence of marketing campaigns

that may have encouraged customers to enroll in green energy contracts at higher rates during those months. The coefficients of interest are β_{1,r_i}, which are allowed to vary for each retailer r_i. If $\beta_{1,r_i} > 0$ for both retailers, this would be evidence that low-income customers are more likely to choose green energy contracts, regardless of their choice of retailer.

1.5 Results

1.5.1 Do low-income customers pay significantly higher prices?

To assess whether trends differ before and after Connecticut regulators enacted new consumer protection laws, I split the zip code-level data into two periods before and after January 2016. The results from estimating Equations 1-3 using data from each period are presented in Table 1.7. Additionally, I test whether the coefficient estimates differ between the two periods by estimating the same equations using data from both periods, except that I include a set of interaction terms between each of the explanatory variables and a dummy variable equal to one for observations after January 2016. The estimates for the pooled regressions are reported in Table 1.8.

The estimates in Table 1.7 demonstrate that there is a statistically significant relationship between the median household income of a customer's zip code and their competitive electric generation service contract price. Customers living in zip codes with median household income below $50,000 paid the highest prices between 2011-2018—about $3 more per month more from 2011-2015 and $5 more per month more from 2016-2018 than customers living in zip codes with median household income between $80,000-$100,000. The difference in prices paid between customers living in wealthy zip codes (with median household income above $100,000) and customers living in zip codes with median household income between $80,000-$100,000 is not statistically significant. Thus, the relation-

ship between zip code income and electricity prices appears to disappear as zip code income increases.

Columns (2) and (5) show that hardship customers paid significantly higher prices than non-hardship customers, which further supports the finding that low-income customers in pay the highest electricity prices in Connecticut. Columns (3) and (6) confirm the hypothesis that electric heating customers paid lower prices ($\beta_2 < 0$). Surprisingly, the coefficient for *Hardship* $\times$ *Heating* is positive and statistically significant, which suggests that the gap in prices paid by hardship and non-hardship customers does not shrink with a customer's expected usage ($\beta_3 > 0$). Furthermore, the difference between the prices paid by high-usage and low-usage hardship customers, $\beta_2 + \beta_3$, is not statistically significant ($F = 2.06,\ p = 0.1527$ for 2011-2015, and $F = 0.84,\ p = 0.3606$ for 2016-2018).

The estimates from Table 1.8 show that the gap between prices paid by high- and low-income customers increased following the adoption of the new consumer protection laws, and that the gap approximately doubled for customers living in the poorest zip codes and utility hardship customers. The coefficient estimates for *Post* $\times$ *Zip Income Above 100k* and *Post* $\times$ *Electric Heating* are both negative and statistically significant, which suggests that non-hardship, high-usage customers, as well as customers living in the richest zip codes, may have benefited from the consumer protection laws. By contrast, the prices paid by high-usage hardship customers increased over time.

1.5.2 Does inertia explain why low-income customers pay more?

The results from estimating Equation 1.4 using the zip code-level data on new and existing customer accounts from June and September 2018 are reported in Table 1.9. On average, customers in the poorest zip codes enroll in contracts that cost over $10 more per month than the contracts chosen by customers living in the richest zip codes. Additionally, the coefficient on *New* $\times$ *Hardship* is statistically significant in Columns (2) and

(3), which further supports the finding that low-income customers enroll in contracts that are significantly more expensive than those chosen by high-income customers. The magnitude of these price differences is both statistically and economically significant. In 2018, the average monthly cost of the Eversource standard service option was $70, so a $10 price difference represents about 14 percent of a customer's monthly electric supply bill.

The estimates in Table 1.9 confirm that new customers paid significantly lower prices than existing customers, across all zip code groups. This pattern suggests that customers of all income levels demonstrate substantial inertia, because some existing customers could have paid less by switching to new, cheaper contracts. Interestingly, the gap between new and existing customers' contract prices is larger among customers living in above median income zip codes than among customers living in the poorest zip code group (about $12-$14 per month vs. $8 per month in the above median income and poorest zip code groups, respectively). Thus, the data do not provide any indication that customers in poor zip codes exhibit higher inertia than customers in above median income zip codes.

Finally, the coefficient estimates that correspond with β_3 in Equation 1.4 for customers living in below median income zip codes, *New* $\times$ *Zip Income Below 50k* and *New* $\times$ *Zip Income Between 50-80k*, are positive and statistically significant across all specifications. This result implies the gap between high- and low-income customers' new contract prices—which reflect active enrollment decisions—is larger than the gap in existing customers' contract prices. Thus, the reason that low-income customers pay more for electricity than high-income customers seems to be that low-income customers choose more expensive contracts, and not because they display a higher degree of retailer inertia.

1.5.3 Do low-income customers choose to pay more?

Table 1.10 reports the estimates for Equations 1.1-1.3, where the dependent variable is a binary variable equal to one if a customer chooses a contract that is more expensive

than the Eversource standard service option at the time of enrollment. Customers living in Connecticut's poorest zip codes, as well as hardship customers, are significantly more likely to choose contracts that cost more than the current standard service option than non-hardship customers in above median income zip codes. These patterns may be rationalized if low-income customers choose to purchase competitive electric supply contracts for their non-price attributes, rather than the short-term cost savings that they offer.

Another interesting result is that the coefficient estimate for *Hardship* × *Heating* is positive and statistically significant, which means that hardship customers with high electric usage are more likely than low-usage hardship customers to choose contracts that offer negative savings at the time of enrollment. One explanation for this finding could be that low-income electric heating customers are more likely to enroll in long-term fixed-price contracts that cost more than the standard service option at the time of enrollment, but eliminate uncertainty over the price of electricity during future high-usage months.

1.5.4 Do low-income customers pay more for price stability?

Table 1.11 presents the coefficient estimates for Equation 1.5 using data on the estimated contract length for customers who enrolled in new competitive generation service contracts between November 2016 and October 2017. The estimation sample is limited to these customers because the data do not permit measurement of whether a contract sold after October 2017 had a term that was longer than 18 months. The summary statistics for the estimation sample are similar to those for the full data set (see Table 1.12).

The estimated coefficients for *Long Term* in Table 1.11 indicate that customers who chose long term, fixed-price contracts pay about $3-4 more per month than customers who chose contracts that were shorter than 18 months in length. However, the estimates for *Hardship* × *Long Term* are not statistically significant, and even change sign between Columns (3) and (4), so it is not possible to conclude whether hardship customers paid

higher prices for long term, fixed-price contracts than non-hardship customers. Furthermore, *Hardship* is positive and statistically significant across all specifications, which indicates that long term price stability cannot fully explain why hardship customers chose contracts that cost more than non-hardship customers.

1.5.5 Do low-income customers pay more for renewable energy?

I estimated Equation 1.6 using the Town Square Energy and Discount Power green energy contract data merged with the zip code-level data on customer accounts from June and September 2018. I prefer to use data from these months because they identify new customers, which allows me to focus on customers' active contract choices. These data also allow me to proxy for a customer's income using their zip code's median household income. To maximize statistical power, I use two zip code groups, depending on whether the zip code median household income was reported as being above or below $80,000 in the American Community Survey data for 2015-2019.

The results are reported in Table 1.13. For both retailers, the coefficient estimates for *Zip Income* $<$ *\$80k* are positive and statistically significant across all specifications, regardless of whether green customer controls are included. This pattern suggests that green energy explains some but not all of the gap in prices paid by Town Square Energy and Discount Power customers living in above and below median income zip codes.

For Town Square Energy, the coefficient for *Zip Income* $<$ *\$80k* shrinks from \$7.742 to \$1.998 when green customer controls are included. This suggests that Town Square Energy customers in below median income zip codes paid more than customers in above median income zip codes because they were more likely to enroll in green energy contracts. However, the coefficient on *Below \$80k* $\times$ *Green* is negative and statistically significant, which means that, conditional on choosing a green energy contract, customers living in below median income zip codes chose slightly cheaper contracts.

For Discount Power, the coefficient estimates for *Zip Income* $<$ *\$80k* change only slightly with the addition of green customer controls, and the coefficient on *Below \$80k* $\times$ *Green* is not statistically significant. Thus, it seems unlikely that low-income customers' preferences for green energy explains why Discount Power customers living in below median income zip codes chose more expensive contracts than customers living in above median income zip codes.

The results from estimating Equation 1.7 using data on both new and existing customers of Town Square Energy and Discount Power from June and September 2018 are reported in Table 1.14. The estimates confirm that customers living in below median income zip codes were more likely to enroll in green energy contracts if they were served by Town Square Energy, but less likely to enroll in green energy contracts if they were served by Discount Power. As a result, it is not possible to speculate whether the relationship between customers' income and their willingness to pay for green energy is positive or negative for other retailers.

1.6 Discussion

The results in this paper demonstrate that Connecticut households who live in the state's poorest zip codes pay significantly more for competitive electric generation service than customers living in above median income zip codes. While other researchers have observed similar patterns in other deregulated retail electricity markets, their explanations for why low-income customers pay higher prices do not explain the patterns I observe in Connecticut's retail electricity market. For example, Hortaçsu et al. (2017) and Byrne and Martin (2021) attribute regressive electricity prices in Texas and Victoria, Australia, to the fact that low-income customers exhibit greater inertia and face higher search costs. Yet, neither inertia nor search costs can fully explain why over half of customers living in

Connecticut's lowest income zip codes actively chose competitive electric generation service contracts that cost more than the cost of their utility's default service option.

This paper considered whether customers in low-income zip codes chose to pay more for electricity than customers in high-income zip codes because they valued the non-price attributes of competitive generation service contracts, such as renewable energy and price stability. My results suggest that customers living in below median income zip codes may have paid more for green energy because they chose more expensive green retailers, and not because they have a higher willingness to pay for green energy. Customers in below median income zip codes were four times more likely to purchase green energy contracts from a retailer that charged an average markup of $30 per month for its renewable energy contracts, but less likely to purchase a comparable renewable energy contract from a retailer that charged an average markup of $6 per month. This finding suggests that customers living in lower income zip codes face substantial search costs, otherwise they would have chosen to purchase green energy contracts from the cheaper retailer. Alternatively, customers may have underestimated the additional cost associated with their contract choice because retailers report prices in cents per kilowatt-hour.

The results are inconclusive about whether low-income customers paid significantly more for long-term contracts, perhaps because of measurement error in how I estimate a customer's contract length. The results are sensitive to the inclusion of retailer fixed effects, which supports the hypothesis that the price gap between high- and low-income customers is related to how customers sort across retailers.

According to the Connecticut Office of Consumer Counsel, some retailers offered $50 gift cards and other gifts, such as movie tickets or smart thermostats, to new customers. While the data analyzed in this paper do not indicate whether a customer received a gift for enrolling in their contract, it is clear that, for many customers, the markup they pay for competitive electric generation service exceeds the value of these gifts. For example,

more than one quarter of customers paid $17 or more per month more than they would have paid under their utility's default service option between 2011-2018. These customers would have incurred an additional cost greater than the value of a $50 gift card within four months of service. Similar calculations led state regulators to conclude that "value added" promotions such as gift cards do not add any actual "value" to customers.[14] Even if the lure of gift cards can explain why some customers chose more expensive contracts, it therefore seems unlikely that these contracts improved consumer welfare.

Another interesting finding is that the gap between prices paid by customers living in above and below median income zip codes increased over time, despite the implementation of new consumer protection laws in 2014 and 2015. It is possible that higher income customers benefited more from the new regulations, which included an information intervention to increase consumer awareness of their contract terms. Another reason could be that retailers increased their door-to-door sales and telemarketing activities after the new regulations, and were more likely to send their sales agents to enroll customers in lower income neighborhoods.

1.7 Conclusion

The findings in this paper demonstrate that the distribution of prices in Connecticut's retail electricity market are regressive: customers in the state's poorest zip codes and utility hardship customers paid significantly higher prices than customers living in above median income zip codes between 2011 and 2018. Although similar patterns have been documented in other jurisdictions with retail electric competition, the Connecticut market differs from those markets in important ways. First, customers are permitted but not required to choose a competitive electric retailer; yet 57 percent of customers in below median income zip codes actively *chose* to pay more for electricity than they would have paid

[14] See the Final Decision in PURA Docket 18-06-02

had they not chosen any competitive retailer, compared to just 19 percent of customers living in above median income zip codes. Second, Connecticut enacted consumer protection laws that aimed to improve customer's awareness of their contract terms and prohibit retailers from selling contracts that profit from consumer inattention and inertia. Yet, my analysis shows that the average markup paid by competitive retailer customers increased after these laws were implemented. One explanation for this is that the total number of customers purchasing from competitive retailers declined over time, and customers who continued to purchase competitive electric generation service either had extremely high inertia, a high willingness to pay for competitive retailers' services, or were persuaded to enroll in contracts via retailers' marketing activities.

The Connecticut Public Utilities Regulatory Authority recently completed its investigations into the marketing practices of several major retailers and established a new Office of Education and Enforcement to aid with ongoing monitoring of the market. This provides an opportunity for future researchers to evaluate the benefit of the state's increased enforcement activities for consumers.

1.8 Figures and Tables

Figure 1.1: Competitive Electric Generation Service Contract Prices Over Time

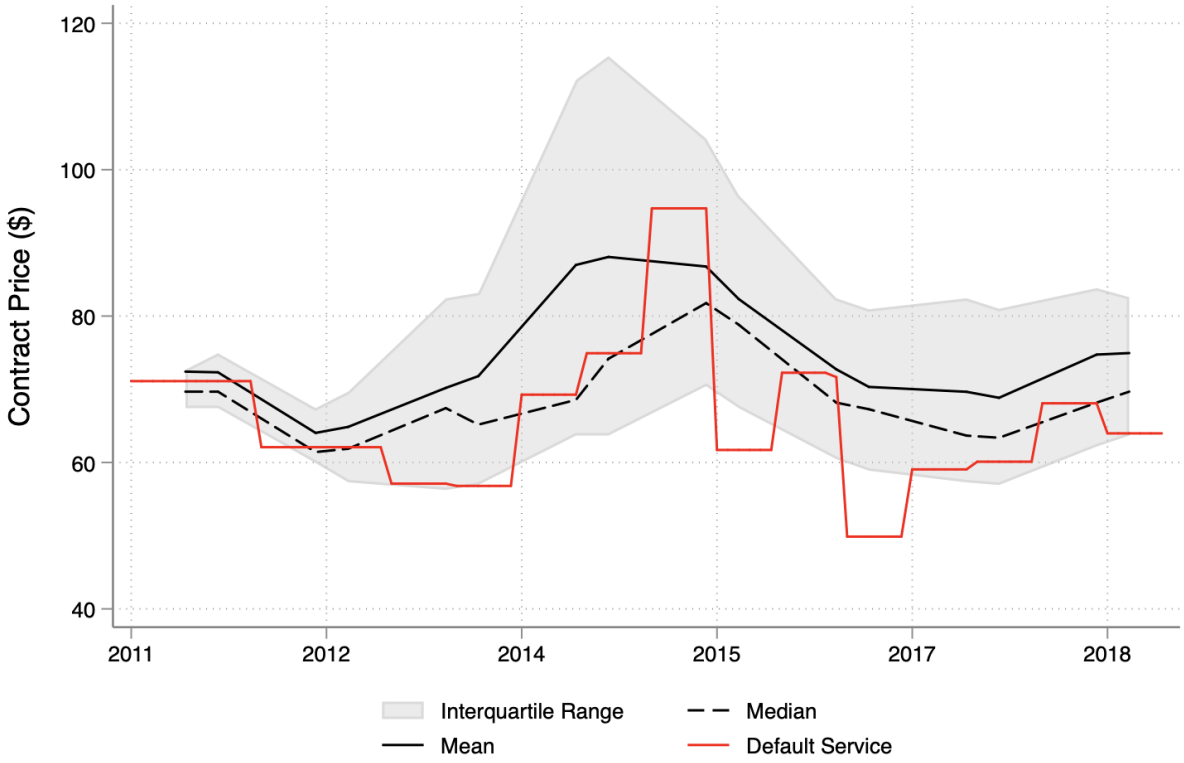


Figure 1.2: Competitive Electric Generation Service Contract Prices by Zip Code Income

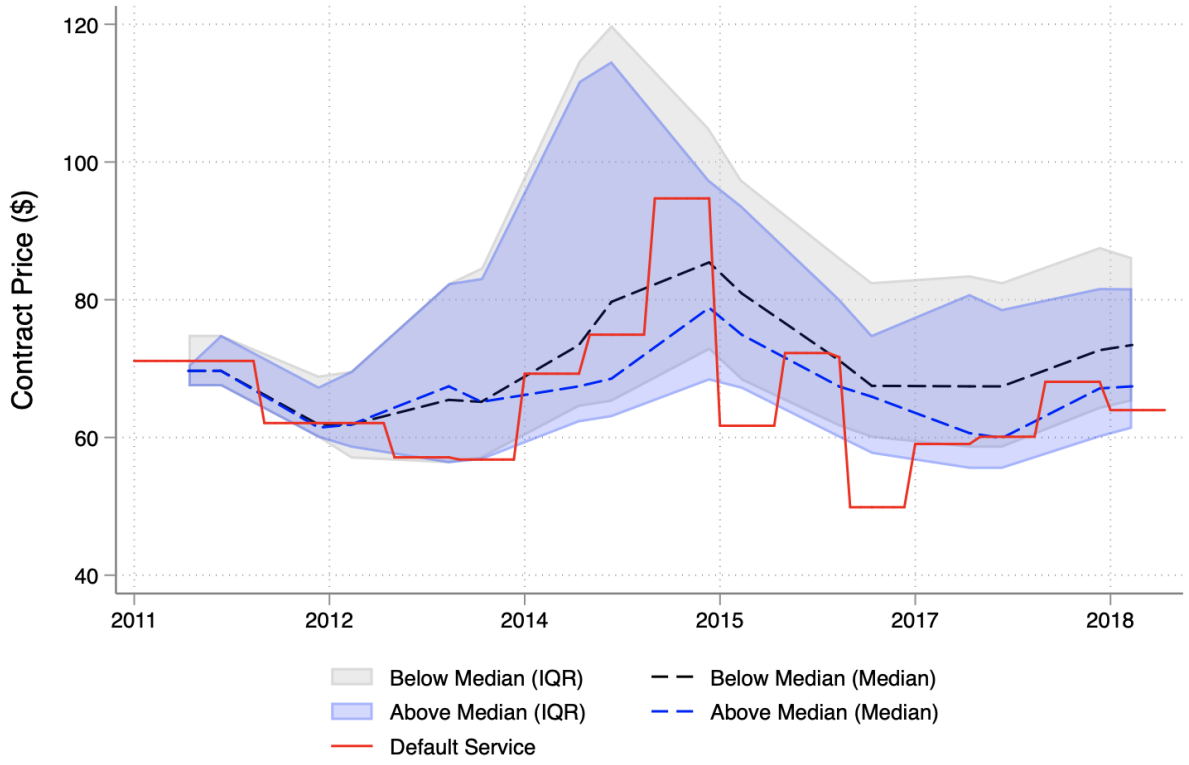


Table 1.1: Overview of Data Set Periods Covered

Year	Zip Code-Level	Contract Length	Green Customer
2011	June, Sept		
2012	June, Sept		
2013	June, Sept		
2014	June, Sept		
2015	June, Sept		
2016	June, Sept	Nov-Dec	
2017	June, Sept	All Months	All Months
2018	June*, Sept*	All Months	All Months

*The Zip-level Accounts data for June and September 2018 include an extra variable for new customer accounts.

Table 1.2: Overview of Data Set Variables

Variable	**Description**	**Zip**	**Length**	**Green**
Contract Characteristics				
Competitive Supplier	Name of competitive electric supplier	Yes	Yes	Yes
New Customer	Equals one for customers who begin their contract in the current month	Yes*	Yes	No
Contract Price	Price per kilowatthour of electricity consumed in current month	Yes	Yes	Yes
Contract Length	Number of months that the initial contract price is fixed	No	Yes	No
Green Energy Content	Equals one for customers purchasing 100% renewable energy	No	No	Yes
Customer Characteristics				
Electric Heating	Equals one for customers who use electric heating	Yes	Yes	No
Hardship	Equals one for utility hardship customers	Yes	Yes	No
Zip Code	Zip code where customer lives	Yes	No	No

Table 1.3: Summary of Zip Code-Level Data for Above and Below Median Income Zip Codes

	2011-2015			2016-2018		
	Below Median	Above Median	Difference	Below Median	Above Median	Difference
Contract Price						
Mean	76.54	74.75	1.79	73.35	69.99	3.36
First quartile	61.88	61.42	0.45	59.92	58.50	1.42
Second quartile	69.68	68.77	0.91	69.75	65.92	3.83
Third quartile	84.54	80.93	3.61	84.67	79.50	5.17
Standard deviation	20.58	19.58	1.00	16.25	16.87	-0.62
Additional Cost/Savings						
Mean	8.71	6.83	1.88	11.26	7.83	3.43
First quartile	-3.69	-4.09	0.40	-1.61	-4.84	3.23
Second quartile	0.95	-0.22	1.17	8.37	3.12	5.25
Third quartile	17.51	13.21	4.30	22.47	18.09	4.38
Standard deviation	20.25	19.85	0.39	17.26	17.83	-0.57
Above Default Price	0.52	0.47	0.05	0.68	0.57	0.11
Electric Heating	0.13	0.14	-0.01	0.13	0.14	-0.01
Hardship Customer	0.11	0.03	0.08	0.10	0.03	0.08
Observations	2,155,105	1,957,470		1,010,686	888,064	

Table 1.4: Summary of Zip Code-Level Data for June and September 2018

	New Customers			Existing Customers		
	Below Median	Above Median	Difference	Below Median	Above Median	Difference
Contract Price (1 Month)						
Mean	71.92	61.43	10.49	76.76	73.74	3.02
First quartile	63.67	53.10	10.58	64.13	60.00	4.13
Second quartile	71.17	59.85	11.32	73.43	67.20	6.23
Third quartile	82.42	67.13	15.30	88.43	82.58	5.85
Standard deviation	12.22	10.71	1.51	16.75	17.89	-1.14
Total Contract Cost (4 Months)						
Mean	287.68	245.73	41.95	307.02	294.96	12.06
First quartile	254.70	212.40	42.30	256.50	240.00	16.50
Second quartile	284.70	239.40	45.30	293.70	268.80	24.90
Third quartile	329.70	268.50	61.20	353.70	330.30	23.40
Standard deviation	48.87	42.84	6.03	67.00	71.55	-4.55
Expected Markup (1 Month)						
Mean	5.55	-5.20	10.75	8.67	5.66	3.01
First quartile	-4.12	-12.66	8.54	-3.96	-8.08	4.12
Second quartile	3.84	-4.41	8.25	5.34	-0.89	6.23
Third quartile	14.70	-0.58	15.28	20.34	14.49	5.85
Standard deviation	12.73	10.98	1.75	16.75	17.89	-1.14
Above Default Price	0.57	0.19	0.38	0.58	0.44	0.14
Electric Heating	0.13	0.16	-0.03	0.12	0.14	-0.02
Hardship	0.16	0.03	0.13	0.10	0.02	0.08
Observations	15,350	9,414	24,764	150,832	134,492	285,324

Notes: The table summarizes the contract choices of Eversource Energy residential customers who purchased electric generation service from a competitive supplier in June and September 2018. The Below/Above Median groups includes customers living in zip codes with median household income less/more than $80,000. Contract Price is the customer's contract price per 750 kWh usage. Markup is the additional cost or savings that the customer pays given their contract choice, compared to how much they would have paid under their utility's default service option (assuming 750 kWh monthly usage). Above Default Price is a dummy variable equal to one if a customer's contract price is above the utility's default service price. Electric Heating and Hardship equal one for electric heating and utility hardship customers, respectively.

Table 1.5: Summary of Contract Length Data by Customer Type

	Non-Hardship		Hardship	
	Non-Heating	Heating	Non-Heating	Heating
Contract Price	64.51	61.62	74.53	71.16
	(14.43)	(13.37)	(13.29)	(14.45)
Term (Years)	1.492	1.328	1.272	1.172
	(0.528)	(0.560)	(0.524)	(0.494)
Under 6 Months	0.0121	0.0168	0.0400	0.0710
	(0.109)	(0.129)	(0.196)	(0.257)
6 Months	0.0107	0.0235	0.0277	0.0386
	(0.103)	(0.152)	(0.164)	(0.193)
12-17 Months	0.360	0.360	0.303	0.293
	(0.480)	(0.480)	(0.460)	(0.455)
18-23 Months	0.241	0.165	0.244	0.323
	(0.428)	(0.371)	(0.429)	(0.468)
24-35 Months	0.244	0.183	0.133	0.0357
	(0.429)	(0.386)	(0.339)	(0.186)
New Customers	97375	19008	9348	2126
% New Customers	76.1	14.9	7.3	1.7

Notes: Main table values are the average of each variable. Standard deviation in parentheses. Contract Price is price per 750kWh. Sample includes new competitive electric retailer customers between November 2016-December 2018.

Table 1.6: Summary of Green Customer and Zip Code-Level New Customer Data

Panel A. Town Square Energy

	Green			Non-Green		
	Below Median	Above Median	Difference	Below Median	Above Median	Difference
Contract Price						
Mean	87.97	89.34	-1.37	58.13	56.02	2.11
First quartile	82.65	88.88	-6.22	52.28	51.30	0.97
Second quartile	88.88	90.75	-1.88	55.42	52.88	2.55
Third quartile	90.75	90.75	0.00	59.78	59.78	0.00
Standard deviation	4.73	3.86	0.88	7.71	5.79	1.92
Expected Markup						
Mean	22.15	24.61	-2.46	-8.05	-10.14	2.08
First quartile	14.57	24.90	-10.33	-15.21	-15.21	0.00
Second quartile	24.90	26.78	-1.88	-11.70	-11.70	0.00
Third quartile	25.59	26.78	-1.19	-4.20	-4.20	0.00
Standard deviation	5.89	4.85	1.04	8.53	6.78	1.74
Above Default Price	1.00	1.00	0.00	0.11	0.08	0.03
Hardship	0.20	0.15	0.05	0.06	0.01	0.05
Electric Heating	0.09	0.07	0.02	0.15	0.14	0.01
Total New Customers	408	60		1,441	2,090	
% Total New Customers	0.10	0.02		0.36	0.52	

Panel B. Discount Power

	Green			Non-Green		
	Below Median	Above Median	Difference	Below Median	Above Median	Difference
Contract Price						
Mean	71.39	66.98	4.41	65.17	60.00	5.17
First quartile	63.75	63.75	0.00	56.17	53.10	3.08
Second quartile	70.43	65.18	5.25	67.88	56.17	11.70
Third quartile	70.43	70.43	0.00	67.88	67.88	0.00
Standard deviation	8.90	4.81	4.09	7.49	7.51	-0.02
Expected Markup						
Mean	5.18	1.17	4.01	-2.01	-7.23	5.22
First quartile	-0.22	-0.22	0.00	-11.55	-11.91	0.36
Second quartile	2.34	-0.22	2.57	-0.21	-11.55	11.34
Third quartile	2.34	2.34	0.00	-0.21	-0.21	0.00
Standard deviation	9.56	4.59	4.98	8.14	7.68	0.45
Above Default Price	0.69	0.36	0.33	0.16	0.06	0.10
Hardship	0.11	0.01	0.10	0.12	0.03	0.09
Electric Heating	0.09	0.17	-0.08	0.13	0.16	-0.03
Total New Customers	81	83		1,912	1,572	
% Total New Customers	0.02	0.02		0.52	0.43	

Table 1.7: Relationship between Zip Code Income and Electricity Contract Prices

	2011-2015			2016-2018		
	(1)	(2)	(3)	(4)	(5)	(6)
Zip Income Below 50k	3.080***	2.888***	2.719***	5.407***	5.060***	4.754***
	(0.343)	(0.351)	(0.325)	(0.490)	(0.514)	(0.461)
Zip Income Between 50-80k	1.299***	1.246***	1.181***	2.186***	2.087***	1.973***
	(0.310)	(0.308)	(0.297)	(0.417)	(0.414)	(0.375)
Zip Income Above 100k	-0.247	-0.223	-0.353	-0.564	-0.519	-0.736**
	(0.254)	(0.253)	(0.239)	(0.361)	(0.360)	(0.317)
Hardship Customer		1.231***	0.858***		2.483***	2.077***
		(0.142)	(0.149)		(0.221)	(0.226)
Electric Heating			-2.620***			-3.969***
			(0.125)			(0.152)
Hardship × Heating			2.916***			3.738***
			(0.231)			(0.298)
Observations	4112575	4112575	4112575	1898750	1898750	1898750
R-Sq	0.191	0.191	0.193	0.0343	0.0357	0.0415
F-Stat	43.97	92.88	175.8	70.69	134.5	262.3
Sample Month FE	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses

Notes: The dependent variable is the customer's monthly supply cost given their contract choice, assuming 750kWh usage. The sample includes Eversource Energy residential customers who purchased electric generation service from a competitive supplier in select months between 2011-2018. Zip Income refers to the median household income of the zip code in which a customer lives. Customers living in zip codes with median household income between $80,000-100,000 are the base category. Electric Heating is a dummy variable equal to one if a customer uses electricity as their primary heating source. Hardship is a dummy variable equal to one if a customer is protected from having their power shut off due to nonpayment during winter months. Standard errors are clustered at the zip code level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 1.8: Relationship between Zip Code Income and Electricity Contract Prices Over Time

	(1)	(2)	(3)
Zip Income Below 50k	3.080***	2.888***	2.719***
	(0.343)	(0.351)	(0.325)
Zip Income Between 50-80k	1.299***	1.246***	1.181***
	(0.310)	(0.308)	(0.297)
Zip Income Above 100k	-0.247	-0.223	-0.353
	(0.254)	(0.253)	(0.239)
Post × Zip Income Below 50k	2.327***	2.172***	2.035***
	(0.264)	(0.267)	(0.251)
Post × Zip Income Between 50-80k	0.887***	0.841***	0.792***
	(0.208)	(0.208)	(0.190)
Post × Zip Income Above 100k	-0.316	-0.296	-0.384**
	(0.206)	(0.206)	(0.190)
Hardship Customer		1.231***	0.858***
		(0.142)	(0.149)
Post × Hardship Customer		1.251***	1.219***
		(0.133)	(0.147)
Electric Heating			-2.620***
			(0.125)
Hardship × Heating			2.916***
			(0.231)
Post × Electric Heating			-1.349***
			(0.113)
Post × Hardship × Heating			0.822***
			(0.266)
Observations	6011325	6011325	6011325
R-Sq	0.161	0.161	0.164
F-Stat	36.83	75.32	155.5
Sample Month FE	Yes	Yes	Yes

Standard errors in parentheses

Notes: The dependent variable is the customer's monthly supply cost given their contract choice, assuming 750kWh usage. The sample includes Eversource Energy residential customers who purchased electric generation service from a competitive supplier in select months between 2011-2018. Zip Income refers to the median household income of the zip code in which a customer lives. Customers living in zip codes with median household income between $80,000-100,000 are the base category. Electric Heating is a dummy variable equal to one if a customer uses electricity as their primary heating source. Hardship is a dummy variable equal to one if a customer is protected from having their power shut off due to nonpayment during winter months. Standard errors are clustered at the zip code level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 1.9: Difference in Initial and Current Contract Prices by Customer Income

	(1)	(2)
New Customer	-12.05***	-12.15***
	(0.409)	(0.406)
Zip Income Below 50k	4.896***	4.549***
	(0.490)	(0.506)
Zip Income Between 50-80k	2.142***	2.041***
	(0.428)	(0.425)
Zip Income Above 100k	-0.436	-0.392
	(0.415)	(0.414)
New × Zip Income Below 50k	8.290***	7.654***
	(0.560)	(0.568)
New × Zip Income Between 50-80k	3.787***	3.542***
	(0.611)	(0.588)
New × Zip Income Above 100k	-2.354***	-2.301***
	(0.530)	(0.523)
September	-0.0940**	-0.0975**
	(0.0462)	(0.0458)
New × September	3.048***	3.010***
	(0.372)	(0.366)
Hardship		2.610***
		(0.226)
New × Hardship		2.508***
		(0.441)
Constant	73.90***	73.82***
	(0.347)	(0.347)
Observations	582805	582805
R-Sq	0.0236	0.0252
F-Stat	381.7	392.7

Standard errors in parentheses

Notes: The dependent variable is the customer's monthly supply cost given their contract choice, assuming 750kWh usage. The sample includes Eversource Energy residential customers who purchased electric generation service from a competitive supplier in June and September 2018. Zip Income refers to the median household income of the zip code in which a customer lives. Customers living in zip codes with median household income between $80,000-100,000 are the base category. Heating is a dummy variable equal to one if a customer uses electricity as their primary heating source. Hardship is a dummy variable equal to one if a customer is protected from having their power shut off due to nonpayment during winter months. Standard errors are clustered at the zip code level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 1.10: Probability of enrolling in a contract that costs more than the default option

	(1)	(2)	(3)
Zip Income Below 50k	0.467***	0.432***	0.429***
	(0.00814)	(0.00835)	(0.00842)
Zip Income Between 50-80k	0.212***	0.200***	0.199***
	(0.00773)	(0.00770)	(0.00770)
Zip Income Above 100k	-0.0817***	-0.0783***	-0.0803***
	(0.00776)	(0.00771)	(0.00774)
September	0.307***	0.306***	0.306***
	(0.00565)	(0.00561)	(0.00561)
Hardship		0.182***	0.170***
		(0.00883)	(0.00959)
Heating			-0.0352***
			(0.00830)
Hardship × Heating			0.0813***
			(0.0228)
Observations	24764	24764	24764
R-Sq	0.276	0.288	0.289
F-Stat	3638.9	3123.7	2240.8

Standard errors in parentheses

Notes: The dependent variable is a dummy variable equal to one if the customer's contract price is higher than the current default generation service price. The sample includes Eversource Energy residential customers who enrolled in a contract for competitive electric generation service in June and September 2018. Low SES is a dummy variable equal to one if a customer lives in a zip code with a median household income below 80,000 USD. Heating is a dummy variable equal to one if a customer uses electricity as their primary heating source. Hardship is a dummy variable equal to one if a customer is protected from having their power shut off due to nonpayment during winter months. Robust standard errors are reported in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 1.11: Relationship between contract price, length, and customer hardship status

	(1) Price	(2) Price	(3) Price	(4) Price
Hardship	8.707**	5.218***	10.69***	5.005**
	(3.902)	(1.630)	(3.932)	(1.946)
Heating	-2.745	-1.375	-2.304	-1.022
	(3.438)	(0.956)	(3.337)	(0.868)
Hardship × Heating	-0.158	-0.0277	-0.521	-0.174
	(5.026)	(2.180)	(4.953)	(2.087)
Long Term			4.427**	3.045**
			(2.078)	(1.370)
Hardship × Long Term			-4.682	0.966
			(3.269)	(1.830)
Observations	109138	109137	109138	109137
R-Sq	0.162	0.600	0.176	0.606
F-Stat	4.245	7.365	3.597	6.730
Sample Month FE	Yes	Yes	Yes	Yes
Supplier FE	No	Yes	No	Yes

Standard errors in parentheses

Notes: The dependent variable is the customer's initial contract price assuming 750kWh usage. The sample includes Eversource Energy residential customers who began a new competitive generation service contract between October 2016 and October 2017. *Long Term* is a dummy variable equal to one if a customer's contract term is longer than 18 months. Hardship is a dummy variable equal to one if a customer is protected from having their power shut off due to nonpayment during winter months. Heating is a dummy variable equal to one if a customer uses electricity as their primary heating source. Standard errors are clustered by the customer's supplier and customer type.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 1.12: Summary of Estimation Sample for Contract Length Data

	Non-Hardship		Hardship	
	Non-Heating	Heating	Non-Heating	Heating
Contract Price	64.51	61.62	74.53	71.16
	(14.43)	(13.37)	(13.29)	(14.45)
Term (Years)	1.492	1.328	1.272	1.172
	(0.528)	(0.560)	(0.524)	(0.494)
Under 6 Months	0.0121	0.0168	0.0400	0.0710
	(0.109)	(0.129)	(0.196)	(0.257)
6 Months	0.0107	0.0235	0.0277	0.0386
	(0.103)	(0.152)	(0.164)	(0.193)
12-17 Months	0.360	0.360	0.303	0.293
	(0.480)	(0.480)	(0.460)	(0.455)
18-23 Months	0.241	0.165	0.244	0.323
	(0.428)	(0.371)	(0.429)	(0.468)
24-35 Months	0.244	0.183	0.133	0.0357
	(0.429)	(0.386)	(0.339)	(0.186)
New Customers	97375	19008	9348	2126
% New Customers	76.1	14.9	7.3	1.7

Table 1.13: Relationship between customer income, contract prices, and renewable energy content

	Town Square Energy				Discount Power			
	(1) Price	(2) Price	(3) Price	(4) Price	(5) Price	(6) Price	(7) Price	(8) Price
Zip Income < $80k	7.742***	1.998***	1.789***	1.926***	5.059***	5.116***	4.764***	4.792***
	(0.363)	(0.217)	(0.212)	(0.223)	(0.250)	(0.247)	(0.253)	(0.260)
September	5.645***	4.189***	4.170***	4.139***	3.037***	2.623***	2.669***	2.665***
	(0.350)	(0.198)	(0.197)	(0.197)	(0.375)	(0.374)	(0.370)	(0.371)
Green		29.83***	29.25***	31.27***		5.839***	5.871***	6.175***
		(0.264)	(0.281)	(0.497)		(0.580)	(0.581)	(0.632)
Hardship			4.064***	4.041***			3.145***	3.143***
			(0.643)	(0.642)			(0.428)	(0.428)
Heating			-0.503	-0.506			-1.132**	-1.136**
			(0.280)	(0.280)			(0.382)	(0.382)
Hardship × Heating			-0.683	-0.554			3.361***	3.372***
			(1.307)	(1.300)			(0.818)	(0.818)
Below $80k × Green				-2.381***				-0.609
				(0.560)				(1.162)
Constant	54.25***	54.12***	54.16***	54.12***	59.66***	59.46***	59.49***	59.48***
	(0.204)	(0.136)	(0.142)	(0.143)	(0.187)	(0.187)	(0.195)	(0.196)
Observations	3999	3999	3999	3999	3648	3648	3648	3999

Standard errors in parentheses

Notes: The dependent variable is the customer's initial contract price, measured as the cost of consuming 750 kWh. The estimation sample uses new Town Square Energy and Discount Power customers who are matched on the basis of their contract price to the zip code-level data from June and September 2018. *Green* equals one if a customer purchases a green energy contract. Green energy contracts include contracts for which over one half of customers purchase 100% renewable electricity. Robust standard errors are reported in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.14: Relationship between a customer's income and whether they choose a green energy contract

	Town Square Energy			Discount Power		
	(1) Green	(2) Green	(3) Green	(4) Green	(5) Green	(6) Green
Zip Income < \$80k	0.119***	0.105***	0.105***	-0.00620**	-0.00570**	-0.00581**
	(0.00262)	(0.00260)	(0.00260)	(0.00189)	(0.00190)	(0.00189)
Hardship		0.175***	0.175***		0.000493	0.000488
		(0.00987)	(0.00986)		(0.00385)	(0.00385)
Heating		-0.0384***	-0.0379***		-0.00374	-0.00398
		(0.00300)	(0.00300)		(0.00283)	(0.00283)
Hardship × Heating		-0.0803***	-0.0800***		-0.00554	-0.00543
		(0.0184)	(0.0183)		(0.00796)	(0.00796)
September		-0.0126***	-0.0179***		0.0509***	0.0520***
		(0.00263)	(0.00273)		(0.00188)	(0.00188)
New Customer (June 2018)			0.00909			0.0204***
			(0.00633)			(0.00323)
New Customer (September 2018)			0.0644***			0.0189
			(0.0100)			(0.0110)
Constant	0.0259***	0.0353***	0.0341***	0.0397***	0.0149***	0.0121***
	(0.00109)	(0.00193)	(0.00209)	(0.00144)	(0.00137)	(0.00135)
Observations	43122	43122	43122	39744	39744	39744

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: The dependent variable is a dummy variable equal to one if a customer enrolls in a green energy contract. The data set used is the zip code-level accounts data merged with the Town Square Energy and Discount Power green customer data. Robust standard errors are reported in parentheses.

1.9 Data Appendix

The data sets used in this paper were compiled using files published in the Connecticut Public Utilities Regulatory Authority online docket system. Table 1.15 below lists the dockets where the raw files can be located for each data set.

Table 1.15: Notes on Data Sources Used

Data Set	Docket	Interrogatories Used
Zip Code-Level	18-06-02	OCC-6, OCC-7, OCC-8, OCC-9
Contract Length	18-06-02	OCC-3,OCC-4, RESA-EDC-36, RESA-EDC-37
Green Customer (Direct) Green Customer (Town Square)	13-07-17 10-03-11RE03	AG-5 OCC-15

Bibliography

Abito, Jose Miguel and Yuval Salant, "The Effect of Product Misperception on Economic Outcomes: Evidence from the Extended Warranty Market," *Review of Economic Studies*, 2019, *86* (6), 2285–2318.

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Perez, "Automated Linking of Historical Data," *NBER Working Paper*, 2019.

_ , Leah Platt Boustan, and Katherine Eriksson, "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration," *American Economic Review*, 5 2012, *102* (5), 1832–1856.

_ , _ , and _ , "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration," *Journal of Political Economy*, 2014, *122* (3), 467–506.

_ , _ , and _ , "To the New World and Back Again: Return Migrants in the Age of Mass Migration," *ILR Review*, 2019, *72* (2), 300–322.

_ , Roy Mill, and Santiago Perez, "Linking Individuals Across Historical Sources: a Fully Automated Approach," Technical Report 24324, National Bureau of Economic Research 2 2018.

Aizer, Anna, Shari Eli, Joseph Ferrie, and Adriana Lleras-Muney, "The Long-Run Impact of Cash Transfers to Poor Families," *American Economic Review*, 4 2016, *106* (4), 935–971.

Anderson, Rachel, Bo Honoré, and Adriana Lleras-Muney, "Estimation and inference using imperfectly matched data," *Working paper*, 8 2022.

Auffhammer, Maximilian and Edward Rubin, "Natural Gas Price Elasticities and Optimal Cost Recovery Under Consumer Heterogeneity: Evidence from 300 million natural gas bills," *National Bureau of Economic Research*, 2018.

Averch, Harvey and Leland L Johnson, "American Economic Association Behavior of the Firm Under Regulatory Constraint," *Source: The American Economic Review*, 1962, *52* (5), 1052–1069.

Bailey, Martha, Connor Cole, Morgan Henderson, and Catherine Massey, "How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data," Technical Report 24019, National Bureau of Economic Research 11 2017.

Baldwin, Susan M, "Are Consumers Benefiting from Competition? An Analysis of the Individual Residential Electric Supply Market in Massachusetts," 2018.

Brown, David P. and Andrew Eckert, "The effect of default rates on retail competition and pricing decisions of competitive retailers: The case of Alberta," *Energy Policy*, 2018, *118* (December 2017), 298–311.

Burger, Scott P., Jesse D. Jenkins, Carlos Batlle, and Ignacio J. Pérez-Arriaga, "Restructuring revisited part 1: Competition in electricity distribution systems," *Energy Journal*, 2019, *40* (3), 31–54.

Bushnell, James, Erin T Mansur, and Kevin Novan, "Review of the Economics Literature on US Electricity Restructuring," Technical Report 2017.

Byrne, David P. and Leslie A. Martin, "Consumer search and income inequality," *International Journal of Industrial Organization*, 2021.

_ , _ , and Jia Sheen Nah, "Price Discrimination by Negotiation: A Field Experiment in Retail Electricity," *SSRN Electronic Journal*, 3 2022.

Chamberlain, Gary, "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 3 1987, *34* (3), 305–334.

Chetty, Raj, John N Friedman, and Jonah E Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," Technical Report 19424, National Bureau of Economic Research 9 2013.

_ , _ , Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan, "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star," *The Quarterly Journal of Economics*, 2011, *126* (4), 1593–1660.

Chretien, Larry, Daria Mark, Eugenia Gibbons, Yaima Braga, and Kate Marcellino, "Green Municipal Aggregation in Massachusetts," 2020, (February), 46.

Collins, William J and Marianne H Wanamaker, "Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data," Technical Report 19124, National Bureau of Economic Research 6 2013.

Dempster, A P, N M Laird, and D B Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1977, *39* (1), 1–38.

Deryugina, Tatyana, Alexander MacKay, and Julian Reif, "The long-run dynamics of electricity demand: Evidence from municipal aggregation," *American Economic Journal: Applied Economics*, 2020, *12* (1), 86–114.

Distributed Energy Financial Group LLC, "Annual Baseline Assessment of Choice in Canada and the United States," Technical Report July, Distributed Energy Financial Group LLC 2015.

Dobbie, Will and Roland G Fryer, "Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone," *American Economic Journal: Applied Economics*, 2011, *3* (3), 158–187.

Dormady, Noah, Matthew Hoyt, Alfredo Roa-Henriquez, and William Welch, "Who pays for retail electric deregulation? Evidence of cross-subsidization from complete bill data," *Energy Journal*, 2019, *40* (2), 161–194.

EIA, "Status of State Electric Industry Restructuring Activity as of February 2003," 2003, (February).

_ , "Annual Electric Power Industry Report, Form EIA-861 Detailed Data Files," 2014.

Enamorado, Ted, Benjamin Fifield, and Kosuke Imai, "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records," *American Political Science Review*, 2019, *113* (2), 353?371.

Fellegi, I P and A B Sunter, "A Theory for Record Linkage," *Journal of the American Statistical Association*, 1969, *64*, 1183–1210.

Ferrie, Joseph and Karen Rolf, "Socioeconomic status in childhood and health after age 70: A new longitudinal analysis for the U.S., 1895G2005," *Explorations in Economic History*, 2011, *48* (4), 445–460.

Ferrie, Joseph P, "A New Sample of Males Linked from the Public Use Microdata Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules," *Historical Methods*, 1996, *29* (4), 141–156.

Giulietti, Monica, Jesus Otero, and Michael Waterson, "Pricing behaviour under competition in the UK electricity supply industry," *New Series*, 2010, *62* (3), 478–503.

Hansen, Lars Peter, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 1982, *50* (4), 1029–1054.

Hortaçsu, Ali, Seyed Ali Madanizadeh, and Steven L. Puller, "Power to choose? An analysis of consumer inertia in the residential electricity market," *American Economic Journal: Economic Policy*, 2017, *9* (4), 192–226.

Hsu, David, "choice aggregation and its spread to other states Version spread to other states," 2022.

Jaro, Matthew A, "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, 1989, *84* (406), 414–420.

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz, "Experimental Analysis of Neighborhood Effects," *Econometrica*, 2007, *75* (1), 83–119.

_ , Jens Ludwig, and Lawrence F Katz, "Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment," *The Quarterly Journal of Economics*, 2005, *120* (1), 87–130.

Lahiri, P and Michael D Larsen, "Regression Analysis with Linked Data," *Journal of the American Statistical Association*, 2005, *100* (469), 222–230.

Larsen, Michael D and Donald B Rubin, "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association*, 2001, *96* (453), 32?41.

Littlechild, Stephen, "The regulation of retail competition in US residential electricity markets," Technical Report 2018.

MacKay, Alexander and Ignacia Mercadal, "Deregulation, Market Power, and Prices: Evidence from the Electricity Sector," *SSRN Electronic Journal*, 2021.

Neter, John, E Scott Maynes, and R Ramanathan, "The Effect of Mismatching on the Measurement of Response Errors," *Journal of the American Statistical Association*, 1965, *60* (312), 1005–1027.

Newey, Whitney K., "Consistency of two-step sample selection estimators despite misspecification of distribution," *Economics Letters*, 5 1999, *63* (2), 129–132.

Newey, Whitney K and Daniel McFadden, "Large Sample Estimation and Hypothesis Testing," in Robert F Engle and Daniel L McFadden, eds., *Handbook of Econometrics*, number 4. In 'Handbooks in Economics,.', Amsterdam, London and New York: Elsevier, {N}orth-{H}olland, 1994, pp. 2111–2245.

Office of Consumer Counsel, "OCC FACT SHEET: In Support of Increased Consumer Protection for the Residential Retail Choice Electric Supply Market," Technical Report, State of Connecticut Office of Consumer Counsel 2020.

O'Shaughnessy, Eric, Jenny Heeter, Julien Gattaciecca, Jenny Sauer, Kelly Trumbull, and Emily Chen, "Empowered communities: The rise of community choice aggregation in the United States," *Energy Policy*, 9 2019, *132*, 1110–1119.

SAS, "SAS 9.4 Functions and CALL Routines: Reference, Fifth Edition," Technical Report, SAS Institute Inc. 2019.

Scheuren, Fritz and William Winkler, "Regression analysis of data files that are computer matched," *Survey Methodology*, 1993, *19.*

Singer, Stephen, "Why is Connecticut No. 1 in the cost of electricity among the 48 lower states? Here are 7 reasons," 4 2021.

Su, Xuejuan, "Have customers benefited from electricity retail competition?," *Journal of Regulatory Economics*, 2015, *47* (2), 146–182.

Swadley, Adam and Mine Yücel, "Did residential electricity rates fall after retail competition? A dynamic panel analysis," *Energy Policy*, 12 2011, *39* (12), 7702–7711.

Tsai, Chen-Hao and Yi-Lin Tsai, "Competitive Retail Electricity Market Under Continuous Price Regulation," *SSRN Electronic Journal*, 2017.

Upton, Gregory, Kenneth M. Rosen, and Brittany Tarufelli, "Electricity Market Restructuring and Retail Rates," *SSRN Electronic Journal*, 3 2021.

U.S. Energy Information Administration, "Electric Power Annual 2020," Technical Report 2020.

Wang, Xueting, "Long term contracts in retail electricity," 2021.

Winkler, William E, "Overview of record linkage and current research directions," Technical Report, U.S. Census Bureau 2006.

Chapter 2

Do Regulators Make Markets? The Effects of Regulatory Policy on Retail Electricity Markets

2.1 Introduction

For households in 13 U.S. states, the most visible outcome of electric sector restructuring was the introduction of consumer choice in retail electricity markets. In those states, consumers may enter contracts with competitive electric retailers, in the same way that they choose a cellular service provider. Unlike in other consumer choice markets, however, residential electric users are assigned a default service provider that sells electricity at regulated prices. The effect of this default option on consumers' choices is very strong: in 2020—nearly two decades after most states enabled retail electric choice—less than 20 percent of eligible households switched away from their default service provider (U.S. Energy Information Administration, 2020).

The fact that so few consumers choose to purchase competitive retail electric service could be one reason why economists have found limited evidence that retail electricity market deregulation has benefited U.S. consumers (Upton et al., 2021; MacKay and Mer-

cadal, 2021; Bushnell et al., 2017; Su, 2015; Swadley and Yücel, 2011). If there is insufficient consumer demand for competitive retail contracts, then firms may lack incentives to offer innovative or lower cost services. However, some states, including Massachusetts, Connecticut, Illinois, and Ohio, have experienced significant competitive retail electric market growth over the last decade, each with nearly half of customers choosing to purchase from competitive electric retailers in some years.[1] Thus, it may be possible to learn whether retail electric competition has benefited consumers in those states, if one can also account for why more customers purchase competitive retail electric service. This paper provides an explanation—those states adopted regulatory policies that encouraged their competitive retail electricity markets to grow.

To illustrate the importance of regulatory policies for competitive retail electricity market development, it is useful to compare states with similar electricity market structures, regulatory environments, and public policy goals. Consider Connecticut and Rhode Island, which opened their retail electricity markets to competition around 2000. Both states belong to the ISO-New England wholesale electricity market and have renewable portfolio standards that require 35 percent of in-state electricity sales to come from renewable energy sources by 2030. Yet, in recent years, competitive electric retailers served nearly half of residential customers in Connecticut, compared to only 2.5 percent of customers in Rhode Island (EIA, 2014). One reason for this difference is that Connecticut regulators required their utilities to adopt "purchase of receivables" rules, which guarantee that competitive retailers are paid even if their residential customers fail to pay their electric bills, while regulators in Rhode Island did not.

The econometric methods in this paper generalize the idea of comparing retail electricity market outcomes for states that have similar electricity market structures and regulatory environments, but differ in terms of whether and when they adopted pro-competitive

[1]The majority of residential customers also purchase competitive retail electric service in Texas, however there is no regulated default service option for residential customers in Texas.

regulatory policies. To perform the analysis, I construct a unique panel data set that contains measures of competitive retail electricity market participation for each utility- or state-wide market where residential customers gained the ability to choose their electric retailer. I combine these data with hand-collected information on whether and when three regulatory policies were implemented in each market. I then test for the impact of the policies on the share of customers choosing to purchase competitive retail electric service, entry and exit by competitive retailers, and competitive retailer prices over time.

Importantly, the empirical analysis in this paper is based on detailed descriptions of states' restructuring timelines, including the dates for when each utility within a state adopted specific regulatory policies to encourage competitive retail electric market growth. This is necessary to properly identify the effect of pro-competitive regulatory policies on competitive retail electric market outcomes because utilities operating within the same state did not necessarily adopt the same set of policies. Nor did they implement them at the same time. Thus, a major contribution of this paper is that it provides a more comprehensive analysis of competitive retail electric market development in the United States compared to previous research in this area (Upton et al., 2021; Su, 2015; Swadley and Yücel, 2011). Additionally, the data used in this paper span from 2001 to 2019, which is crucial for the analysis because residential customer participation did not take off in most states until the mid-2010s.

My analysis also complements recent studies that evaluate the effect of U.S. electric sector restructuring on retail electricity prices, such as MacKay and Mercadal (2021), Upton et al. (2021), and Dormady et al. (2019). These papers find that default service provider prices increased as a result of restructuring, however they do not consider whether the benefits of restructuring may have accrued to customers through lower prices for *competitive* retail electric service. While the data used in my analysis do not allow the comparison of default service provider and competitive electric retailer prices, my find-

ings suggest that at least one regulatory policy—the authorization of opt-out municipal aggregation programs—may have lowered competitive retailer prices, compared to other restructured states that did not authorize these programs.

The rest of the paper is arranged as follows. Section 2.2 describes states' approach to retail electric market restructuring in the U.S., the barriers to retail electric competition that emerged as a result, and the regulatory policies that were adopted to address these barriers. Section 2.3 describes the data, and Section 2.4 discusses the empirical framework for the analysis. Section 2.5 presents the estimation results, which are discussed in Section 2.6. Section 2.7 concludes.

2.2 Background

In this section, I provide background on U.S. electric industry restructuring, with a focus on retail electricity market deregulation, and the regulatory policies that were adopted to foster retail electric market competition.

2.2.1 Background on U.S. Electricity Sector Restructuring

The electric supply industry consists of four distinct operations: electric power generation, high-voltage transmission wires, lower-voltage distribution wires, and retail sales to customers. For most of the 20th century, over 75 percent of U.S. customers received bundled electric service from investor-owned utilities that were closely regulated by state-level public service commissions.[2] These utilities operated as local monopolies in all four sectors of the electric industry, and were required to provide retail electric service to customers living in the footprint of their exclusive distribution service territories.

[2]The rest of consumers received power from government-owned utilities (municipal, state, and federal) and nonprofit electric cooperatives, some of which were regulated at the state level.

Most of the investor-owned utilities operated under cost-of-service regulation. Under cost-of-service regulation, state utility commissions approved electric rates that provided for the recovery of prudently incurred operating costs plus a regulated return on capital expenditures, such as the construction of new power plants and improvements in the utility's distribution network. Although cost-of-service regulation offers benefits such as electric system reliability and resource adequacy, it may also encourage utilities to over-invest in capital—a phenomenon known as "gold plating" in the industry—if the rate of return allowed by the regulator is too high (Averch and Johnson, 1962).

Inspired by the deregulation of other network industries, such as telecommunications, airlines, railroads, and gasoline retailing, many states saw an opportunity to lower the cost of electricity for end users by encouraging competition in the generation and retail sales functions of the electric industry.[3] Proponents of restructuring challenged the view that electric utilities constituted a natural monopoly over generation because of technological changes that reduced the minimum efficient scale of power plants.[4] Additionally, competition in retailing was viewed as necessary to support the pass through of lower generation costs to lower electricity prices for end users. Thus, about half of U.S. states began a process of partial deregulation and regulatory reform in the late 1990s, commonly referred to as electricity market "restructuring".

The primary goal of states' restructuring measures was to increase competition in electric generation through the creation of organized wholesale electricity markets. States required their investor-owned utilities to forfeit control of their transmission operations to an independent system operator (ISO) or regional transmission organization (RTO). These ISOs and RTOs became responsible for coordinating the dispatch of generators—which

[3]The distribution and transmission functions of the electric industry were largely ignored, as these were still believed to merit natural monopoly status due to economies of scale.

[4]These changes included the deployment of cost-effective combined cycle natural gas-fired power plants and improvements to the transmission system that allowed more flexibility in the geographic siting of generation units.

guaranteed nondiscriminatory grid access for non-utility, independent power producers—and running centralized auctions to determine the wholesale price of electricity. Because the U.S. transmission system crosses state lines, the geographic footprints of organized wholesale electricity markets cross state boundaries as well (see Figure 2.1).[5]

Some states took additional measures to enable competition in electric retailing, which allows customers to purchase "energy-only" service from non-utility, competitive electric retailers, who produce or procure power in wholesale electricity markets to sell to end users. These states typically required their investor-owned utilities to divest or structurally separate their electric generation operations from their distribution and retail businesses. The utilities continued to operate as regulated monopolies for their wires-based services (e.g., transmission and distribution) because these activities were seen as benefiting from economies of scale.[6] The utilities also continued to provide retail service for their distribution customers, however their retail rates were divided into a regulated charge for "delivery" service and a market-based charge for the "generation" or "energy" portion of customers' bills. These changes allowed customers to receive "delivery-only" service from their distribution utility, should they choose to purchase "energy-only" service from a non-utility, competitive electric retailer.

Initially, about half of U.S. states enacted legislation to enable retail electric competition. However, many states halted or even reversed their decisions because of the blackouts and record high electricity prices in California shortly after the state opened its market to competition in 2000-2001. Other states limited the classes of customers for which retail competition would be introduced. Ultimately, only 13 states and the District of Columbia enabled retail competition for all electric customers (see Figure 2.2). Unlike

[5]The geographic boundaries of organized wholesale electricity markets have expanded over time. For example, parts of Arkansas, Louisiana, and Mississippi were incorporated into the Midcontinent ISO (MISO) market in 2013. Participants in the MISO market include vertically integrated utilities operating under cost-of-service regulation, independent power producers, and competitive electric retailers.

[6]Today, the emergence of distributed energy resources has sparked debates over whether the distribution function should be opened to competition (Burger et al., 2019).

wholesale electricity markets, the boundaries of retail electricity markets do not cross state lines. As a result, the states that enabled retail competition comprise a subset of the states that restructured their wholesale electricity markets.

The process by which these states implemented retail electric competition is described as follows. First, the state legislatures passed laws that enabled regulators to implement retail choice and outlined general provisions, such as which customers would be eligible to purchase from competitive electric retailers and dates when retail market access would begin. States then phased in retail competition gradually, beginning with large industrial and commercial electric users, followed by small commercial and residential customers, sometimes several years later. Some states created special offices within their regulatory commissions to focus specifically on electric and natural gas retail market issues. The first such office was the New York Office of Retail Market Development, which was established in December 2003. That office published a 2004 Policy Statement that included a list of best practices for promoting retail electric market competition that continue to inform regulatory policy nearly two decades later. I describe three of the most influential of these policies in detail below.

2.2.2 Market-Based Default Service Rates

To ease the transition to retail electric competition for small electric users, states required their distribution utilities to offer a regulated, default service tariff for residential and small commercial customers who did not choose a competitive electric retailer. During the first years of retail competition, most states imposed multi-year price caps on utilities' default service rates for these customers (EIA, 2003). In some states, the price caps were set below utilities' pre-restructuring prices because of stipulations in their restructuring legislation that mandated rate reductions for residential customers. As a result, entrant,

non-utility electric retailers could not compete with the incumbent utilities' default service option while the price caps were in place.

Although the intention was to eventually phase out the default service tariffs over time, only Texas ended retail electricity rate regulation entirely.[7] The remaining states required their utilities to adopt market-based default service prices that were, in theory, designed to strike a balance between protecting small electric users and promoting retail electric competition. In practice, regulators established rules for setting a utility's default service prices to reflect near-term wholesale electricity market conditions plus a markup to create a more level playing field for non-utility competitive electric retailers.

Depending on the state, regulators update utilities' default service rates on either a monthly or biannual basis (Littlechild, 2018). This encourages customers who prefer to know the price of electricity in advance to purchase long-term fixed-price contracts from competitive electric retailers (Wang, 2021; Brown and Eckert, 2018). Additionally, customers may search for contracts that offer short-term savings if wholesale electricity prices decline between default service rate updates, or if the markup embedded in the default service price is especially high (Tsai and Tsai, 2017). Thus, the adoption of "market-based" default service prices is one regulatory policy that regulators have used to increase small electric customer demand for competitive retail electric service.

2.2.3 Purchase of Receivables Rules

Whereas the adoption of market-based default service prices aimed to increase the demand for competitive retail electric service contracts, the second policy that I discuss aimed to increase the number of competitive retailers that serve small electric customers.

[7]Electric distribution utilities in Texas were required to divest their retail operations to form affiliated, competitive retail companies. When the Texas market opened to competition on January 1, 2002, all customers were moved from their regulated electric utility to the affiliated retail electric provider. Until December 31, 2006, the affiliate companies continued to offer a default service for residential customers at rates that were regulated by the Public Utility Commission of Texas. After that period, rate regulation was eliminated and customers who had not chosen any provider remained with the affiliate company.

In addition to providing a default service option, distribution utilities in restructured retail electricity markets are responsible for billing small electric customers on behalf of competitive electric retailers. The reason for this is that distribution utilities use ratepayer-funded systems for collecting payments from customers, including the ability to terminate a customer's service if they default on their electricity bill. Competitive electric retailers, by comparison, do not have the ability to penalize their customers for nonpayment. This places them at a cost disadvantage compared to the distribution utility, because they must perform costly credit screening processes or lengthy enrollment processes to manage the risk of customer nonpayment (Littlechild, 2018).

To address this problem, regulators in many states have adopted purchase of receivables rules to remove the risk of nonpayment for competitive electric retailers whose customers fail to pay their electric bills. A typical purchase of receivables program requires the regulated distribution utility to purchase a competitive electric retailer's total accounts receivable (less a small percentage discount), regardless of whether the retailer's customers have paid their bills in full. Such programs arguably create a moral hazard for competitive electric retailers, because they incentivize maximizing sales without consideration of whether customers will pay or not. Meanwhile, the costs incurred by the non-paying customers are socialized across all electric distribution customers through regulated delivery service rates.

Nine of the retail choice states and the District of Columbia have implemented a purchase of receivables program for at least one of their investor-owned utilities. The rules for these programs differ both by state and utility. For example, distribution utilities in New Jersey only collect payments for customers of competitive retailers whose charges are less than 120 days past due. After 120 days, the distribution utility stops billing the non-paying customer on behalf of the competitive retailer, thereby ending its obligation to purchase the retailer's accounts receivable. When this happens, competitive retailers are

allowed to terminate service with the delinquent customer, upon which the customer is returned to their utility's default service option. The number of days that must pass before a retailer is allowed to end its relationship with the delinquent customer differs by utility in New Jersey, and remains an ongoing source of dispute.[8]

To summarize, regulators adopted purchase of receivables programs to encourage competitive retailers to serve small electric customers, by making such customers less risky to serve.

2.2.4 Opt-Out Aggregation Programs

Another policy that regulators have used to increase small customer participation in competitive retail electric markets is the authorization of opt-out aggregation programs. Such programs allow a municipality or community to negotiate the purchase of electricity with a competitive retailer on behalf of its residents. This negotiated price becomes the new default price for all residents. Residents who wish to a purchase a different retail product, including their default service provider's regulated option, must "opt-out" of the aggregation program.

In addition to making small electric users cheaper to serve (because these customers' load profiles are more predictable when aggregated), aggregation programs have the potential to benefit consumers because they increase their bargaining power and eliminate search frictions. For example, Deryugina et al. (2020) show that aggregation program customers in Illinois paid lower prices than the cost of their distribution utility's default service option. Although these realized savings came largely from the timing of the programs—wholesale electricity prices declined while the distribution utility's default service prices were fixed—they still led to lower prices for participating customers who otherwise would not have shopped for a competitive retail electric contract on their own.

[8]See, for example, the Supplemental Comments from Gateway Energy Services Corporation filed on October 5, 2010, before the New Jersey Board of Public Utilities.

Seven of the full retail choice states have enacted legislation to authorize the creation of opt-out aggregation programs to date. However, participation in aggregation programs differs across these states. For example, aggregation programs have been cited as the primary driver of competitive retail electricity market growth in Illinois, Ohio, and Massachusetts, but not in New York, New Jersey, or Rhode Island (Distributed Energy Financial Group LLC, 2015).[9] These differences may be attributed to differences in states' initial aggregation laws, which most states have strengthened over time. For example, Rhode Island initially permitted aggregation programs in its restructuring legislation, passed in 1996. However, the law was amended in 2002 and 2017 to remove legal roadblocks to opt-out aggregation. Additionally, some states place restrictions on aggregation contracts. For example, aggregation contract prices in New Jersey must be lower than the local distribution utility's default service rate, unless they include renewable energy content exceeding the state's renewable portfolio standard.[10] This requirement, coupled with the fact that the default service rates in New Jersey are set via competitive procurement auctions, may leave too little room for retailers to offer lower prices.

2.3 Data

In this section, I describe the data set that I use to measure the effect of pro-competitive regulatory policies on deregulated retail electricity markets in the United States.

Most of the key variables in the analysis come from the Energy Information Administration Form 861 (EIA-861), which contains detailed measures of retail sales for all distribution utilities and non-utility, competitive retail electric providers in the United States.

[9]The seventh state is New Hampshire, which authorized aggregation programs in 2019 and thus has experienced limited participation to date. Additionally California and Virginia permit aggregation programs, however there is no retail competition for residential customers outside of these programs.

[10]NJ Rev Stat § 48:3-92 (2013)

The data are reported annually for years between 1990-2020, however I use observations from 2001 to 2020 because most states did not enable retail electric market competition for residential customers until after 2001.

I define the sample to include observations from U.S. jurisdictions that both allow retail electric choice for residential customers and provide a regulated default service option for customers who do not choose a competitive electric retailer. These jurisdictions include the District of Columbia and all the full retail choice states in Figure 2.2 except for Texas, because the default service provider in Texas is an unregulated affiliate of the customer's distribution utility, and Maine, which I exclude because of inconsistencies in the EIA-861 data. The purpose of defining the sample in this way is because the outcomes that I measure only apply to states where retail electric competition exists.

The unit of analysis is a retail electricity market, which I define as either the distribution service territory of an investor-owned utility or a state-wide retail electricity market, depending on the part of analysis. Whenever possible, I use a utility's distribution service territory as the unit of analysis, because most states have multiple utilities and regulators may enforce different rules for different utilities. For example, Pennsylvania has 11 investor-owned utilities, of which only five were required to adopt purchase of receivables programs. Additionally, some investor-owned utilities have distribution operations in multiple states, however I count these observations as distinct utility-state entities because utilities are regulated at the state-level. Utilities that merge at any point are treated as a single, merged entity throughout the sample. The final sample includes 43 distribution utility markets in 11 states and the District of Columbia.

For each utility in the sample, I hand-collect data on whether and when the utility was required to adopt each of the three regulatory policies discussed in Section 2.2. Sources include the Energy Information Administration, National Renewable Energy Laboratory, news articles, press releases, state archives, and peer-reviewed research papers such as Up-

ton et al. (2021) and Su (2015). Because parts of the analysis use a state-wide retail electricity market as the unit of analysis, I also assign policy adoption dates for each state (see Table 2.1 for the dates used). Because some states implemented policies at different times for different utilities, this creates the potential for measurement error, which I consider when analyzing the results in Section 2.5.

2.3.1 Key Variables

The main outcome of interest is the share of utility customers who choose to purchase competitive retail electric service. This metric is called the customer "switching share", because it measures the fraction of customers who actively switched their retail electric service provider. States use aggregate switching shares to monitor their retail electricity markets and measure customer engagement with competitive electric retailers. This metric is also straightforward to compute using the EIA-861 data, which identify the number of customers and sales in megawatthours by customer class (residential, commercial, or industrial) and service type (delivery-only or bundled service) for each utility. I am therefore able to compute the switching share for each utility by dividing the number of delivery-only customers (or sales) by the total number of customers (or sales) for each customer class.

The other outcome variables are the number of competitive electric retailers that operate in each state and the average price that they charge for retail electric service. The EIA-861 data reports these variables for each state and customer class, but not for each local utility. I therefore use this data in the portion of the analysis that uses a state-wide retail electricity market as the unit of analysis. Also, the competitive retailer price data are missing observations, because there are some states and years where no residential customers purchase competitive retail electric service. This can be seen by plotting the data in Figure 2.3.

Finally, I obtain data on prices for a utility's bundled and delivery-only services from the EIA-861 data to use as additional explanatory variables in the analysis. I select these variables because a utility's bundled service price includes the cost of both its energy and delivery services. Thus, the difference between a utility's bundled service price and its delivery-only service price should approximately equal the cost of the utility's energy-only service price (e.g., the price of its default service option). This measure cannot be computed for all utilities and years, because utilities do not report delivery-only service prices if they do not make any delivery-only sales. Also, the delivery-only service prices include outliers, as well as negative values. For these observations, I use the state average delivery-only service price in lieu of the utility's delivery-only service price.

The utility-level data set is summarized in Table 2.2. Summary statistics are reported separately for residential and non-residential customers, where non-residential customers include commercial and industrial customers combined. Statistics are also reported separately for observations from 2002, 2011, and 2019, to illustrate trends over time. These years correspond to (i) the first year that all states' retail markets were open to competition, (ii) the first year where all utilities had adopted market-based default service rates, and (iii) the last year of the sample.

2.4 Methods

2.4.1 Effects of Policies on Competitive Retail Market Size

First, I evaluate the effects of adopting pro-competitive regulatory policies on the size of a competitive retail market, which I measure using the fraction of utility customers who purchase generation service from a competitive retailer. As discussed in the previous section, this outcome is also referred to as the customer switching share because it measures

the share of customers who actively "switch" away from the utility's default generation service option.

To begin, I estimate the following equation using the utility-level data,

$$Switching\ Share_{ijt} = \alpha_{ij} + \lambda_{jt} + \beta_{1j} Market\ Rates_{it} + \beta_{2j} POR_{it} + \beta_{3j} Agg_{it} + u_{ijt} \quad (2.1)$$

where $Switching\ Share_{ijt}$ is the fraction of utility i's customers of type j who purchase energy-only service from a competitive retailer during year t. The subscript j refers to a customer type, which is either residential, commercial, or industrial. *Market Rates, POR,* and *Agg* are dummy variables equal to one for years during which a utility i has adopted market-based default service prices, a purchase of receivables program, or opt-out municipal aggregation programs, respectively. The fixed effects, α_{ij}, which are defined for each utility and customer class combination, control for differences in the fixed characteristics of utility markets, such as the total number of utility customers or the physical size of a utility's service territory. The fixed effects, λ_{jt}, which are defined for each customer type and year combination, control for macroeconomic trends, such as changes in natural gas prices or federal regulations over time. Standard errors are clustered at the state level to allow for serial correlation and to account for the fact that the error terms u_{ijt} may be correlated for utility markets that are located within the same state.

The coefficients of interest are β_{1j}, β_{2j} and β_{3j}, and they measure the average change in the outcome variable after a utility adopts one of the three pro-competitive regulatory policies. I purposely do not include any control variables in this baseline specification to provide a descriptive analysis of what actually occurred in restructured retail electricity markets before and after implementing each policy.

Because the stated purpose of each policy change was to encourage competitive retail electric market growth, I expect these coefficients to be (weakly) positive. I expect that the adoption of market-based default service rates had a larger effect for commercial

and industrial customers than for residential customers. This is because utilities typically adopted market-based default service rates that were less favorable for large electric customers, since they have more resources to search for competitive retail products. Also, commercial and industrial customers use more electricity, and therefore have a strong financial incentives to respond to changes in electric rate design.

I expect that β_{2j} and β_{3j}, which measure the effects of purchase of receivables and opt-out aggregation programs on the customer switching share, will be positive for residential and commercial customers, but not significantly different from zero for industrial customers. The reason for this prediction is purely technical: these programs only applied to customers who consume a sufficiently small quantity of electricity.

There are multiple reasons that the coefficient estimates could differ from the "true" coefficients in Equation (2.1). The first is omitted variable bias, caused by the fact that a utility's default service price affects customers' incentive to shop for competitive retail electric service. Although the EIA-861 data do not include default service prices, I can proxy for a utility's default service price using its bundled and delivery-only service prices. This is because a utility's bundled service price reflects the cost of providing energy and delivery services, and a utility's default service price is the cost of its regulated, energy service. Thus, the difference between a utility's bundled and delivery-only service prices should approximately equal its default service price. I therefore estimate a version of (2.1) that includes the utility's current bundled and delivery-only service prices as controls.

Another reason that the coefficient estimates in (2.1) could be biased is if the regulatory policies are correlated with the error terms u_{ijt}. This would be the case if the policies were adopted because customer switching was particularly low in the preceding years. I do not suspect this to be an issue for the policies that I study, however, because these policies were developed through regulatory proceedings that spanned several years, and so the date for when each policy was took effect was usually known well in advance of its implemen-

tation. For example, the date for when a utility's transitional default service prices would expire was often specified in a state's restructuring legislation. In some states, purchase of receivables programs were debated for years before they were adopted.[11] Statutory provisions to authorize the formation of opt-out aggregation programs were either included in state's restructuring legislation, or adopted due to the lobbying efforts of individuals who were generally not affiliated with the electric industry (Hsu, 2022).

Testing for delayed effects of regulatory policies

One limitation of the model in (2.1) is that the parameters measure the average change in the share of customers choosing to purchase competitive retail electric service before and after the adoption of each regulatory policy. However, some of the regulatory policies may have a delayed effect on customer switching shares, because they aimed to increase entry by competitive retailers, and the retailers may take months or years to apply for and obtain the requisite licenses to operate in a new state market. I therefore estimate another model that allows for the effect of the regulatory policies to differ between the first two years of policy adoption versus later years,

$$Switching\ Share_{ijt} = \alpha_{ij} + \lambda_{jt} + \tilde{\beta}_j(Years\ 1{-}2) \times X_{it} + \beta_j(Years\ 3{+}) \times X_{it} + Z'_{ijt}\delta_j + u_{it} \quad (2.2)$$

where *Years 1-2* is a vector of indicator variables such that, when interacted with the policy variables X_{it}, equals one if utility i adopted a given policy during year t or $t-1$ and zero for all other years. The vector $Years\ 3{+}$ is defined analogously, but for years after the first two years of each policy's implementation. The vector Z_{ijt} is a set of controls including the utility's bundled and delivery-only service prices, which, as discussed previously, proxy for the utility's default service price.

[11] For example, Massachusetts opened a proceeding to establish rules for implementing purchase of receivables programs for all of its utilities in 2010, but did not implement those programs until nearly five years later.

The coefficients of interest are $\tilde{\beta}_j$ and β_j. For policies that satisfy $\tilde{\beta}_j < \beta_j$, this means that the effect of the policy increased over time. I expect this to be the case for purchase of receivables programs for residential and commercial customers, because they aimed to increase entry by competitive retailers to serve these customer classes. I also expect the effect of authorizing opt-out aggregation programs to increase over time, because a municipality must hold an election for voters to approve its aggregation program before it can begin the process of switching its residents to a competitive electric retailer. With regards to the adoption of market-based default service prices, I expect the effect to be increasing over time for residential customers, because few competitive retailers served residential customers before utility's transitional rates expired, and residential demand responds slowly to changes in pricing (Deryugina et al., 2020). For commercial and industrial customers, I do not necessarily expect the effect of adopting market-based rates to be increasing over time, because these customers are more price elastic and virtually all utilities had already adopted market-based rates for large electric users by the time that they implemented market-based default service prices for small electric customers (which is how I defined *Market Rates*).

2.4.2 Effects of Policies on Competitive Retailer Entry and Exit

The first part of this section described methods to evaluate the impact of pro-competitive regulatory policies on the extent to which consumers participate in competitive retail electric markets. For the next part of the analysis, I evaluate whether the policies increased competition among suppliers, in the sense that they encouraged additional retailers to enter a market. Because the EIA-861 data report sales for each competitive retailer at the state-level, I use a state's entire retail electric market as the unit of analysis.

Specifically, I estimate,

$$\Delta \#Retailers_{sjt} = \alpha_{sj} + \lambda_{jt} + \beta_{1j} MarketRates_{st-1} + \beta_{2j} POR_{st-1} + \beta_{3j} Agg_{st-1} + u_{sjt} \quad (2.3)$$

where $\Delta Retailers_{sjt} = Retailers_{sjt} - Retailers_{sjt-1}$ measures the change in the number of competitive retailers serving customers of type j in state s between year t and $t-1$. The coefficients α_{sj} are state- and customer class-specific fixed effects, and are included to control for unobserved factors that affect retailer entry, such as the fact that states have different licensing requirements for competitive electric retailers. The coefficients λ_{jt} are year- and customer class-specific fixed effects, and control for changes in federal regulations and macroeconomic trends, such as changes in interest rates that could affect firm's abilities to obtain credit. Standard errors are clustered by state to allow for serial correlation over time.

The coefficients of interest are β_{1j}, β_{2j}, and β_{3j}, and they measure the increase (or decrease) in number of competitive retailers in years following the adoption of the policies. As before, the coefficients are allowed to vary by customer class, because the policies targeted small electric customers. As a result I do not expect any of the coefficients to be significantly different from zero for industrial customers, unless there are spillover effects from retailer entry into the commercial and residential markets. For residential and commercial customers, I only expect purchase of receivables programs to have a positive effect on retailer entry because the explicit aim of the policy is to increase retailer entry by eliminating the financial risk of serving small customers. Although β_{1j} could be positive, I do not expect the magnitude of the coefficient to be very large, because market-based default service prices increase consumer demand for competitive retail products but do not address the issue that small customers are more costly to serve. Finally, the coefficient β_{3j} could be negative, because all of the customers in an aggregation program are served by a

single retailer that is selected through a competitive process, which may favor incumbent retailers rather than encourage retailer entry.[12]

2.4.3 Effects of Policies on Competitive Retailer Prices

The final part of the analysis aims to evaluate whether the pro-competitive policies led to lower competitive retailer prices for residential consumers. To do this, I regress the difference between state average energy-only service prices for residential and non-residential customers on the regulatory policies, as follows:

$$Price_{res,st} - Price_{nonres,st} = \alpha_s + \lambda_t + \beta_1 Market\ Rates_{st} + \beta_2 POR_{st} + \beta_3 Agg_{st} + u_{st} \quad (2.4)$$

where $Price_{res,st}$ is the average energy-only price for residential customers in state s during year t and $Price_{nonres,st}$ is the average energy-only price for commercial or industrial customers in the same state and year. The parameters α_s and λ_t are state and time fixed effects, respectively. The coefficients β_1, β_2, and β_3 measure the differential effect of the regulatory policies on competitive retailer prices for residential customers. While there is no clear theoretical prediction for whether the introduction of market-based default service prices or adoption of purchase of receivables rules should lead to lower prices for customers, aggregation programs are likely to result in (weakly) lower prices for customers. This is because aggregators have more resources to negotiate for better prices than customers would obtain if they were to shop for competitive retail electric service on their own.

The reason for taking the difference between residential and non-residential customer prices is to control for state-specific, time-varying factors that influence competitive retailer prices and are unrelated to the regulatory policies. These factors include changes

[12]See, for example, Pennsylvania Public Utility Commission Chairman Robert F. Powelson's municipal aggregation statement published on March 17, 2011.

in wholesale electricity prices, as well as changes in policies that increase the cost of providing competitive retail electric service, such as the cost of complying with a state's renewable portfolio standard. This strategy is valid if energy-only prices for non-residential customers serve as a plausible counterfactual for residential prices in absence of the regulatory policies, which I verify graphically in Figure 2.3. Additionally, I estimate equation (2.4) using state average bundled and delivery-only service prices in place of energy-only prices. Because bundled and delivery-only service prices are regulated and are not affected by the pro-competitive policies, the coefficients for purchase of receivables rules and aggregation programs should be zero; however, utilities may have adopted market-based prices that shrank or widened the gap between residential and commercial customer prices. The results from these auxiliary regressions are discussed with the main results in Section 2.5.

2.5 Results

2.5.1 Effects of Policies on Competitive Retail Market Size

The results from estimating Equation (2.1) are presented in Table 2.3. Of the three policies tested, the adoption of market-based rates appears to have had the largest effect on residential customer switching, followed by the adoption of purchase of receivables rules. As expected, the estimated coefficients for *Market Rates* are positive and statistically significant for all customer classes, and the effect size is larger for industrial customers than for residential customers. Also, the coefficient estimates for *POR* are positive and statistically significant for residential and commercial customers, but not industrial users. The pattern is the same for *Aggregation*, however the coefficients are not statistically significant.

Interestingly, the coefficient estimates in Table 2.3 are similar for residential and commercial customers. This makes sense because the outcome variable is the proportion of

utility customers who choose to purchase competitive retail electric service, and the majority of commercial customers are small electric users, who are very similar to residential customers (from a retailer/utility and policymaker perspective). I therefore estimate Equation (2.1) using the utility's sales-weighted switching share as the dependent variable, which weights customers according to their electric consumption. The estimates are reported in Table 2.4, and show that the results change significantly for commercial customers. Specifically, the coefficient estimates for *Market Rates* are much larger (closer to the estimates for industrial customers), and the estimates for *Aggregation* are much smaller. The results for residential and industrial customers are essentially unchanged, because these customers within these groups are more homogeneous in terms of their electric usage.

Table 2.5 presents the estimates for Equation (2.2). Again, the variable is the customer switching share, as in Table 2.3, so the results for residential and commercial customers are very similar. As predicted, purchase of receivables and opt-out aggregation programs appear to have had a delayed effect on customer switching for residential and commercial customers, while neither policy had a statistically significant impact on industrial customer switching. The difference between the estimates for *Market Rates (Years 1-2)* and *Market Rates (Later Years)* is not statistically significant for any customer classes, and so there is no evidence that the effect of adopting market-based rates increased over time.

2.5.2 Effects of Policies on Retailer Entry and Exit

Table 2.6 presents the estimates for Equation (2.3) using the data on the number of competitive retailers operating in each state. The most interesting result is that the coefficients for *Aggregation* are also positive and statistically significant for residential and commercial customers. This finding differs from the predicted effect of aggregation programs on retailer entry, which was based on the idea that aggregation programs favor incumbent

retailers and therefore deter entry. To the contrary, the estimates suggest that aggregation programs encouraged entry.

As expected, the coefficient estimates for *Market Rates* are not statistically significant, which suggests that the adoption of market-based default service prices did not have a significant impact on retailer entry. This is consistent with the idea that the introduction of market-based default service increased consumer demand for competitive retail electric service, but did not lower the cost of serving small electric customers, and therefore did not encourage additional retailer entry. Also as expected, the coefficients for *POR* are positive and statistically significant for residential and commercial customers, which suggests that these policies had the desired effect of increasing competitive retailer entry. None of the policies appears to have had a significant effect on the number of retailers serving industrial customers.

2.5.3 Effects of Policies on Competitive Retailer Prices

The coefficient estimates for Equation (2.4) are reported in Table 2.7. Columns (1)-(3) use the difference between residential and commercial customer prices as the dependent variable, and Columns (4)-(6) use the difference between residential and industrial customer prices. The estimation sample is smaller than the sample used in the previous section because there are some years for which no residential energy-only prices are reported. As before, standard errors are clustered at the state level.

The only statistically significant estimate in Table 2.7 is the coefficient on *Aggregation* in Column (4), which is the specification where the dependent variable is the difference between the average energy-only prices for residential and industrial customers. If industrial prices are a valid counterfactual for residential prices, then this estimate suggests that the authorization of opt-out aggregation programs reduced competitive retailer prices by

about 1.3 cents per kWh for residential customers (the sample average energy-only price for residential customers is 8.85 cents per kWh, for comparison).

The fact that none of the other estimates in Table 2.7 are statistically significant may be seen as evidence that energy-only prices for industrial customers are a valid counterfactual for energy-only residential customer prices. As previously discussed, the estimated effects of the regulatory policies on customer switching shares and competitive retailer entry for residential and commercial customers were very similar. If these policies also had a similar effect on competitive retailer prices for residential and commercial customers, then the coefficients in Column (1) should be approximately zero (which they are). Furthermore, none of the regulatory policies appear to have had an effect on the difference between residential and non-residential customers' bundled and delivery-only service prices, which is to be expected given that these prices are regulated and thus should be unaffected by changes in policies affecting competitive retailers.

2.6 Discussion

The results from this paper demonstrate the importance of pro-competitive regulatory policies for establishing functional competitive retail electricity markets in the United States. Of the policies tested, the adoption of market-based default service prices had the largest effect on residential customer participation in competitive retail electric markets. Purchase of receivables rules also had a small positive effect on residential customer participation, however they had a more significant impact on increasing the number of competitive retailers serving residential customers, as regulators intended. Aggregation programs, in addition to increasing the share of residential customers served by competitive retailers and the number of retailers serving them, appear to have led to lower prices for competitive retail electric service. These patterns suggest that, of the three policies, aggregation programs are the most likely to have increased consumer welfare.

The intuition for this finding is that opt-out aggregation programs increase residential customer demand for competitive retail electric service, while encouraging competitive retailers to offer lower prices. Aggregators use competitive procurement processes that allow them to negotiate better prices for their participating customers, compared to the prices that customers would be offered if they were to shop for competitive retail electric service contracts on their own. Additionally, aggregating residential customers' electric usage results in an electric load profile that is more predictable, which makes residential customers cheaper to serve. Thus, aggregation programs benefit consumers because they lower the cost of providing retail electric service to residential customers, while the competitive procurement processes used by aggregators ensure that these cost savings are passed through to consumers via lower retail electricity prices.

In general, it is not possible to conclude whether residential customers have benefited from retail electric market competition based on competitive retailer prices alone. This is because competitive retailer customers may still pay higher prices than they would pay for their utility's default service option.[13] Nor is it possible to evaluate whether retail electricity market deregulation benefits consumers using data on state-level average prices, as was used in this paper and in previous literature (Upton et al., 2021; MacKay and Mercadal, 2021; Su, 2015). This is because customers in deregulated retail electricity markets pay different prices depending on the terms of their individual retail electric service contracts, such that customers with the highest search costs pay the highest prices (Hortaçsu et al., 2017; Byrne and Martin, 2021; Giulietti et al., 2010). Furthermore, customers may benefit from retail electric service contracts that include non-price attributes, such as renewable electricity, that are not reflected in analyses based on retail electricity prices alone. Thus, a comprehensive evaluation of whether U.S. consumers have benefited from retail electric

[13]For example, the Connecticut Office of the Consumer Counsel's analysis of utility bills found that residential customers paid more than $240 million more to competitive retailers than if they had remained on their utility's default service option between 2015-2019.

market competition requires data on customers' individual retail electric service contract terms, which are usually confidential.

Finally, it is worth highlighting that while purchase of receivables programs increased residential customer participation in competitive retail electric markets and encouraged entry by competitive retailers, they may have done so at a significant cost to customers. This is because purchase of receivables programs encourage competitive retailers to sell more expensive contracts to customers, without considering the risk that customers may not be able to pay their bills.[14] Given this moral hazard problem, it is perhaps unsurprising that some retailers have resorted to deceptive marketing practices to enroll customers in markets with such rules.[15]

2.7 Conclusion

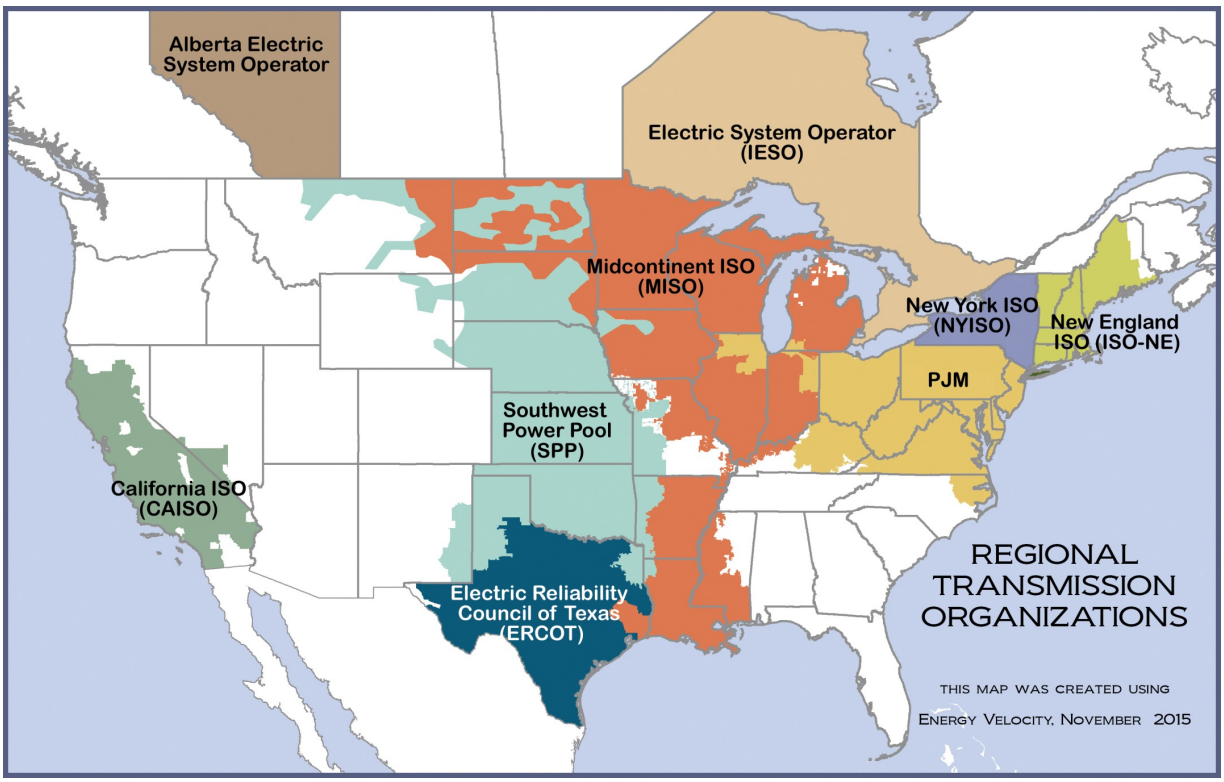
This paper showed that regulatory policies were crucial for driving the development of competitive residential retail electricity markets in the United States. Understanding the role of regulatory policy in these markets is timely, given recent, renewed policy interest in using retail electric competition to facilitate the integration of distributed energy resources such as roof-top solar and electric vehicles (Burger et al., 2019). At the same time, Massachusetts is considering whether to re-regulate their residential retail electricity markets, in light of recent findings that the state's customers pay significantly more for competitive retail electric service than they would have paid for their utility's default service option (Baldwin, 2018). Both sides of these policy debates would benefit from additional research into these markets. This paper provides an important first step.

[14]Purchase of receivables programs also introduce regulatory certainty into the competitive retail electric market, which is counter to traditional economic grounds for deregulation.

[15]See, for example, the investigations listed in the Connecticut Office of the Consumer Counsel's Fact Sheet on *In Support of Increased Consumer Protection for the Residential Retail Choice Electric Supply Market*, published on January 30, 2020.

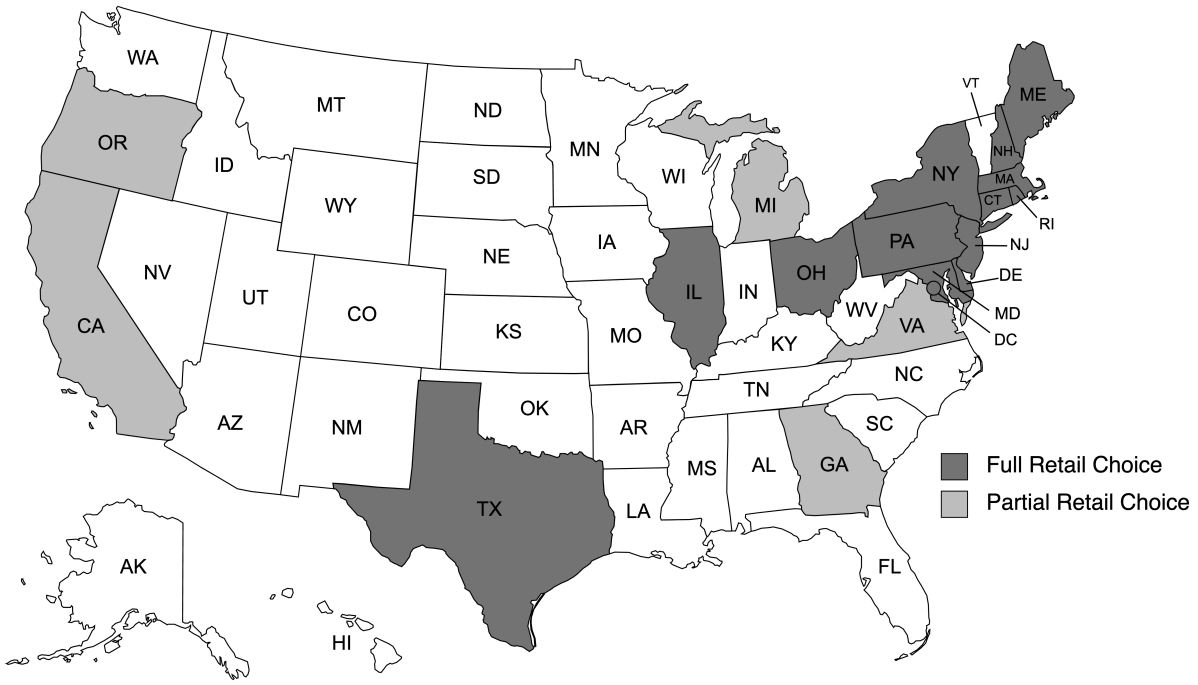
2.8 Figures and Tables

Figure 2.1: Restructured Wholesale Electricity Markets in the United States



Source: Federal Energy Regulatory Commission

Figure 2.2: Restructured Retail Electricity Markets in the United States



Notes: Full Retail Choice states allow retail electric competition for residential customers. Partial Retail Choice states allow retail electric competition for some non-residential customers, but not for residential customers. This figure was created using mapchart.net, combined with information obtained from the American Coalition of Competitive Energy Suppliers website.

Figure 2.3: Competitive Retailer Prices by Customer Class

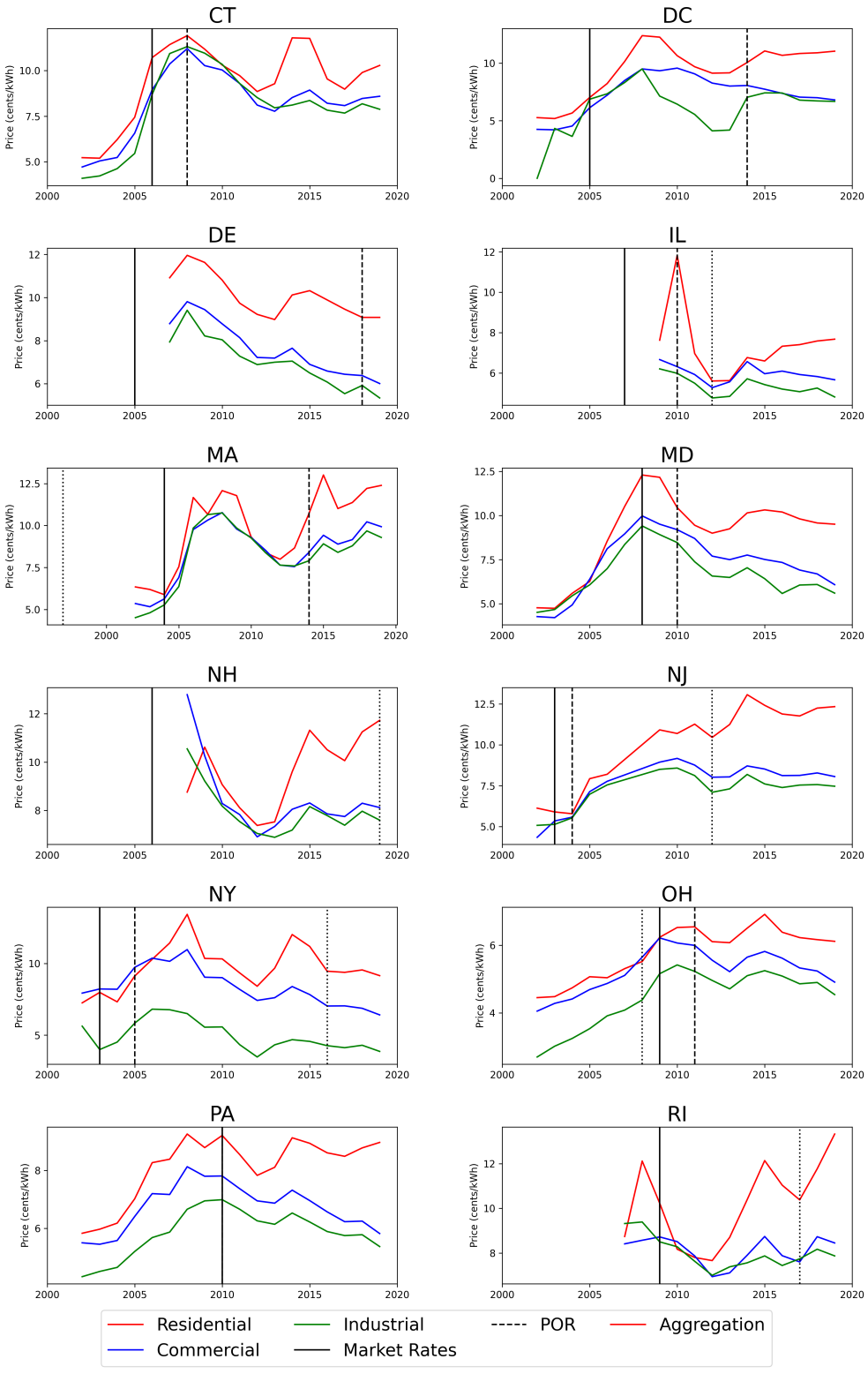


Table 2.1: State Retail Electric Restructuring Policy Dates

State	Legislation	Market Access	End Transition	POR	Aggregation
CT	1998	2000	2006	2008	N/A
DC	1999	2001	2005	2014	N/A
DE	1999	2000	2005	2018	N/A
IL	1997	2002	2007	2011	2012
MA	1997	1998	2004	2014	1997
MD	1999	2000	2008	2010	N/A
ME	1997	2000	2005	N/A	N/A
NH	1996	2001	2006	N/A	2019
NJ	1997	1999	2003	2004	2012
NY	1997	2001	2003	2005	2016
OH	1999	2001	2009	2011	2008
PA	1996	2001	2010	2010	N/A
RI	1996	1998	2009	N/A	2017

Policy	Definition	Sources
Market Access	Date when residential customers legally allowed to choose a competitive supplier	Energy Information Administration, state restructuring legislation
End Transition	Date when major utilities' transition rates expire and small electric customers exposed to market based default prices	Upton et al. (2021), Su (2015), industry reports
Purchase of Receivables (POR)	First date that major utility implemented a POR program	news articles, press releases, state public utility commissions
Aggregation	Date when state adopted legislation to strengthen its opt-out municipal aggregation provisions, or date of first aggregation program (see Table 2.8 for details)	O'Shaughnessy et al. (2019), Hsu (2022)

Table 2.2: Summary Statistics for Key Variables by Customer Class

	Residential Customers					Non-Residential Customers				
	Mean	SD	Min	Median	Max	Mean	SD	Min	Median	Max
2002										
Customer Switching Share	0.06	0.13	0.00	0.00	0.56	0.06	0.10	0.00	0.01	0.48
Sales Switching Share	0.06	0.12	0.00	0.00	0.55	0.12	0.14	0.00	0.05	0.58
Bundled Price (¢/kWh)	9.89	2.08	6.61	9.87	16.86	12.57	3.74	6.14	11.53	22.49
Delivery Price (¢/kWh)	5.25	2.21	0.00	6.11	7.87	3.51	1.09	0.88	3.47	5.20
Log(Customers)	12.66	1.59	8.16	13.01	15.00	10.65	1.49	6.78	10.97	13.02
Log(Sales)	14.86	1.57	10.20	15.23	17.13	14.89	1.69	10.43	15.24	17.27
2011										
Customer Switching Share	0.17	0.20	0.00	0.09	0.75	0.30	0.18	0.00	0.25	0.74
Sales Switching Share	0.19	0.21	0.00	0.11	0.76	0.63	0.18	0.00	0.66	0.87
Bundled Price (¢/kWh)	14.23	3.12	9.83	13.54	25.50	15.22	3.72	9.98	14.40	28.01
Delivery Price (¢/kWh)	5.96	2.12	3.03	4.79	9.93	3.84	1.85	1.21	3.08	6.85
Log(Customers)	12.70	1.57	8.21	13.09	15.05	10.74	1.47	6.90	11.05	13.12
Log(Sales)	14.95	1.56	10.30	15.32	17.16	14.97	1.70	10.45	15.44	17.50
2019										
Customer Switching Share	0.28	0.21	0.00	0.23	0.90	0.43	0.18	0.07	0.40	0.88
Sales Switching Share	0.29	0.21	0.00	0.24	0.90	0.73	0.15	0.22	0.74	0.96
Bundled Price (¢/kWh)	15.57	4.76	9.58	13.39	25.37	14.18	3.83	8.86	12.90	24.70
Delivery Price (¢/kWh)	8.00	2.81	4.64	6.50	12.88	4.75	2.02	2.25	4.39	8.20
Log(Customers)	12.74	1.57	8.25	13.11	15.11	10.79	1.47	6.93	11.04	13.23
Log(Sales)	14.93	1.54	10.30	15.21	17.10	14.94	1.71	10.40	15.22	17.47
Total Observations			1720					1720		
Utilities			43					43		
Years			20					20		

Notes: Sample includes investor-owned utilities in CT, DE, IL, MA, MD, NH, NH, NJ, NY, OH, PA, RI and the District of Columbia. Non-Residential customers include commercial and industrial customers. *Switching Share* is the fraction of utility customers or sales that purchase competitive retail electric service. *Bundled Price* is a utility's average monthly price for bundled electric service (across all months of the calendar year). *Delivery Price* is the utility's average price for delivery-only service, where outliers are replaced with state average delivery-only prices. Data come from the Energy Information Administration Form 861.

Table 2.3: Estimated Effects of Policies on Customer Switching Share

	Residential		Commercial		Industrial	
	(1)	(2)	(3)	(4)	(5)	(6)
Market Rates	12.24***	11.01**	14.41***	11.63***	25.32***	24.28***
	(3.459)	(3.839)	(3.030)	(3.476)	(5.244)	(5.316)
POR	8.925**	8.316**	8.125**	7.771**	0.341	0.229
	(3.726)	(3.153)	(3.301)	(3.343)	(5.339)	(5.134)
Aggregation	6.866	7.675	5.539	4.952	-2.494	-2.376
	(5.876)	(5.552)	(4.555)	(3.677)	(4.194)	(4.632)
Bundled Price		1.228		1.383**		0.117
		(0.882)		(0.533)		(0.337)
Delivery-Only Price		-0.272		-0.462*		-0.788
		(0.226)		(0.215)		(0.886)
Observations	860	860	860	816	816	814
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Utility FE	Yes	Yes	Yes	Yes	Yes	Yes
F-Stat	11.75	8.243	9.299	99.51	13.41	11.44
R-sq.	0.739	0.747	0.813	0.821	0.823	0.825

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Notes: The dependent variable is the utility's customer switching share (e.g., the share of utility customers who purchase competitive retail electric service). Market Rates, POR, and Aggregation are dummy variables equal to one for years after a utility began using market-based default service rates, implemented a purchase of receivables rule, or served opt-out municipal aggregation program participants. The sample includes investor-owned utilities in states that enabled retail choice for all customer classes. The data span 2001-2019. Standard errors are clustered at the state level.

Table 2.4: Estimated Effects of Policies on Sales Switching Share

	Residential		Commercial		Industrial	
	(1)	(2)	(3)	(4)	(5)	(6)
Market Rates	11.02**	9.809**	24.24***	19.71***	34.22***	29.38***
	(3.582)	(3.956)	(4.318)	(4.428)	(6.282)	(6.664)
POR	10.18**	9.599**	7.589*	6.307*	0.0410	-0.447
	(3.937)	(3.423)	(3.905)	(3.331)	(6.224)	(6.004)
Aggregation	6.855	7.627	0.672	2.284	-6.784	-4.589
	(5.836)	(5.564)	(2.909)	(2.656)	(5.490)	(5.628)
Bundled Price		1.195		2.967***		1.701***
		(0.887)		(0.540)		(0.480)
Delivery-Only Price		-0.329		-0.902**		-2.200*
		(0.233)		(0.308)		(1.060)
Observations	860	860	860	816	816	814
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Utility FE	Yes	Yes	Yes	Yes	Yes	Yes
F-Stat	10.12	6.941	11.43	16.88	13.08	18.17
R-sq.	0.747	0.754	0.875	0.887	0.796	0.809

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Notes: The dependent variable is the utility's sales switching share, which measures the share of customers who purchase competitive retail electric service, but weights customers according to their electric consumption. Market Rates, POR, and Aggregation are dummy variables equal to one for years after a utility began using market-based default service rates, implemented a purchase of receivables rule, or served opt-out municipal aggregation program participants. The sample includes investor-owned utilities in states that enabled retail choice for all customer classes. The data span 2001-2019. Standard errors are clustered at the state level.

Table 2.5: Results for Delayed Effects of Policies on Customer Switching Share

	Residential		Commercial		Industrial	
	(1)	(2)	(3)	(4)	(5)	(6)
Market Rates (Years 1-2)	10.91***	9.254***	12.64***	9.843***	23.14***	22.56***
	(2.498)	(2.699)	(3.013)	(2.101)	(6.625)	(6.404)
Market Rates (Later Years)	16.37***	14.68***	18.46***	15.20***	27.21***	25.71***
	(3.584)	(4.132)	(3.304)	(3.846)	(5.070)	(5.112)
POR (Years 1-2)	0.195	-0.334	1.848	1.607	0.0980	0.150
	(0.905)	(0.806)	(1.186)	(0.905)	(3.925)	(4.020)
POR (Later Years)	11.09**	10.82**	9.071**	9.267**	-0.173	0.364
	(4.454)	(3.978)	(3.959)	(4.108)	(4.045)	(3.901)
Aggregation (Years 1-2)	1.747	1.988	1.228	0.818	-6.718	-6.562
	(4.430)	(3.839)	(4.981)	(3.724)	(4.298)	(4.617)
Aggregation (Later Years)	10.80	12.39*	8.864*	8.785*	0.816	0.577
	(6.703)	(6.282)	(4.790)	(4.371)	(6.004)	(6.544)
Bundled Price		1.484		1.560**		0.0507
		(0.912)		(0.621)		(0.311)
Delivery-Only Price		-0.279		-0.494**		-0.610
		(0.258)		(0.203)		(0.879)
Observations	860	860	860	816	836	820
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Utility FE	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses

Notes: The dependent variable is the share of utility customers served by competitive electric retailers. Market Rates, POR, and Aggregation are dummy variables equal to one for years after a utility began using market-based default service rates, implemented a purchase of receivables program, or served opt-out municipal aggregation program participants. The sample includes investor-owned utilities in states that enabled retail choice for all customer classes. The data span 2001-2019. Standard errors are clustered at the state level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 2.6: OLS Results for Effect of Policies on Retailer Entry and Exit

	Residential		Commercial		Industrial	
	(1)	(2)	(3)	(4)	(5)	(6)
Market Rates (t-1)	-0.803	-0.759	-1.167	-1.490	-0.706*	-0.691*
	(1.086)	(0.908)	(1.043)	(0.917)	(0.381)	(0.376)
POR (t-1)	2.604***	2.222***	2.108***	1.801***	0.259	0.183
	(0.805)	(0.686)	(0.655)	(0.524)	(0.272)	(0.215)
Aggregation (t-1)	2.157**	2.103***	2.368***	1.910**	0.711*	0.609*
	(0.755)	(0.661)	(0.739)	(0.731)	(0.342)	(0.305)
Bundled Price (t-1)		0.218		0.133		0.0198
		(0.227)		(0.264)		(0.0304)
Delivery Price (t-1)		-0.345*		-0.998**		-0.209
		(0.191)		(0.443)		(0.151)
Observations	228	228	228	228	228	228
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes
F-Stat	5.578	9.388	20.96	28.00	7.257	7.399
R-sq.	0.697	0.712	0.548	0.575	0.214	0.225

Standard errors in parentheses

Notes: The dependent variable is the number of competitive retailers operating in a state retail electricity market. Market Rates, POR, and Aggregation are dummy variables equal to one for years after a utility began using market-based default service rates, implemented a purchase of receivables program, or served opt-out municipal aggregation program participants. The sample includes all states that enabled retail choice for all customer classes. The data span 2001-2019. Standard errors are clustered at the state level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 2.7: OLS Results for Effect of Policies on Retailer Prices

	Residential – Commercial			Residential – Industrial		
	(1) Energy	(2) Bundled	(3) Delivery	(4) Energy	(5) Bundled	(6) Delivery
Market Rates	-0.0627	-0.223	-0.256	-0.501	-0.0567	-0.713
	(0.344)	(0.642)	(0.370)	(0.348)	(0.652)	(0.671)
POR	-0.317	0.445	0.179	0.0354	-0.0234	0.0831
	(0.631)	(0.599)	(0.404)	(0.733)	(1.072)	(0.619)
Aggregation	-0.890	-0.488	0.160	-1.325***	-0.170	0.472
	(0.478)	(0.274)	(0.360)	(0.347)	(0.562)	(0.572)
Observations	125	125	125	125	125	125
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes
F-Stat	3.807	1.150	1.144	16.85	0.0568	1.359
R-sq.	0.603	0.718	0.768	0.715	0.815	0.790

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Notes: The dependent variables in Columns (1)-(3) are the difference between the residential and commercial state average prices for energy-only, bundled, and delivery-only services, respectively. The dependent variables in Columns (4)-(6) are the difference between state average residential and industrial customer prices. Market Rates, POR, and Aggregation are dummy variables equal to one for years after a utility began using market-based default service rates, implemented a purchase of receivables rule, or served opt-out municipal aggregation program participants. The sample includes states that enabled retail choice for all customer classes. The data span 2001-2019, and exclude states and years where no average prices are reported. Standard errors are clustered at the state level.

Table 2.8: Notes on Opt-out Aggregation Program Authorization Dates Used

State	Notes
Illinois	The Illinois Power Agency Act was amended in 2009 to allow for municipal aggregation, and went into effect on January 1, 2010. Townships gained this ability in 2012, which resulted in a surge of aggregation communities, and hence I use this date (Deryugina et al., 2020).
Massachusetts	Aggregation enabled in 1997.
New Hampshire	Enabling legislation became effective in October 2019.
New Jersey	The Government Energy Aggregation Act of 2003 authorized municipalities and/or counties of New Jersey to establish Government Energy Aggregation programs after passing an ordinance or a resolution. Due to procedural barriers, however, the first aggregation programs did not start until 2012, so I use this date (Chretien et al., 2020).
New York	The New York Public Service Commission issued an order on April 21, 2016, which enabled statewide aggregation programs.
Ohio	Initial legislation enacted in 1999. Senate Bill 221 helped catalyze aggregation in 2008 by requiring utilities to support large-scale programs (Chretien et al., 2020), so I use this date.
Pennsylvania	Opt-out aggregation programs must be approved by the state public utility, and are only approved in "unique circumstances where it is clearly in the public interest to do so." The utility-level data set accounts for one aggregation program that was approved in the Pike County Light & Power service territory in 2006.
Rhode Island	Enabling legislation was passed in 2017, but aggregation programs only serve governmental customers (e.g., municipal buildings).

Bibliography

Abito, Jose Miguel and Yuval Salant, "The Effect of Product Misperception on Economic Outcomes: Evidence from the Extended Warranty Market," *Review of Economic Studies*, 2019, *86* (6), 2285–2318.

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Perez, "Automated Linking of Historical Data," *NBER Working Paper*, 2019.

_ , Leah Platt Boustan, and Katherine Eriksson, "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration," *American Economic Review*, 5 2012, *102* (5), 1832–1856.

_ , _ , and _ , "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration," *Journal of Political Economy*, 2014, *122* (3), 467–506.

_ , _ , and _ , "To the New World and Back Again: Return Migrants in the Age of Mass Migration," *ILR Review*, 2019, *72* (2), 300–322.

_ , Roy Mill, and Santiago Perez, "Linking Individuals Across Historical Sources: a Fully Automated Approach," Technical Report 24324, National Bureau of Economic Research 2 2018.

Aizer, Anna, Shari Eli, Joseph Ferrie, and Adriana Lleras-Muney, "The Long-Run Impact of Cash Transfers to Poor Families," *American Economic Review*, 4 2016, *106* (4), 935–971.

Anderson, Rachel, Bo Honoré, and Adriana Lleras-Muney, "Estimation and inference using imperfectly matched data," *Working paper*, 8 2022.

Auffhammer, Maximilian and Edward Rubin, "Natural Gas Price Elasticities and Optimal Cost Recovery Under Consumer Heterogeneity: Evidence from 300 million natural gas bills," *National Bureau of Economic Research*, 2018.

Averch, Harvey and Leland L Johnson, "American Economic Association Behavior of the Firm Under Regulatory Constraint," *Source: The American Economic Review*, 1962, *52* (5), 1052–1069.

Bailey, Martha, Connor Cole, Morgan Henderson, and Catherine Massey, "How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data," Technical Report 24019, National Bureau of Economic Research 11 2017.

Baldwin, Susan M, "Are Consumers Benefiting from Competition? An Analysis of the Individual Residential Electric Supply Market in Massachusetts," 2018.

Brown, David P. and Andrew Eckert, "The effect of default rates on retail competition and pricing decisions of competitive retailers: The case of Alberta," *Energy Policy*, 2018, *118* (December 2017), 298–311.

Burger, Scott P., Jesse D. Jenkins, Carlos Batlle, and Ignacio J. Pérez-Arriaga, "Restructuring revisited part 1: Competition in electricity distribution systems," *Energy Journal*, 2019, *40* (3), 31–54.

Bushnell, James, Erin T Mansur, and Kevin Novan, "Review of the Economics Literature on US Electricity Restructuring," Technical Report 2017.

Byrne, David P. and Leslie A. Martin, "Consumer search and income inequality," *International Journal of Industrial Organization*, 2021.

_ , _ , and Jia Sheen Nah, "Price Discrimination by Negotiation: A Field Experiment in Retail Electricity," *SSRN Electronic Journal*, 3 2022.

Chamberlain, Gary, "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 3 1987, *34* (3), 305–334.

Chetty, Raj, John N Friedman, and Jonah E Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," Technical Report 19424, National Bureau of Economic Research 9 2013.

_ , _ , Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan, "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star," *The Quarterly Journal of Economics*, 2011, *126* (4), 1593–1660.

Chretien, Larry, Daria Mark, Eugenia Gibbons, Yaima Braga, and Kate Marcellino, "Green Municipal Aggregation in Massachusetts," 2020, (February), 46.

Collins, William J and Marianne H Wanamaker, "Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data," Technical Report 19124, National Bureau of Economic Research 6 2013.

Dempster, A P, N M Laird, and D B Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1977, *39* (1), 1–38.

Deryugina, Tatyana, Alexander MacKay, and Julian Reif, "The long-run dynamics of electricity demand: Evidence from municipal aggregation," *American Economic Journal: Applied Economics*, 2020, *12* (1), 86–114.

Distributed Energy Financial Group LLC, "Annual Baseline Assessment of Choice in Canada and the United States," Technical Report July, Distributed Energy Financial Group LLC 2015.

Dobbie, Will and Roland G Fryer, "Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone," *American Economic Journal: Applied Economics*, 2011, *3* (3), 158–187.

Dormady, Noah, Matthew Hoyt, Alfredo Roa-Henriquez, and William Welch, "Who pays for retail electric deregulation? Evidence of cross-subsidization from complete bill data," *Energy Journal*, 2019, *40* (2), 161–194.

EIA, "Status of State Electric Industry Restructuring Activity as of February 2003," 2003, (February).

_ , "Annual Electric Power Industry Report, Form EIA-861 Detailed Data Files," 2014.

Enamorado, Ted, Benjamin Fifield, and Kosuke Imai, "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records," *American Political Science Review*, 2019, *113* (2), 353?371.

Fellegi, I P and A B Sunter, "A Theory for Record Linkage," *Journal of the American Statistical Association*, 1969, *64*, 1183–1210.

Ferrie, Joseph and Karen Rolf, "Socioeconomic status in childhood and health after age 70: A new longitudinal analysis for the U.S., 1895G2005," *Explorations in Economic History*, 2011, *48* (4), 445–460.

Ferrie, Joseph P, "A New Sample of Males Linked from the Public Use Microdata Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules," *Historical Methods*, 1996, *29* (4), 141–156.

Giulietti, Monica, Jesus Otero, and Michael Waterson, "Pricing behaviour under competition in the UK electricity supply industry," *New Series*, 2010, *62* (3), 478–503.

Hansen, Lars Peter, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 1982, *50* (4), 1029–1054.

Hortaçsu, Ali, Seyed Ali Madanizadeh, and Steven L. Puller, "Power to choose? An analysis of consumer inertia in the residential electricity market," *American Economic Journal: Economic Policy*, 2017, *9* (4), 192–226.

Hsu, David, "choice aggregation and its spread to other states Version spread to other states," 2022.

Jaro, Matthew A, "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, 1989, *84* (406), 414–420.

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz, "Experimental Analysis of Neighborhood Effects," *Econometrica*, 2007, *75* (1), 83–119.

_ , Jens Ludwig, and Lawrence F Katz, "Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment," *The Quarterly Journal of Economics*, 2005, *120* (1), 87–130.

Lahiri, P and Michael D Larsen, "Regression Analysis with Linked Data," *Journal of the American Statistical Association*, 2005, *100* (469), 222–230.

Larsen, Michael D and Donald B Rubin, "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association*, 2001, *96* (453), 32?41.

Littlechild, Stephen, "The regulation of retail competition in US residential electricity markets," Technical Report 2018.

MacKay, Alexander and Ignacia Mercadal, "Deregulation, Market Power, and Prices: Evidence from the Electricity Sector," *SSRN Electronic Journal*, 2021.

Neter, John, E Scott Maynes, and R Ramanathan, "The Effect of Mismatching on the Measurement of Response Errors," *Journal of the American Statistical Association*, 1965, *60* (312), 1005–1027.

Newey, Whitney K., "Consistency of two-step sample selection estimators despite misspecification of distribution," *Economics Letters*, 5 1999, *63* (2), 129–132.

Newey, Whitney K and Daniel McFadden, "Large Sample Estimation and Hypothesis Testing," in Robert F Engle and Daniel L McFadden, eds., *Handbook of Econometrics*, number 4. In 'Handbooks in Economics,.', Amsterdam, London and New York: Elsevier, {N}orth-{H}olland, 1994, pp. 2111–2245.

Office of Consumer Counsel, "OCC FACT SHEET: In Support of Increased Consumer Protection for the Residential Retail Choice Electric Supply Market," Technical Report, State of Connecticut Office of Consumer Counsel 2020.

O'Shaughnessy, Eric, Jenny Heeter, Julien Gattaciecca, Jenny Sauer, Kelly Trumbull, and Emily Chen, "Empowered communities: The rise of community choice aggregation in the United States," *Energy Policy*, 9 2019, *132*, 1110–1119.

SAS, "SAS 9.4 Functions and CALL Routines: Reference, Fifth Edition," Technical Report, SAS Institute Inc. 2019.

Scheuren, Fritz and William Winkler, "Regression analysis of data files that are computer matched," *Survey Methodology*, 1993, *19*.

Singer, Stephen, "Why is Connecticut No. 1 in the cost of electricity among the 48 lower states? Here are 7 reasons," 4 2021.

Su, Xuejuan, "Have customers benefited from electricity retail competition?," *Journal of Regulatory Economics*, 2015, *47* (2), 146–182.

Swadley, Adam and Mine Yücel, "Did residential electricity rates fall after retail competition? A dynamic panel analysis," *Energy Policy*, 12 2011, *39* (12), 7702–7711.

Tsai, Chen-Hao and Yi-Lin Tsai, "Competitive Retail Electricity Market Under Continuous Price Regulation," *SSRN Electronic Journal*, 2017.

Upton, Gregory, Kenneth M. Rosen, and Brittany Tarufelli, "Electricity Market Restructuring and Retail Rates," *SSRN Electronic Journal*, 3 2021.

U.S. Energy Information Administration, "Electric Power Annual 2020," Technical Report 2020.

Wang, Xueting, "Long term contracts in retail electricity," 2021.

Winkler, William E, "Overview of record linkage and current research directions," Technical Report, U.S. Census Bureau 2006.

Chapter 3

Estimation and inference using imperfectly matched data[1]

3.1 Introduction

The recent availability of large administrative data sets has increased the use of matched data in economic applications. In these studies, the standard approach to analyzing matched data is to exclude observations that do not have unique matches, and assume that the remaining observations have been linked without error. Inference and estimation then proceed by treating the merged data set as if it came from a single source.

The primary disadvantage of this approach is that matched data sets may contain three types of errors that cause standard estimators to be biased. First, some observations may not be matched to an outcome, which results in missing data. Second, some observations may be linked to multiple matches, corresponding with multiple outcomes, when it is not possible to identify unique matches. Finally, there might be measurement error in

[1]This chapter was co-authored with Bo Honoré and Adriana Lleras-Muney. The material is based upon work supported by the National Institutes of Health Grant No. R01 HD077227-01.

matching: even when an observation is linked to a unique match, that match might not correspond to the correct match.

In this paper, we develop a theoretical framework that formalizes these ideas, and propose methods for analyzing matched data sets that allow for the use of multiple matches per observation. We show that it is straightforward to modify standard estimators to accommodate data where some observations are linked to multiple possible values of the dependent variable. Furthermore, we demonstrate that it is relatively simple to compute the efficient version of our estimator by weighting observations according to the number of corresponding matches. Results from a small Monte Carlo study indicate that this efficient version of our estimator performs as well or better than standard estimators computed using only the subset of observations that are linked to unique matches under the data generating processes that we test.

Our methods are likely to be useful for many applications, including settings where researchers could increase the size of the matched data set by allowing for multiple matches per observation. Such challenges frequently arise in studies by economic historians and historical demographers that analyze merged data. For example, Abramitzky et al. (2012) link Norwegian-born men from the Norwegian census of 1865 to the 1900 U.S. Census of Population and the 1900 Norwegian census, finding matches for only 26% of observations from the base year. Ferrie and Rolf (2011) link males born between 1895 and 1900 from the 5% Integrated Public Use Microdata Series (IPUMS) sample of the 1900 U.S. Census of Population to the Social Security Death Master File, successfully linking 29% of observations. Collins and Wanamaker (2013) are able to link 21% of black, Southern-born males from the 1% IPUMS sample of the 1910 U.S. Census of Population to the 1930 U.S. Census of Population.

There are also several prominent examples of program evaluations that use matched administrative data sets, and could potentially benefit from our methods. For example,

the Moving to Opportunity experiment, which offered vouchers to poor families to move to low-poverty areas, was evaluated by matching data on participants to multiple administrative data sets, including state unemployment insurance and public assistance program data, juvenile and criminal justice administrative data, National Student Clearinghouse data on college attendance, and school data, among others (Kling et al., 2007, 2005). Chetty et al. (2013) look at the long-run effects of smaller classes and better teachers by matching experimental data from the Tennessee Project Student/Teacher Achievement Ratio to administrative tax records from when the participants were adults. Chetty et al. (2011) match 2.5 million New York City public school records to Internal Revenue Service data to study the long-run impacts of teacher quality on students' earnings as adults. Dobbie and Fryer (2011) evaluate the effect of the Harlem Children's Zone by matching data from the program to National Student Clearinghouse college enrollment data.

Finally, there are also many publicly available and commonly used data sets that are constructed from matching administrative data sets. Examples include the Linked Birth and Infant Death data (which matches birth certificates and death certificates from population databases), the IPUMS Linked Representative Samples for 1850-1930 (which link individuals across censuses to create a longitudinal data set), the National Longitudinal Mortality Survey (which matches Current Population Survey data to death certificates from the National Death Index, the National Health Interview Survey Linked Mortality File (which is the National Health Interview Survey linked to the National Death Index), the General Social Survey-National Death Index, and the National Health and Nutrition Examination Survey linked mortality files. Even in these high quality data sets, which contain individuals' social security numbers, there are a non-trivial number of multiple matches and linking error.

The rest of this paper proceeds as follows. The next section introduces the data generating process that results in a matched data set for which our methods can be applied. In Section 3.3, we derive the estimator that allows for multiple matches in linear instrumental variables models, and describe in detail how the estimator should be implemented in practice. In Section 4.4, we present the results from a small Monte Carlo study that evaluates the performance of our proposed methods. Finally, Section 3.5 discusses how our methods can be generalized to other econometric models, and Section 4.6 concludes.

3.2 Data Generating Process

We start by describing the data generating process that results in a matched data set for which our methods can be applied. We assume that the researcher is interested in analyzing a data set which is the result of linking two distinct data sets. The first data set contains the covariates (X, Z) for a random sample of individuals (we refer to this as the "x-data set"). The second data set contains the outcomes Y from a larger supersample (the "y-data set"). The data are linked in a way such that each observation of (X, Z) is associated with a unique identifier W and is matched to L values of Y that are associated with the same identifier W. One of these L values of Y is the true outcome for the same individual as (X, Z), while the other $L - 1$ values are "false" matches.

One example of such a data set comes from Aizer et al. (2016), who study the effect of providing cash transfers to single mothers on the life expectancy of their children. The covariates (X, Z) are obtained from mothers' welfare program applications, and include a binary variable equal to 1 if an individual's mother received a cash transfer, and other demographic characteristics of the individual. The outcomes Y are obtained from a database of death records. The values of (X, Z) are linked to the outcomes Y using identifiers W that are recorded in both data files. They include an individual's first name, middle initial, and

birth date, such that individuals with common names cannot be linked to a unique death record. As a result, the merged data set links some applicants to multiple death records.

More formally, the data generating process can be described as follows. Let F be the distribution of the vector (X, Z, Y, W, L), where W is continuously distributed and L is independent of (X, Z, Y) conditional on W. L is an integer and takes on values in $\{1, 2, .., L_{max}\}$. All expectations and conditional distributions discussed herein are with respect to the distribution F.

We imagine a random sample of $\left(X_i, Z_i, \{Y_{i\ell}\}_{\ell=1}^{L_i}, W_i\right)$ for $i = 1, \ldots, n$, generated as follows:

- First $\left(X_i, Z_i, \widetilde{Y}_i, W_i, L_i\right)$ is drawn from F, where L_i is an integer that takes values from 1 to $L_{\max}$.
- Next, $L_i - 1$ independent draws (denoted $\widetilde{Y}_{i\ell}$ for $\ell = 2, ..., L_i$) are made from the distribution of $\widetilde{Y}$ given W.
- The set $\{Y_{i\ell}\}_{\ell=1}^{L_i}$ is a random reordering of $\left\{\widetilde{Y}_i, \left\{\widetilde{Y}_{i\ell}\right\}_{\ell=2}^{L_i}\right\}$.
- The econometrician observes $\left(X_i, Z_i, \{Y_{i\ell}\}_{\ell=1}^{L_i}, W_i\right)$.

We think of this data generating process as being appropriate when the econometrician has access to the "matched" data set, but not the underlying "x-data set" and "y-data set". This will be the focus of the analysis below.

3.3 Linear Instrumental Variables Estimation

In this section, we present methods for the identification and estimation of parameters in an instrumental variables model when the data are generated according to the process described in Section 3.2. Specifically, we consider a "textbook" linear instrumental variables model where the orthogonality condition holds conditional on the variables used for

matching, W_i,

$$\widetilde{Y}_i = X_i'\beta + \varepsilon_i \qquad E\left[Z_i\varepsilon_i \middle| W_i\right] = 0 \tag{3.1}$$

where X_i is a k-dimensional vector and Z_i is an m- dimensional vector such that $m \geq k$.

In the next subsection, we derive a set of population moments that are implied by (3.1) and identify β, but do not require knowledge of the true matches, $\widetilde{Y}$. After that, we present strategies for estimating β.

3.3.1 Identification

To derive a set of valid population moments that can be used to identify β, we write the model in (3.1) as

$$E\left[Z_i\left(\widetilde{Y}_i - X_i'\beta\right)\middle| W_i, L_i\right] = 0, \tag{3.2}$$

or

$$E\left[Z_i\widetilde{Y}_i\middle| W_i\right] = E\left[Z_iX_i' \middle| W_i, L_i\right]\beta,$$

which follows from the fact that L_i is independent of (X_i, Z_i, Y_i) conditional on W_i.

Let $\kappa(L_i)$ be a function that assigns a weight to each observation according to its number of matches. The set of matches $\{Y_{i\ell}\}_{\ell=1}^{L_i}$ is a random reordering of the correct match $\widetilde{Y}_i$ and the false matches $\left\{\widetilde{Y}_{i\ell}\right\}_{\ell=2}^{L_i}$. Thus,

$$\begin{aligned}
E\left[\sum_{\ell=1}^{L_i}\kappa(L_i)\,Z_iY_{i\ell}\right] &= E\left[\kappa(L_i)\,Z_i\widetilde{Y}_i + \kappa(L_i)\sum_{\ell=2}^{L_i}Z_i\widetilde{Y}_{i\ell}\right] \\
&= E\left[\kappa(L_i)\,Z_i\widetilde{Y}_i\right] + E\left[\kappa(L_i)\sum_{\ell=2}^{L_i}Z_i\widetilde{Y}_{i\ell}\right] \\
&= E\left[\kappa(L_i)\,Z_i\widetilde{Y}_i\right] + E\left[E\left[\kappa(L_i)\sum_{\ell=2}^{L_i}Z_i\widetilde{Y}_{i\ell}\middle| W_i, L_i\right]\right] \\
&= E\left[\kappa(L_i)\,Z_i\widetilde{Y}_i\right] + E\left[\kappa(L_i)\sum_{\ell=2}^{L_i}E\left[Z_i\widetilde{Y}_{i\ell}\middle| W_i, L_i\right]\right].
\end{aligned}$$

Each false match $\widetilde{Y}_{i\ell}$ is independent of Z_i conditional on W_i, because they refer to two independent draws from (X, Y, Z), so the last line can be rewritten as,

$$E\left[\sum_{\ell=1}^{L_i}\kappa\left(L_i\right)Z_iY_{i\ell}\right] = E\left[\kappa\left(L_i\right)Z_i\widetilde{Y}_i\right] + E\left[\kappa\left(L_i\right)\sum_{\ell=2}^{L_i}E\left[Z_i\middle|W_i,L_i\right]E\left[\widetilde{Y}_{i\ell}\middle|W_i,L_i\right]\right].$$

Letting $g\left(W_i\right) = E\left[Y_i\middle|W_i,L_i\right] = E\left[Y_i\middle|W_i\right]$,

$$\begin{aligned} E\left[\sum_{\ell=1}^{L_i}\kappa\left(L_i\right)Z_iY_{i\ell}\right] &= E\left[\kappa\left(L_i\right)Z_i\widetilde{Y}_i\right] + E\left[\kappa\left(L_i\right)\sum_{\ell=2}^{L_i}E\left[Z_i\middle|W_i,L_i\right]g\left(W_i\right)\right] \\ &= E\left[\kappa\left(L_i\right)Z_i\widetilde{Y}_i\right] + E\left[\kappa\left(L_i\right)\left(L_i-1\right)Z_ig\left(W_i\right)\right] \end{aligned}$$

Summarizing, we have

$$E\left[\sum_{\ell=1}^{L_i}\kappa\left(L_i\right)Z_iY_{i\ell}\right] = E\left[\kappa\left(L_i\right)Z_i\widetilde{Y}_i\right] + E\left[\kappa\left(L_i\right)\left(L_i-1\right)Z_ig\left(W_i\right)\right]$$

or, rearranging,

$$\begin{aligned} E\left[\kappa\left(L_i\right)Z_i\widetilde{Y}_i\right] &= E\left[\sum_{\ell=1}^{L_i}\kappa\left(L_i\right)Z_i\left\{Y_{i\ell}-\left(L_i-1\right)g\left(W_i\right)\right\}\right] \\ &= E\left[\kappa\left(L_i\right)Z_i\left(\sum_{\ell=1}^{L_i}Y_{i\ell}-\left(L_i-1\right)g\left(W_i\right)\right)\right] \end{aligned}$$

Also, (3.1) implies that

$$\begin{aligned} E\left[\kappa\left(L_i\right)Z_i\widetilde{Y}_i\right] &= E\left[E\left[\kappa\left(L_i\right)Z_i\widetilde{Y}_i\middle|W_i\right]\right] = E\left[E\left[\kappa\left(L_i\right)\middle|W_i\right]E\left[Z_i\widetilde{Y}_i\middle|W_i\right]\right] \\ &= E\left[E\left[\kappa\left(L_i\right)\middle|W_i\right]E\left[Z_iX_i'\beta\middle|W_i\right]\right] = E\left[\kappa\left(L_i\right)Z_iX_i'\right]\beta \end{aligned}$$

so β satisfies the moment conditions

$$E\left[\kappa\left(L_i\right)Z_i\left(\sum_{\ell=1}^{L_i}Y_{i\ell}-\left(L_i-1\right)g\left(W_i\right)\right)\right]-E\left[\kappa\left(L_i\right)Z_iX_i'\right]\beta=0 \tag{3.3}$$

When $\dim(Z_i)=\dim(X_i)$, the explicit expression for β is

$$\beta=E\left[\kappa\left(L_i\right)Z_iX_i'\right]^{-1}E\left[\kappa\left(L_i\right)Z_i\left(\sum_{\ell=1}^{L_i}Y_{i\ell}-\left(L_i-1\right)g\left(W_i\right)\right)\right] \tag{3.4}$$

When $Z_i=X_i$ and $L_i=1$ for all i, (3.4) is the population version of the least squares estimator in a regression of Y_i on X_i. More generally, when $Z_i=X_i$ and L_i is not necessarily 1, it is the population version of the weighted least squares estimator in a regression of $\left(\sum_{\ell=1}^{L_i}Y_{i\ell}-\left(L_i-1\right)g\left(W_i\right)\right)$ on X_i that uses $\kappa\left(L_i\right)$ as the weight for observation i.

If the model is over–identified in the sense that $\dim\left(Z_i\right)>\dim\left(X_i\right)$, then

$$\beta=\left(PE\left[\kappa\left(L_i\right)Z_iX_i'\right]\right)^{-1}PE\left[\kappa\left(L_i\right)Z_i\left(\sum_{\ell=1}^{L_i}Y_{i\ell}-\left(L_i-1\right)g\left(W_i\right)\right)\right] \tag{3.5}$$

for any $k\times m$-matrix such that $E\left[P\kappa\left(L_i\right)Z_iX_i'\right]^{-1}$exists. The choice of P is equivalent to the choice of weighting matrix in Generalized Method of Moments (GMM) estimation. For example when $L_i=1$ for all i, choosing P to be $E\left[X_iZ_i'\right]E\left[Z_iZ_i'\right]^{-1}$ yields the population version of the usual two stage least squares estimator.

3.3.2 Estimation

We now turn to the problem of converting (3.4) and (3.5) into estimators. The first complication in this is that $g\left(\cdot\right)$ must be estimated. In many applications, the y–data set will be much larger than the x–data set. In such cases, it is reasonable to treat $g\left(\cdot\right)$ as if it is known. This is the initial approach considered in this subsection. Later, in Section 3.3.3, we treat $g\left(\cdot\right)$ as an object to be estimated.

The second complication is that the unconditional moment conditions in (3.3) are derived from the conditional moment conditions

$$E\left[Z_i\left(\widetilde{Y}_i - X_i'\beta\right)\middle| W_i, L_i\right] = 0 \tag{3.6}$$

or

$$E\left[\widetilde{Y}_i - X_i'\beta\middle| Z_i, W_i, L_i\right] = 0. \tag{3.7}$$

In those cases, it is possible to estimate β more efficiently by weighting observations, in the same way that $\kappa(L_i)$ assigns weights to observations based on their number of matches. Applying the arguments from the previous section, (3.7) can be rewritten as

$$\begin{aligned} & E\left[\left(\sum_{\ell=1}^{L_i} Y_{i\ell} - (L_i - 1)\, g\left(W_i\right)\right) - X_i'\beta\middle| Z_i, W_i, L_i\right] \\ = \;& E\left[\widetilde{Y}_i - X_i'\beta\middle| Z_i, W_i, L_i\right] + E\left[(L_i - 1)\left(\widetilde{Y}_{i\ell} - g\left(W_i\right)\right)\middle| Z_i, W_i, L_i\right] = 0. \end{aligned}$$

Per the arguments in Chamberlain (1987) and Newey and McFadden (1994), these conditional moments imply the following unconditional moments that yield the efficient GMM estimator:

$$E\left[A\left(W_i, L_i, Z_i\right)\left(\left(\sum_{\ell=1}^{L_i} Y_{i\ell} - (L_i - 1)\, g\left(W_i\right)\right) - X_i'\beta\right)\right] = 0$$

where

$$A\left(W_i, L_i, Z_i\right) = E\left[X_i\middle| W_i, L_i, Z_i\right] V\left[\left(\sum_{\ell=1}^{L_i} Y_{i\ell} - (L_i - 1)\, g\left(W_i\right)\right) - X_i'\beta\middle| W_i, L_i, Z_i\right]^{-1}.$$

The asymptotic distribution of the resulting GMM estimator for β follows from standard arguments (see, for example, Hansen (1982)).

Because few empirical applications perform GMM estimation using the efficient choice of unconditional moment conditions implied by conditional moments such as (3.6) or (3.7),

we focus on deriving an expression for the efficient unconditional moment conditions given

$$E\left[Z_i\left(\left(\sum_{\ell=1}^{L_i} Y_{i\ell} - (L_i - 1)\, g\left(W_i\right)\right) - X_i'\beta\right)\middle| L_i\right] = 0. \tag{3.8}$$

Applying the results in Chamberlain (1987) to (3.8) yields

$$E\left[A\left(L_i\right) Z_i\left(\left(\sum_{\ell=1}^{L_i} Y_{i\ell} - (L_i - 1)\, g\left(W_i\right)\right) - X_i'\beta\right)\right] = 0 \tag{3.9}$$

where

$$A\left(L_i\right) = E\left[X_i Z_i' \middle| L_i\right] E\left[Z_i\left(\left(\sum_{\ell=1}^{L_i} Y_{i\ell} - (L_i - 1)\, g\left(W_i\right)\right) - X_i'\beta\right)^2 Z_i' \middle| L_i\right]^{-1}.$$

Letting $Y_i^* = \left(\sum_{\ell=1}^{L_i} Y_{i\ell} - (L_i - 1)\, g\left(W_i\right)\right)$ and $\varepsilon_i^* = (Y_i^* - X_i'\beta)$, this implies

$$\beta = E\left[A\left(L_i\right) Z_i X_i'\right]^{-1} E\left[A\left(L_i\right) Z_i Y_i^*\right],$$

where

$$\begin{aligned} A\left(L_i\right) &= E\left[X_i Z_i' \middle| L_i\right] E\left[Z_i\left(Y_i^* - X_i'\beta\right)^2 Z_i' \middle| L_i\right]^{-1} \\ &= E\left[X_i Z_i' \middle| L_i\right] E\left[Z_i \varepsilon_i^{*2} Z_i' \middle| L_i\right]^{-1}. \end{aligned}$$

By assumption, $L_i \leq L_{\max}$ (with probability 1), so we only need to be concerned about a finite number of matrices, $A\left(\ell\right)$. Let $\widetilde{\beta}$ be a GMM estimator based on the moment conditions $E\left[B\left(L_i\right) Z_i\left(\left(\sum_{\ell=1}^{L_i} Y_{i\ell} - (L_i - 1)\, g\left(W_i\right)\right) - X_i'\beta\right)\right] = 0$, for some arbitrary $B\left(L_i\right)$. The weight matrices $A\left(\ell\right)$ are consistently estimated by

$$\widehat{A}\left(\ell\right) = \left(\sum_{i:L_i=\ell} X_i Z_i'\right)\left(\sum_{i:L_i=\ell} Z_i\left(Y_i^* - X_i'\widetilde{\beta}\right)^2 Z_i'\right)^{-1}$$

for $\ell = 1, ..., L_{\max}$, and the feasible efficient GMM estimator is

$$\widehat{\beta}_1 = \left(\sum_{i=1}^{n} \widehat{A}(L_i) Z_i X_i' \right)^{-1} \left(\sum_{i=1}^{n} \widehat{A}(L_i) Z_i Y_i^* \right).$$

The asymptotic distribution of $\widehat{\beta}_1$ is

$$\sqrt{n}\left(\widehat{\beta}_1 - \beta\right) \xrightarrow{d} N\left(0, E\left[A(L_i) Z_i X_i'\right]^{-1} E\left[A(L_i) Z_i \varepsilon_i^{*2} Z_i' A(L_i)'\right] \left(E\left[A(L_i) Z_i X_i'\right]^{-1}\right)'\right),$$

following standard arguments for the asymptotic distribution of two-step estimators (see, for example, Newey and McFadden (1994)). Importantly, the asymptotic distribution of $\widehat{\beta}_1$ is not affected by the estimation of $\widehat{A}(\ell)$.

The formula for the asymptotic variance of $\widehat{\beta}_1$ can be simplified. Because $A(L_i) = E\left[X_i Z_i' | L_i\right] E\left[Z_i \varepsilon_i^{*2} Z_i' | L_i\right]^{-1}$, the first and last terms become

$$\begin{aligned} E\left[A(L_i) Z_i X_i'\right]^{-1} &= E\left[A(L_i) E\left[Z_i X_i' | L_i\right]\right]^{-1} \\ &= E\left[E\left[X_i Z_i' | L_i\right] E\left[Z_i \varepsilon_i^{*2} Z_i' \middle| L_i\right]^{-1} E\left[Z_i X_i' | L_i\right]\right]^{-1}. \end{aligned}$$

The middle term becomes

$$\begin{aligned} & E\left[A(L_i) Z_i \varepsilon_i^{*2} Z_i' A(L_i)'\right] \\ = \; & E\left[A(L_i) E\left[Z_i \varepsilon_i^{*2} Z_i' \middle| L_i\right] A(L_i)'\right] \\ = \; & E\left[E\left[X_i Z_i' | L_i\right] E\left[Z_i \varepsilon_i^{*2} Z_i' \middle| L_i\right]^{-1} E\left[Z_i \varepsilon_i^{*2} Z_i' \middle| L_i\right] E\left[Z_i \varepsilon_i^{*2} Z_i' \middle| L_i\right]^{-1} E\left[Z_i X_i' | L_i\right]\right] \\ = \; & E\left[E\left[X_i Z_i' | L_i\right] E\left[Z_i \varepsilon_i^{*2} Z_i' \middle| L_i\right]^{-1} E\left[Z_i X_i' | L_i\right]\right]. \end{aligned}$$

The asymptotic distribution of $\widehat{\beta}_1$ therefore simplifies to

$$\sqrt{n}\left(\widehat{\beta}_1 - \beta\right) \xrightarrow{d} N\left(0, E\left[E\left[X_i Z_i' | L_i\right] E\left[Z_i \varepsilon_i^{*2} Z_i' \middle| L_i\right]^{-1} E\left[Z_i X_i' | L_i\right]\right]^{-1}\right).$$

If, additionally, the errors ε_i^* are homoskedastic within each L_i, such that $E\left[Z_i\varepsilon_i^{*2}Z_i'|\,L_i\right] = \sigma^2\left(L_i\right)E\left[Z_iZ_i'|\,L_i\right]$, then $A\left(L_i\right) = \sigma^2\left(L_i\right)^{-1}E\left[X_iZ_i'|\,L_i\right]E\left[Z_iZ_i'|\,L_i\right]^{-1}$ and the feasible efficient GMM estimator is

$$\widehat{\beta}_2 = \left(\sum_{i=1}^{n}\widehat{A}\left(L_i\right)Z_iX_i'\right)^{-1}\left(\sum_{i=1}^{n}\widehat{A}\left(L_i\right)Z_iY_i^*\right), \tag{3.10}$$

where $\widehat{A}\left(L_i\right) = \widehat{\sigma}^2\left(L_i\right)^{-1}\left(\sum_{i:L_i=\ell}X_iZ_i'\right)\left(\sum_{i:L_i=\ell}Z_iZ_i'\right)^{-1}$, $\widehat{\sigma}^2\left(\ell\right) = n_\ell^{-1}\sum_{i:L_i=\ell}\left(Y_i^* - X_i'\widetilde{\beta}\right)^2$, and n_ℓ is the number of observations for which $L_i = \ell$.

Thus, a more explicit expression for $\widehat{\beta}_2$ is

$$\begin{aligned}
\widehat{\beta}_2 &= \left(\sum_{\ell=1}^{L_i}\widehat{A}\left(\ell\right)\sum_{i:L_i=\ell}Z_iX_i'\right)^{-1}\left(\sum_{\ell=1}^{L_i}\widehat{A}\left(\ell\right)\sum_{i:L_i=\ell}Z_iY_i^*\right) \\
&= \left(\sum_{\ell=1}^{L_i}\widehat{\sigma}^2\left(\ell\right)^{-1}\left(\sum_{i:L_i=\ell}X_iZ_i'\right)\left(\sum_{i:L_i=\ell}Z_iZ_i'\right)^{-1}\sum_{i:L_i=\ell}Z_iX_i'\right)^{-1} \\
&\qquad \left(\sum_{\ell=1}^{L_i}\widehat{\sigma}^2\left(\ell\right)^{-1}\left(\sum_{i:L_i=\ell}X_iZ_i'\right)\left(\sum_{i:L_i=\ell}Z_iZ_i'\right)^{-1}\sum_{i:L_i=\ell}Z_iY_i^*\right).
\end{aligned}$$

Regardless of whether the homoskedasticity assumption is satisfied, the asymptotic distribution of this estimator is

$$\sqrt{n}\left(\widehat{\beta}_2 - \beta\right) \xrightarrow{d} N\left(0, E\left[A\left(L_i\right)Z_iX_i'\right]^{-1}E\left[A\left(L_i\right)Z_i\varepsilon_i^{*2}Z_i'A\left(L_i\right)'\right]\left(E\left[A\left(L_i\right)Z_iX_i'\right]^{-1}\right)'\right).$$

Plugging in $E\left[A\left(L_i\right)Z_iX_i'\right]^{-1} = E\left[\sigma^2\left(L_i\right)^{-1}E\left[X_iZ_i'|\,L_i\right]E\left[Z_iZ_i'|\,L_i\right]^{-1}E\left[Z_iX_i'|\,L_i\right]\right]^{-1}$ and

$$\begin{aligned}
&E\left[A\left(L_i\right)Z_i\varepsilon_i^{*2}Z_i'A\left(L_i\right)'\right] = \\
&E\left[\sigma^2\left(L_i\right)^{-2}E\left[X_iZ_i'|\,L_i\right]E\left[Z_iZ_i'|\,L_i\right]^{-1}E\left[Z_i\varepsilon_i^{*2}Z_i'\middle|\,L_i\right]E\left[Z_iZ_i'|\,L_i\right]^{-1}E\left[Z_iX_i'|\,L_i\right]\right],
\end{aligned}$$

we have

$$\sqrt{n}\left(\widehat{\beta}_2-\beta\right)\xrightarrow{d}N\Bigg(0,E\left[\sigma^2\left(L_i\right)^{-1}E\left[X_iZ_i'|\,L_i\right]E\left[Z_iZ_i'|\,L_i\right]^{-1}E\left[Z_iX_i'|\,L_i\right]\right]^{-1} \tag{3.11}$$
$$E\left[\sigma^2\left(L_i\right)^{-2}E\left[X_iZ_i'|\,L_i\right]E\left[Z_iZ_i'|\,L_i\right]^{-1}E\left[Z_i\varepsilon_i^{*2}Z_i'\big|\,L_i\right]E\left[Z_iZ_i'|\,L_i\right]^{-1}E\left[Z_iX_i'|\,L_i\right]\right]$$
$$\left(E\left[\sigma^2\left(L_i\right)^{-1}E\left[X_iZ_i'|\,L_i\right]E\left[Z_iZ_i'|\,L_i\right]^{-1}E\left[Z_iX_i'|\,L_i\right]\right]^{-1}\right)'\Bigg).$$

The asymptotic variance in (3.11) is valid whether or not the homoskedasticity assumption is satisfied. With homoskedasticity, the asymptotic variance simplifies, such that

$$\sqrt{n}\left(\widehat{\beta}_2-\beta\right)\xrightarrow{d}N\Big(0,E\left[\sigma^2\left(L_i\right)^{-1}E\left[X_iZ_i'|\,L_i\right]E\left[Z_iZ_i'|\,L_i\right]^{-1}E\left[Z_iX_i'|\,L_i\right]\right]^{-1}\Big).$$

A desirable property of the estimator in (3.10) is that, when L_i is the same for all of the observations, it becomes the two stage least squares estimator that uses Y_i^* as the dependent variable. If the explanatory variables are exogeneous such that $Z_i=X_i$, then $\widehat{\beta}_2$ becomes the Generalized Least Squares (GLS) estimator that uses Y_i^* as the dependent variable:

$$\widehat{\beta}_2=\left(\sum_{i=1}^{n}\widehat{\sigma}^2\left(L_i\right)^{-1}\left(X_iX_i'\right)\right)^{-1}\left(\sum_{i=1}^{n}\widehat{\sigma}^2\left(L_i\right)^{-1}X_iY_i^*\right)$$

3.3.3 Controlling for Estimation of $g\left(\cdot\right)$

We now turn to the case where $g\left(\cdot\right)=E\left[Y_i|\,W_i=\cdot\right]$ is estimated, the data are generated as described in Section 3.2, and $g\left(\cdot\right)$ is parameterized as $g\left(w_i\right)=g\left(w_i;\alpha\right)$.

We assume that α is estimated by a standard estimator that can be written as a solution to a set of moment conditions. For example, the estimator could be an extremum estimator, in which case the (population) first order condition of the appropriately formulated objective function delivers the moment conditions. Formally, we write the population

moment conditions for α as

$$E\left[\sum_{\ell=1}^{L_i}\rho\left(Y_{i\ell},W_i,L_i,\alpha\right)\right]=0.$$

If α and β are estimated sequentially, then there is (essentially) no loss of generality in assuming that α is just-identified from the moment conditions (e.g., $\dim(\alpha)=\dim(\rho)$). The asymptotic distribution of $\left(\widehat{\alpha},\widehat{\beta}\right)$ can be derived by stacking the moment conditions for $\widehat{\alpha}$ and the moment conditions for $\widehat{\beta}$:

$$E\left[\begin{pmatrix} E\left[\sum_{\ell=1}^{L_i}\rho\left(Y_{i\ell},W_i,L_i,\alpha\right)\right] \\ E\left[PZ_i\left(\left(\sum_{\ell}Y_{i\ell}-\left(L_i-1\right)g\left(W_i;\alpha\right)\right)-X_i'\beta\right)\right]\end{pmatrix}\right]=\begin{pmatrix}0\\0\end{pmatrix}.$$

As before, P is a $\dim(X_i)\times\dim(Z_i)$ matrix that determines the weight given to each component of $E\left[Z_i\varepsilon_i|\,W_i,L_i\right]$. The joint asymptotic distribution of $\left(\widehat{\alpha},\widehat{\beta}\right)$, which also delivers the asymptotic distribution of $\widehat{\beta}$, follows from standard theory for nonlinear GMM estimation (Newey, 1999).

If α or β are overidentified, it can be more efficient to estimate α and β jointly. To do this, consider a GMM estimator of (α,β) based on the moment conditions

$$\begin{pmatrix} E\left[\sum_{\ell=1}^{L_i}\rho\left(Y_{i\ell},W_i,\alpha\right)\right] \\ E\left[Z_i\left(\left(\sum_{\ell=1}^{L_i}Y_{i\ell}-\left(L_i-1\right)g\left(W_i;\alpha\right)\right)-X_i'\beta\right)\right]\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}.$$

The asymptotic theory of this estimator follows from standard theory for GMM estimators, as does the choice of efficient weighting matrix.

One could also imagine estimating g nonparametrically, but we leave that for future work.

3.4 Monte Carlo Evidence

In this section, we report the results of a small Monte Carlo study designed to evaluate the benefit of our proposed instrumental variables estimation method compared to two benchmark estimators. The first is the standard instrumental variables estimator that uses the average value of all of the Y's that are matched to an observation as that observation's dependent variable. The second is the standard instrumental variables estimator computed using only the observations that are linked to a unique outcome.

3.4.1 Data Generating Process

In the baseline design, there are three explanatory variables: a constant, X_{i1}, an endogenous explanatory variable, X_{i2}, and an exogenous explanatory variable, X_{i3}. There are four instruments, the first of which is a constant. The number of matches is a positive integer no greater than 4. To avoid reporting artificially "nice" results, we deliberately generated the values of the X' s and Z's from "messy"distributions.

The data consist of either 500, 1000 or independent observations generated as follows.

- W is drawn from a uniform distribution between 0 and $\sqrt{12}$. This implies that the variance of W is 1.

- To generate L, we first generate $L^* = 1 + W/2 + N(0,1))/3$, and then set L equal to the rounded values of L^* censored from below at 1 and from above at 4. The resulting probability distribution of L is approximately: $P(L=1) = 0.294$, $P(L=2) = 0.545$, $P(L=3) = 0.161$, and $P(L=4) = 0.001$.

- To generate Z, we start by generating a four-dimensional Z^*. The first element of Z^* is 1. The remaining three elements are drawn from a multivariate normal distribution, where the marginal distribution of each element has mean 0 and vari-

ance 1, and the covariance between any two elements is 0.5. The elements of Z are set equal to those of Z^*, except that Z_2 is rescaled to make its variance equal to 1: $Z_2 = (Z_2^* + W)/\sqrt{2}$.

- The first element of X is 1. To generate X_2 and X_3, we first generate X_2^* and X_3^* independently from a uniform distribution between 0 and $\sqrt{12}$. We then define

$$X_2 = \left(X_2^*/2 + Z_2/2 + \sum_{m=1}^{4} Z_m/4 + W/\sqrt{2} + \varepsilon \right) \Big/ \sqrt{2}$$
$$X_3 = X_3^*/2 + Z_3/2 + \sum_{m=1}^{4} Z_m/4 + 1\{W > 1.8\}/\sqrt{2}.$$

  The variable ε has a standard normal distribution, and will be the error term in the equation for Y, which makes X_2 endogenous.

- Y is generated according to $Y = X'\beta + \varepsilon$, with $\beta = (1, 1, 1,)'$.

The implied function $g(\cdot) = E[Y|\cdot]$ is depicted in Figure 3.1.

We consider deviations from this baseline design in later subsections.

3.4.2 Estimation

We analyze the performance of four estimators: the two benchmark estimators and two versions of our proposed estimation methods. The first benchmark estimator is a "naive" two stage least squares estimator that uses the average of all of the matched Y's as the dependent variable for each observation. The second benchmark estimator is a "conservative" two stage least squares estimator that uses the observations with a unique match ($L_i = 1$). The two remaining estimators are versions of $\widehat{\beta}_2$ where g has been estimated via nonparametric regression or linear regression using a much larger sample with 1,000,000 observations of W, each of which has between 1 and 4 values of Y (for a total of almost 1.9 million observations).

The results from the baseline design are reported in Table 3.1. For each estimator, we report the median and median absolute error (MAE) of the estimates it produces across all 1,000 replications of the study. The results show that both versions of our estimator are more accurate in small samples ($n = 500$) and have lower MAE than the benchmark estimators. That is, our estimator performs well, regardless of whether g is estimated via nonparametric or linear regression.

Additionally, Table 3.2 presents the results from a design where g is estimated using the matched data set instead of the large data set. The results for the benchmark estimators ("Naive" and "Conservative" in the tables) do not change from Table 3.1, because the matched sample has not changed. Surprisingly, the results for our estimators do not change very much when g is estimated using the much smaller matched data set.

One potential limitation of the baseline design is that g is close to linear. It is possible that this biases the results in the direction of making our proposed estimation methods appear "too nice". We therefore test two designs that modify the data generating process for X_3 and result in making g nonlinear:

(a) $X_3 = X_3^*/2 + Z_3/2 + \sum_{m=1}^{4} Z_m/4 + (3.6 - 2W) \cdot 1\{W > 1.8\}/\sqrt{2}$

(b) $X_3 = X_3^*/2 + Z_3/2 + \sum_{m=1}^{4} Z_m/4 + (5.4 - 3W) \cdot 1\{W > 1.8\}/\sqrt{2}$

Modification (a) changes the implied g as depicted in Figure 3.2. The results in Tables 3.3 and 3.4 show that, regardless of whether the large or small sample is used to estimate g, our methods improve upon the MAE of the conservative benchmark estimator by a factor of three for the parameter that is associated with the endogenous variable, X_2.

Modification (b) has a more dramatic effect on the shape of g (see Figure 3.3), however the results, which are presented in Tables 3.5 and 3.6, are very similar to Modification (a).

The results so far could give the impression that our proposed estimation approach is not sensitive to misspecification of g. To illustrate that this is not the case, we test one additional modified design:

(c) First, the data are generated as in Modification (b). Then, the values of Z_3 are modified such that $Z_3 = Z_3 + W + (3.6 - 2W) \cdot 1\{W > 1.8\}$.

Modification (c) leaves g unchanged. However, when g is estimated with linear regression, the estimation error is now correlated with one of the instruments, Z_3. The results from this modified design are presented in Tables 3.7 and 3.8, and show that the estimator based on (incorrectly) estimating g to be linear is not biased for the β_3. The intuition for this is that X_3 is the explanatory variable that is most closely associated with Z_3, and X_3 is exogenous.

3.5 Possible Generalizations

In this section, we discuss extensions for the methods proposed in this paper, including how one might incorporate observations with multiple matches in nonlinear models using GMM and maximum likelihood estimation. We also discuss the limitations of incorporating (potentially incorrect) information about the quality of an observation's matches into estimators that use multiple matches.

3.5.1 Estimating Nonlinear Models

Generalized Method of Moments Estimation

Our proposed methods can be extended to nonlinear models that take the form

$$E\left[m\left(Y_i, X_i, Z_i, \beta\right) \middle| W_i, L_i\right] = 0. \tag{3.12}$$

In this case, X and Z can be combined into one vector of variables, X, such that we can write $E\left[m\left(Y_i, X_i, \beta\right)\middle| W_i, L_i\right] = 0$. With this, (3.12) implies that

$$E\left[\sum_{\ell=1}^{L_i} m\left(Y_{i\ell}, X_i, \beta\right)\middle| W_i, L_i\right] = E\left[m\left(Y_i, X_i, \beta\right)\middle| W_i, L_i\right] + \left(L_i - 1\right) E\left[m\left(Y_i^*, X_i^*, \beta\right)\middle| W_i, L_i\right]$$

where Y^* and X^* are independent draws from the distributions of Y given (W, L) and of X given (W, L), respectively.

If the moment conditions in (3.12) are a consequence of conditional moment conditions with the form $E\left[m\left(Y_i, X_i, Z_i, \beta\right)\middle| Z_i, W_i, L_i\right] = 0$, then it is possible to construct unconditional moments for efficient GMM, as previously discussed in Section 3.3.2.

Maximum Likelihood Estimation

We next consider the case where the nonlinear model is one that would typically be estimated via maximum likelihood estimation if observations were linked to unique outcomes. Let $f\left(y; \theta\middle| X_i\right)$ be the postulated functional for the distribution of Y given X and let $g\left(y\middle| W_i\right)$ be the distribution of Y conditional on W. The function g can be estimated parametrically or nonparametrically using the y-data set. Since this is typically much larger than the matched data set, we treat g as known in the discussion below.

Using the same notation as before, an observation consists of $\left(X_i, \{Y_{i\ell}\}_{\ell=1}^{L_i}, W_i\right)$. For $\ell = 1, ..., L_i$, there is probability $\frac{1}{L_i}$ that $Y_{i\ell}$ is drawn from the parametric model, $f\left(y; \theta\middle| X_i\right)$, and Y_{ik} is drawn from $g\left(y\middle| W_i\right)$ for $k \neq \ell$. This gives the following likelihood function for a single observation:

$$\sum_{\ell=1}^{L_i}\left\{\frac{1}{L_i} f\left(Y_{i\ell}; \theta\middle| X_i\right)\prod_{k\neq\ell} g\left(Y_{ik}\middle| W_i\right)\right\} = \prod_{k=1}^{L_i} g\left(Y_{ik}\middle| W_i\right)\left(\sum_{\ell=1}^{L_i}\frac{1}{L_i}\frac{f\left(Y_{i\ell}; \theta\middle| X_i\right)}{g\left(Y_{i\ell}\middle| W_i\right)}\right).$$

So, except for a constant, the (pseudo) log–likelihood function is

$$\sum_i \log\left(\sum_\ell \frac{1}{L_i}\frac{f\left(Y_{i\ell};\theta|\, X_i\right)}{g\left(Y_{i\ell}|\, W_i\right)}\right). \tag{3.13}$$

In practice, the challenge is to come up with a good g, which must be application-specific.

3.5.2 Using Information on Match Quality

When presented with data containing observations that are linked to multiple outcomes, the researcher may have information about which of an observation's matches is most likely to be the correct match. In this section, we argue that using such information results in biased estimators, unless the probability that a given match is correct is exactly known. We make this point with a very simple example.

Suppose we want to estimate the mean of a random variable Y, $\mu = E[Y]$. For each i, we have two observations, Y_{1i} and Y_{2i}. One is drawn from the true distribution of Y, which has mean μ and variance σ^2, while the other is drawn from another distribution with *known* mean κ and variance ω^2.

Suppose that the probability that the first observation is drawn from the correct distribution is π. Then $E[Y_{1i}] = \pi\mu + (1-\pi)\kappa$ and $E[Y_{2i}] = \pi\kappa + (1-\pi)\mu$, so

$$E[Y_{1i} + Y_{2i}] - \kappa = \pi(\mu+\kappa) + (1-\pi)(\kappa+\mu) - \kappa = \mu.$$

It therefore follows that

$$\frac{1}{n}\sum_{i=1}^{n}(Y_{1i} + Y_{2i}) - \kappa$$

is a consistent estimator of μ.

More generally, consider an estimator of the form

$$\widehat{\mu} = \frac{a_1}{n}\sum_{i=1}^{n} Y_{1i} + \frac{a_2}{n}\sum_{i=1}^{n} Y_{2i} - a_3\kappa \tag{3.14}$$

Its mean would be

$$\begin{aligned} E\left[\widehat{\mu}\right] &= \pi\left(a_1\mu + a_2\kappa\right) + \left(1-\pi\right)\left(a_1\kappa + a_2\mu\right) - a_3\kappa \\ &= \left(\pi a_1 + \left(1-\pi\right)a_2\right)\mu + \left(\pi a_2 + \left(1-\pi\right)a_1 - a_3\right)\kappa \end{aligned} \tag{3.15}$$

For unbiasedness, we then need

$$\pi a_1 + \left(1-\pi\right)a_2 = 1 \quad \text{or, equivalently, } a_2 = \frac{1-\pi a_1}{1-\pi} = \frac{1}{1-\pi} - \frac{\pi}{1-\pi}a_1, \tag{3.16}$$

as well as

$$\left(\pi a_2 + \left(1-\pi\right)a_1 - a_3\right) = 0. \tag{3.17}$$

The only way to do this without knowing π is to set $a_1 = a_2$. But, in that case, (3.17) implies that $a_1 = a_2 = 1$. In other words, there is no way of using inexact information about the quality of a match without introducing bias.

3.6 Conclusion

Popular methods for linking historical data sets can result in linking individuals to multiple possible matches, such that the estimation of standard econometric models requires dropping observations without a unique match. The methods introduced in this paper demonstrate that it is possible to construct estimators that not only accommodate multiple matches per observation, but also improve upon the standard approach. A key advantage of our methods is that they do not require any information about which of an

observation's matches is most likely to be correct, provided that the true match is included among the set of potential matches. Further research is necessary to determine whether identification is possible in settings where some observations are not linked to their correct match, or whether it is possible to construct bounds in those cases. Additional research is also needed to formalize our methods for cases where g, probability distribution of an observation's false matches, is estimated nonparametrically.

3.7 Figures and Tables

Figure 3.1: g Implied by the Baseline Monte Carlo Design

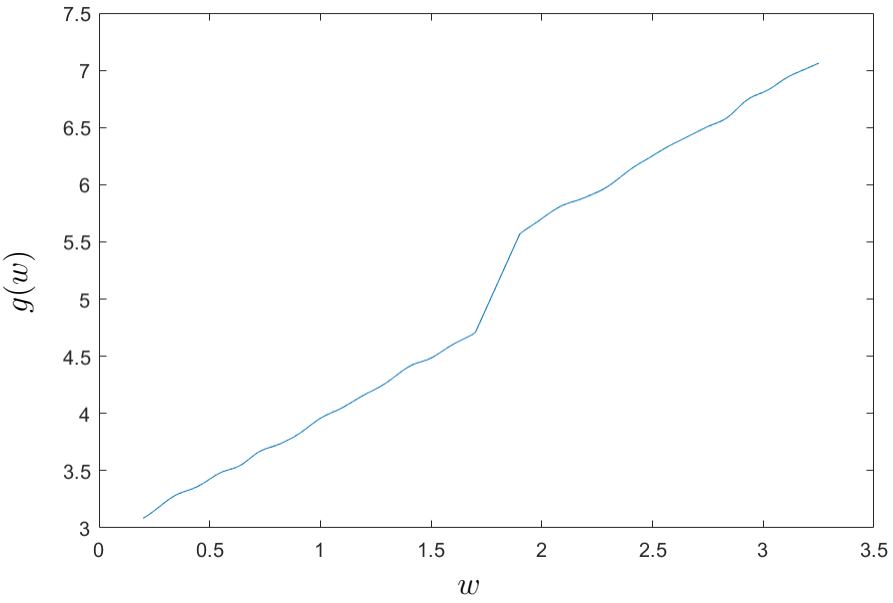


Figure 3.2: g Implied by Modification (a) to the Baseline Design

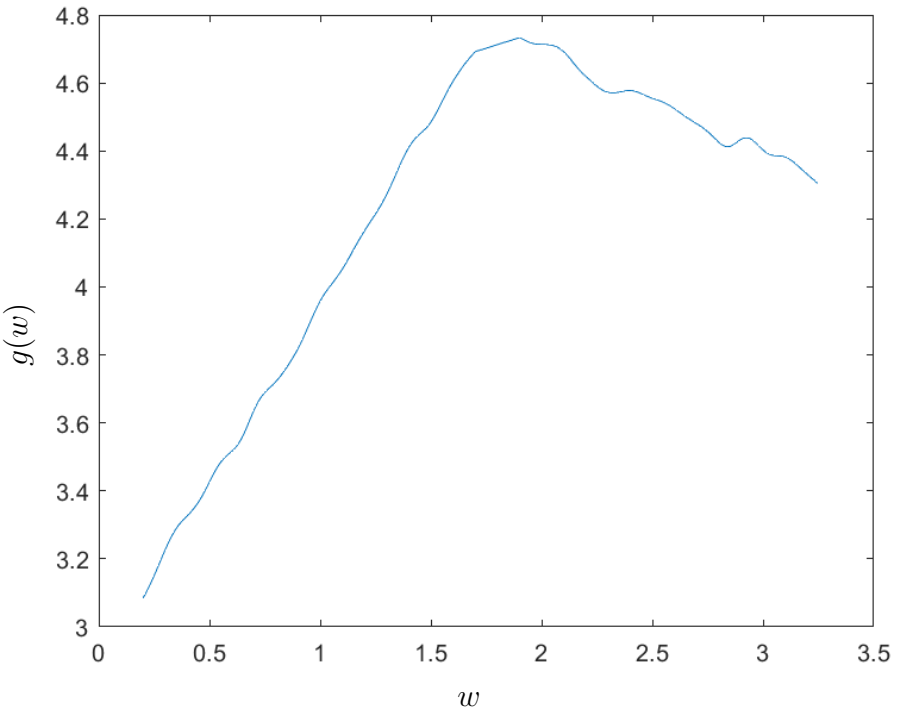


Figure 3.3: g Implied by Modification (b) to the Baseline Design

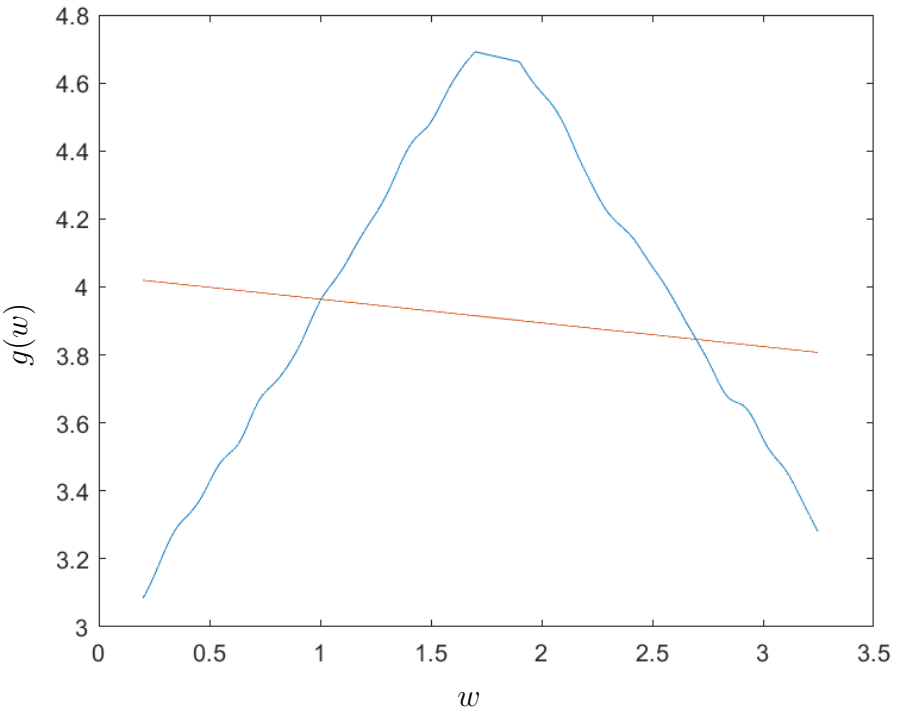


Table 3.1: Baseline Results with Large Sample g Estimates

	Naive		Conservative		Nonparametric g		Linear g	
	Median	MAE	Median	MAE	Median	MAE	Median	MAE
				$n = 500$				
β_1	1.77	0.77	0.86	0.45	0.87	0.30	0.87	0.30
β_2	1.20	0.21	1.15	0.46	1.13	0.23	1.12	0.23
β_3	0.32	0.68	0.93	0.32	0.92	0.20	0.93	0.20
				$n = 1000$				
β_1	1.76	0.76	0.93	0.33	0.94	0.22	0.94	0.22
β_2	1.19	0.19	1.07	0.32	1.08	0.16	1.07	0.16
β_3	0.32	0.68	0.96	0.22	0.95	0.14	0.95	0.14
				$n = 2000$				
β_1	1.80	0.80	0.96	0.22	0.96	0.15	0.96	0.15
β_2	1.18	0.18	1.05	0.23	1.04	0.12	1.04	0.12
β_3	0.32	0.68	0.97	0.15	0.97	0.11	0.97	0.11

Table 3.2: Baseline Results with Small Sample g Estimates

	Naive		Conservative		Nonparametric g		Linear g	
	Median	MAE	Median	MAE	Median	MAE	Median	MAE
				$n = 500$				
β_1	1.77	0.77	0.86	0.45	0.91	0.29	0.88	0.29
β_2	1.20	0.21	1.15	0.46	1.12	0.24	1.11	0.24
β_3	0.32	0.68	0.93	0.32	0.91	0.21	0.93	0.21
				$n = 1000$				
β_1	1.76	0.76	0.93	0.33	0.96	0.21	0.94	0.20
β_2	1.19	0.19	1.07	0.32	1.07	0.17	1.07	0.17
β_3	0.32	0.68	0.96	0.22	0.95	0.15	0.96	0.14
				$n = 2000$				
β_1	1.80	0.80	0.96	0.22	0.97	0.14	0.96	0.15
β_2	1.18	0.18	1.05	0.23	1.04	0.12	1.04	0.12
β_3	0.32	0.68	0.97	0.15	0.97	0.11	0.97	0.11

Table 3.3: Results for Modification (a) with Large Sample g Estimates

	Naive		Conservative		Nonparametric g		Linear g	
	Median	MAE	Median	MAE	Median	MAE	Median	MAE
				$n = 500$				
β_1	1.98	0.98	0.86	0.43	0.88	0.30	0.93	0.30
β_2	0.74	0.26	1.14	0.43	1.06	0.14	1.04	0.14
β_3	0.56	0.44	0.93	0.30	0.99	0.12	1.01	0.12
				$n = 1000$				
β_1	1.99	0.99	0.94	0.31	0.95	0.21	0.99	0.20
β_2	0.73	0.27	1.06	0.30	1.03	0.11	1.01	0.10
β_3	0.56	0.44	0.96	0.21	0.99	0.09	1.01	0.09
				$n = 2000$				
β_1	2.00	1.00	0.96	0.22	0.98	0.15	1.02	0.15
β_2	0.73	0.27	1.05	0.21	1.02	0.07	1.00	0.07
β_3	0.56	0.44	0.97	0.14	1.00	0.07	1.01	0.07

Table 3.4: Results for Modification (a) with Small Sample g Estimates

	Naive		Conservative		Nonparametric g		Linear g	
	Median	MAE	Median	MAE	Median	MAE	Median	MAE
				$n = 500$				
β_1	1.98	0.98	0.86	0.43	0.92	0.30	0.93	0.29
β_2	0.74	0.26	1.14	0.43	1.04	0.15	1.03	0.14
β_3	0.56	0.44	0.93	0.30	0.98	0.13	1.02	0.13
				$n = 1000$				
β_1	1.99	0.99	0.94	0.31	0.96	0.20	0.99	0.20
β_2	0.73	0.27	1.06	0.30	1.02	0.11	1.01	0.10
β_3	0.56	0.44	0.96	0.21	0.98	0.09	1.01	0.09
				$n = 2000$				
β_1	2.00	1.00	0.96	0.22	0.98	0.15	1.02	0.15
β_2	0.73	0.27	1.05	0.21	1.01	0.07	0.99	0.08
β_3	0.56	0.44	0.97	0.14	0.99	0.07	1.01	0.07

Table 3.5: Results for Modification (b) with Large Sample g Estimates

	Naive		Conservative		Nonparametric g		Linear g	
	Median	MAE	Median	MAE	Median	MAE	Median	MAE
				$n = 500$				
β_1	2.04	1.04	0.86	0.43	0.89	0.31	0.94	0.31
β_2	0.65	0.35	1.13	0.42	1.06	0.13	1.03	0.13
β_3	0.61	0.39	0.94	0.30	1.00	0.11	1.03	0.11
				$n = 1000$				
β_1	2.05	1.05	0.94	0.31	0.95	0.21	1.00	0.21
β_2	0.64	0.36	1.06	0.29	1.03	0.10	1.01	0.10
β_3	0.61	0.39	0.96	0.20	0.99	0.08	1.02	0.08
				$n = 2000$				
β_1	2.06	1.06	0.96	0.22	0.98	0.15	1.03	0.15
β_2	0.64	0.36	1.05	0.21	1.01	0.07	0.99	0.07
β_3	0.61	0.39	0.97	0.14	1.00	0.06	1.03	0.06

Table 3.6: Results for Modification (b) with Small Sample g Estimates

	Naive		Conservative		Nonparametric g		Linear g	
	Median	MAE	Median	MAE	Median	MAE	Median	MAE
				$n = 500$				
β_1	2.04	1.04	0.86	0.43	0.93	0.30	0.95	0.30
β_2	0.65	0.35	1.13	0.42	1.04	0.14	1.03	0.13
β_3	0.61	0.39	0.94	0.30	0.99	0.12	1.03	0.11
				$n = 1000$				
β_1	2.05	1.05	0.94	0.31	0.96	0.20	1.00	0.20
β_2	0.64	0.36	1.06	0.29	1.02	0.10	1.00	0.10
β_3	0.61	0.39	0.96	0.20	0.99	0.08	1.03	0.08
				$n = 2000$				
β_1	2.06	1.06	0.96	0.22	0.98	0.15	1.03	0.15
β_2	0.64	0.36	1.05	0.21	1.01	0.07	0.99	0.07
β_3	0.61	0.39	0.97	0.14	0.99	0.06	1.03	0.06

Table 3.7: Results for Modification (c) with Large Sample g Estimates

	Naive		Conservative		Nonparametric g		Linear g	
	Median	MAE	Median	MAE	Median	MAE	Median	MAE
				$n = 500$				
β_1	2.17	1.17	0.81	0.44	0.89	0.30	0.90	0.30
β_2	0.59	0.41	1.24	0.53	1.05	0.14	1.03	0.13
β_3	0.68	0.32	0.85	0.40	1.00	0.11	1.10	0.13
				$n = 1000$				
β_1	2.16	1.16	0.91	0.32	0.94	0.21	0.95	0.21
β_2	0.58	0.42	1.12	0.35	1.03	0.10	1.01	0.10
β_3	0.67	0.33	0.92	0.27	0.99	0.08	1.09	0.09
				$n = 2000$				
β_1	2.18	1.18	0.94	0.22	0.98	0.15	0.99	0.15
β_2	0.58	0.42	1.07	0.24	1.01	0.07	0.99	0.07
β_3	0.68	0.32	0.95	0.19	1.00	0.06	1.09	0.09

Table 3.8: Results for Modification (c) with Small Sample g Estimates

	Naive		Conservative		Nonparametric g		Linear g	
	Median	MAE	Median	MAE	Median	MAE	Median	MAE
				$n = 500$				
β_1	2.17	1.17	0.81	0.44	0.93	0.30	0.91	0.30
β_2	0.59	0.41	1.24	0.53	1.04	0.14	1.03	0.13
β_3	0.68	0.32	0.85	0.40	0.99	0.11	1.10	0.13
				$n = 1000$				
β_1	2.16	1.16	0.91	0.32	0.96	0.20	0.95	0.20
β_2	0.58	0.42	1.12	0.35	1.01	0.10	1.00	0.10
β_3	0.67	0.33	0.92	0.27	0.99	0.08	1.09	0.10
				$n = 2000$				
β_1	2.18	1.18	0.94	0.22	0.98	0.14	0.98	0.15
β_2	0.58	0.42	1.07	0.24	1.01	0.07	0.98	0.07
β_3	0.68	0.32	0.95	0.19	0.99	0.06	1.09	0.10

Bibliography

Abito, Jose Miguel and Yuval Salant, "The Effect of Product Misperception on Economic Outcomes: Evidence from the Extended Warranty Market," *Review of Economic Studies*, 2019, *86* (6), 2285–2318.

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Perez, "Automated Linking of Historical Data," *NBER Working Paper*, 2019.

_ , Leah Platt Boustan, and Katherine Eriksson, "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration," *American Economic Review*, 5 2012, *102* (5), 1832–1856.

_ , _ , and _ , "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration," *Journal of Political Economy*, 2014, *122* (3), 467–506.

_ , _ , and _ , "To the New World and Back Again: Return Migrants in the Age of Mass Migration," *ILR Review*, 2019, *72* (2), 300–322.

_ , Roy Mill, and Santiago Perez, "Linking Individuals Across Historical Sources: a Fully Automated Approach," Technical Report 24324, National Bureau of Economic Research 2 2018.

Aizer, Anna, Shari Eli, Joseph Ferrie, and Adriana Lleras-Muney, "The Long-Run Impact of Cash Transfers to Poor Families," *American Economic Review*, 4 2016, *106* (4), 935–971.

Anderson, Rachel, Bo Honoré, and Adriana Lleras-Muney, "Estimation and inference using imperfectly matched data," *Working paper*, 8 2022.

Auffhammer, Maximilian and Edward Rubin, "Natural Gas Price Elasticities and Optimal Cost Recovery Under Consumer Heterogeneity: Evidence from 300 million natural gas bills," *National Bureau of Economic Research*, 2018.

Averch, Harvey and Leland L Johnson, "American Economic Association Behavior of the Firm Under Regulatory Constraint," *Source: The American Economic Review*, 1962, *52* (5), 1052–1069.

Bailey, Martha, Connor Cole, Morgan Henderson, and Catherine Massey, "How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data," Technical Report 24019, National Bureau of Economic Research 11 2017.

Baldwin, Susan M, "Are Consumers Benefiting from Competition? An Analysis of the Individual Residential Electric Supply Market in Massachusetts," 2018.

Brown, David P. and Andrew Eckert, "The effect of default rates on retail competition and pricing decisions of competitive retailers: The case of Alberta," *Energy Policy*, 2018, *118* (December 2017), 298–311.

Burger, Scott P., Jesse D. Jenkins, Carlos Batlle, and Ignacio J. Pérez-Arriaga, "Restructuring revisited part 1: Competition in electricity distribution systems," *Energy Journal*, 2019, *40* (3), 31–54.

Bushnell, James, Erin T Mansur, and Kevin Novan, "Review of the Economics Literature on US Electricity Restructuring," Technical Report 2017.

Byrne, David P. and Leslie A. Martin, "Consumer search and income inequality," *International Journal of Industrial Organization*, 2021.

_ , _ , and Jia Sheen Nah, "Price Discrimination by Negotiation: A Field Experiment in Retail Electricity," *SSRN Electronic Journal*, 3 2022.

Chamberlain, Gary, "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 3 1987, *34* (3), 305–334.

Chetty, Raj, John N Friedman, and Jonah E Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," Technical Report 19424, National Bureau of Economic Research 9 2013.

_ , _ , Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan, "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star," *The Quarterly Journal of Economics*, 2011, *126* (4), 1593–1660.

Chretien, Larry, Daria Mark, Eugenia Gibbons, Yaima Braga, and Kate Marcellino, "Green Municipal Aggregation in Massachusetts," 2020, (February), 46.

Collins, William J and Marianne H Wanamaker, "Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data," Technical Report 19124, National Bureau of Economic Research 6 2013.

Dempster, A P, N M Laird, and D B Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1977, *39* (1), 1–38.

Deryugina, Tatyana, Alexander MacKay, and Julian Reif, "The long-run dynamics of electricity demand: Evidence from municipal aggregation," *American Economic Journal: Applied Economics*, 2020, *12* (1), 86–114.

Distributed Energy Financial Group LLC, "Annual Baseline Assessment of Choice in Canada and the United States," Technical Report July, Distributed Energy Financial Group LLC 2015.

Dobbie, Will and Roland G Fryer, "Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone," *American Economic Journal: Applied Economics*, 2011, *3* (3), 158–187.

Dormady, Noah, Matthew Hoyt, Alfredo Roa-Henriquez, and William Welch, "Who pays for retail electric deregulation? Evidence of cross-subsidization from complete bill data," *Energy Journal*, 2019, *40* (2), 161–194.

EIA, "Status of State Electric Industry Restructuring Activity as of February 2003," 2003, (February).

_ , "Annual Electric Power Industry Report, Form EIA-861 Detailed Data Files," 2014.

Enamorado, Ted, Benjamin Fifield, and Kosuke Imai, "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records," *American Political Science Review*, 2019, *113* (2), 353?371.

Fellegi, I P and A B Sunter, "A Theory for Record Linkage," *Journal of the American Statistical Association*, 1969, *64*, 1183–1210.

Ferrie, Joseph and Karen Rolf, "Socioeconomic status in childhood and health after age 70: A new longitudinal analysis for the U.S., 1895G2005," *Explorations in Economic History*, 2011, *48* (4), 445–460.

Ferrie, Joseph P, "A New Sample of Males Linked from the Public Use Microdata Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules," *Historical Methods*, 1996, *29* (4), 141–156.

Giulietti, Monica, Jesus Otero, and Michael Waterson, "Pricing behaviour under competition in the UK electricity supply industry," *New Series*, 2010, *62* (3), 478–503.

Hansen, Lars Peter, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 1982, *50* (4), 1029–1054.

Hortaçsu, Ali, Seyed Ali Madanizadeh, and Steven L. Puller, "Power to choose? An analysis of consumer inertia in the residential electricity market," *American Economic Journal: Economic Policy*, 2017, *9* (4), 192–226.

Hsu, David, "choice aggregation and its spread to other states Version spread to other states," 2022.

Jaro, Matthew A, "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, 1989, *84* (406), 414–420.

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz, "Experimental Analysis of Neighborhood Effects," *Econometrica*, 2007, *75* (1), 83–119.

_ , Jens Ludwig, and Lawrence F Katz, "Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment," *The Quarterly Journal of Economics*, 2005, *120* (1), 87–130.

Lahiri, P and Michael D Larsen, "Regression Analysis with Linked Data," *Journal of the American Statistical Association*, 2005, *100* (469), 222–230.

Larsen, Michael D and Donald B Rubin, "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association*, 2001, *96* (453), 32?41.

Littlechild, Stephen, "The regulation of retail competition in US residential electricity markets," Technical Report 2018.

MacKay, Alexander and Ignacia Mercadal, "Deregulation, Market Power, and Prices: Evidence from the Electricity Sector," *SSRN Electronic Journal*, 2021.

Neter, John, E Scott Maynes, and R Ramanathan, "The Effect of Mismatching on the Measurement of Response Errors," *Journal of the American Statistical Association*, 1965, *60* (312), 1005–1027.

Newey, Whitney K., "Consistency of two-step sample selection estimators despite misspecification of distribution," *Economics Letters*, 5 1999, *63* (2), 129–132.

Newey, Whitney K and Daniel McFadden, "Large Sample Estimation and Hypothesis Testing," in Robert F Engle and Daniel L McFadden, eds., *Handbook of Econometrics*, number 4. In 'Handbooks in Economics,.', Amsterdam, London and New York: Elsevier, {N}orth-{H}olland, 1994, pp. 2111–2245.

Office of Consumer Counsel, "OCC FACT SHEET: In Support of Increased Consumer Protection for the Residential Retail Choice Electric Supply Market," Technical Report, State of Connecticut Office of Consumer Counsel 2020.

O'Shaughnessy, Eric, Jenny Heeter, Julien Gattaciecca, Jenny Sauer, Kelly Trumbull, and Emily Chen, "Empowered communities: The rise of community choice aggregation in the United States," *Energy Policy*, 9 2019, *132*, 1110–1119.

SAS, "SAS 9.4 Functions and CALL Routines: Reference, Fifth Edition," Technical Report, SAS Institute Inc. 2019.

Scheuren, Fritz and William Winkler, "Regression analysis of data files that are computer matched," *Survey Methodology*, 1993, *19.*

Singer, Stephen, "Why is Connecticut No. 1 in the cost of electricity among the 48 lower states? Here are 7 reasons," 4 2021.

Su, Xuejuan, "Have customers benefited from electricity retail competition?," *Journal of Regulatory Economics*, 2015, *47* (2), 146–182.

Swadley, Adam and Mine Yücel, "Did residential electricity rates fall after retail competition? A dynamic panel analysis," *Energy Policy*, 12 2011, *39* (12), 7702–7711.

Tsai, Chen-Hao and Yi-Lin Tsai, "Competitive Retail Electricity Market Under Continuous Price Regulation," *SSRN Electronic Journal*, 2017.

Upton, Gregory, Kenneth M. Rosen, and Brittany Tarufelli, "Electricity Market Restructuring and Retail Rates," *SSRN Electronic Journal*, 3 2021.

U.S. Energy Information Administration, "Electric Power Annual 2020," Technical Report 2020.

Wang, Xueting, "Long term contracts in retail electricity," 2021.

Winkler, William E, "Overview of record linkage and current research directions," Technical Report, U.S. Census Bureau 2006.

Chapter 4

Evaluating the joint performance of methods for matching and analyzing linked data

4.1 Introduction

When analyzing multiple data sources with overlapping units, automated matching procedures offer a computationally efficient solution for merging data. These methods allow the researcher to specify a set of matching variables and a decision rule for linking observations to obtain a matched data set in a matter of minutes. If the files to be linked are large, the time saved relative to manual linking or automated linking with clerical review is immense.

In the social sciences, interest in automated matching methods has emerged in response to the increasing availability of administrative data, including recently digitized historical complete count population censuses. These data often contain identifiers that are prone to typographical, duplication, enumeration, and mis-reporting errors. Although methods for merging data sets have been developed in statistics, computer science, operations research, and epidemiology, economic historians and historical demographers have

developed their own matching algorithms out of concern about the accuracy and representativeness of data that are matched using imperfect identifiers (Ferrie, 1996; Abramitzky, Boustan and Eriksson, 2012, 2014, 2019b).

Recent papers by Abramitzky, Boustan, Eriksson, Feigenbaum and Perez (2019a) and Bailey, Cole, Henderson and Massey (2017) evaluate the performance of matching methods that are commonly used in the economics literature. These papers measure the performance of each method by comparing the matched data set that it produces to a "ground truth" data set that is simulated or linked by hand. Both papers document a trade-off between the false positive rate and the (true) match rate across matching procedures, and generally agree that using more conservative algorithms results in a more representative merged data set.

Other examples of papers that compare matching methods are Abramitzky, Mill and Perez (2018) and Enamorado, Fifield and Imai (2019), who study the application of probabilistic matching methods from statistics to match historical and large-scale survey data. The main advantage of using probabilistic matching methods, compared to the deterministic methods studied in Abramitzky, Boustan, Eriksson, Feigenbaum and Perez (2019a), is that they can quantify the uncertainty about the matched data. This extra information can be incorporated into estimation to correct for bias in the Ordinary Least Squares (OLS) estimator for linear regression models that arises from errors in matching (Scheuren and Winkler, 1993; Lahiri and Larsen, 2005).

The ability to combine probabilistic matching methods with estimation methods that take an observation's match probabilities into account illustrates how the choice of matching procedure determines which estimation methods are available for analysis. For example, the methods in Anderson, Honoré and Lleras-Muney (2022) can be used to analyze data sets where observations are linked to multiple possible matches. This issue has received little attention in the literature: although many papers compare the relative

performance of different matching methods, they do not compare the quality of the estimates that they produce during analysis. Thus, researchers could benefit from knowing how different estimation methods for analyzing linked data will perform, prior to choosing whether to use a deterministic or probabilistic matching method, and whether to allow for multiple matches.

To study which combinations of matching and estimation methods work best in practice, this paper conducts a Monte Carlo study to evaluate the joint performance of methods for matching and estimating linear regression models when the variables of interest are recorded in separate data sets. Each replication of the Monte Carlo design involves simulating a "ground truth" data set that is split into separate data files, such that one contains the explanatory variables and the other contains the dependent variable of the linear regression model to be estimated. Each data file contains the same set of identifying variables, however some observation's identifying variables are recorded with error. Also, the identifying variables are generated such that they are not necessarily unique to each observation. This data generating process creates the likelihood that a matching procedure may link observations with error, or to multiple possible matches.

The next step of the Monte Carlo design involves linking the data files using deterministic and probabilistic matching methods to produce a data set with either one match per observation or possibly multiple matches per observation. Then, the linear regression model is estimated using each matched data set via OLS as well as methods that account for how the data were linked. Once all of the replications are finished, various metrics are computed to compare the accuracy of the different matching methods and the estimates that they produce.

The rest of this paper is laid out as follows. The estimation setting and data generating process are described in Section 4.2. Then, Section 4.3 reviews the methods for matching and estimating linear regression models that are tested in this paper. The design of

the Monte Carlo study is described in Section 4.4, and Section 4.5 reports the results. Finally, Section 4.6 concludes.

4.2 Setup

The goal is to estimate β in the linear regression model,

$$y_i = x_i'\beta + \varepsilon_i, \;\; E[\varepsilon_i|x_i] = 0, \;\; E[\varepsilon_i^2|x_i] = \sigma^2 \tag{4.1}$$

using data that satisfy the following assumptions.

Assumption 1. The variables of interest are x and y, which are recorded in separate data files, along with a set of identifiers w. Formally, the raw data consist of an x-data file with N_x observations of (x, w), and a y-data file with N_y observations of (y, w), with $N_y \geq N_x$. Each observation in the x-data file has a unique, corresponding value in the y-data file that satisfies the relationship in (4.1), but the identity of this match is unknown. If, additionally, $N_y > N_x$, some y do not correspond to any observation in the x-data file. Thus, it is necessary to identify which (x, y) pairs describe the same individuals prior to estimating the parameters in (4.1).

Example 1. Aizer, Eli, Ferrie and Lleras-Muney (2016) estimate the impact of providing cash transfers to single mothers on the life expectancy of their children. In their setting, the x-data file consists of mothers' welfare program applications, where x includes a binary variable equal to 1 if the applicant received a cash transfer, and other demographic variables. The y-data file is a universal database of death records, where y is an individual's age at death. The x- and y-data files also contain identifiers w, which include the individual's first and last names, middle initial, and birth date.

Assumption 2. The raw data files are matched using the identifiers w to produce a data set that has the form,

$$\mathcal{D}_n \equiv \left(x_i, w_i, \{y_{i\ell}\}_{\ell=1}^{L_i}, \{\pi_{i\ell}\}_{\ell=1}^{L_i}\right)_{i=1}^{N} \tag{4.2}$$

where $N \leq N_x$, such that some observations in the x-data may not be matched to any value of y. Each observation x_i is linked to the set of outcomes, $\{y_{i\ell}\}_{\ell=1}^{L_i}$. If the matching method produces at most one match per observation, then $L_i = 1$ for all i in the matched data set. If multiple matches per observation are allowed, then $L_i \in \{1, 2, \ldots, L_{max}\}$, and the number of matches varies across observations. If the data are matched via deterministic methods, then the probability that the outcome $y_{i\ell}$ is the true match is $\pi_{i\ell} = \frac{1}{L_i}$. If the data are matched via probabilistic methods, then the match probabilities $\{\pi_{i\ell}\}_{\ell=1}^{L_i}$ are outputted by the matching procedure and sum to 1.

Example 1 (cont'd). Aizer, Eli, Ferrie and Lleras-Muney (2016) link observations using a deterministic method that accounts for potential errors in w. To account for changes in spelling and typographical errors, they convert all names into sounds using a phonetic algorithm, and measure the similarity between two individual's phonetically-spelled names using a string distance metric called SPEDIS (SAS, 2019). They assign as matches all pairs of individuals whose SPEDIS scores and differences in birth dates fall within a prespecified range. Their procedure does not enforce unique matches, so some individuals with common names are matched to multiple death records.

Assumption 3. The observations of x_i and $\{y_{i\ell}\}_{\ell=1}^{L_i}$ included in $\mathcal{D}_n$ comprise random samples conditional on the identifying variables w_i. More specifically, I assume that each x_i is drawn independently from $f_x(x|w)$, its true match is drawn from $f_y(y|x, w)$, and the false matches are drawn independently from $f_y(y|w)$.

Example 1 (cont'd). Assumption 3 requires that all individuals with the same identifying information are equally likely to be included in $\mathcal{D}_n$. This would be violated if, for

example, higher income individuals had a greater probability of appearing in the sample, unless the identifiers w include an individual's income.

4.3 Review of Methods Tested

In this section, I review methods for matching the x- and y- data files to create a merged data set that has the form in (4.2), as well as methods for estimating β in (4.1) that take the matching step into account.

4.3.1 Matching Methods

Formally, a matching method consists of the following steps: (i) selecting and standardizing the identifying variables w, (ii) choosing a metric for quantifying (partial) agreement between the standardized values of w, (iii) choosing which (x, y) pairs to consider as potential matches, and (iv) designating (x, y) pairs as matches.

The purpose of the first step is to increase the probability that an observation is linked to its true match, by correcting for transcription errors that may cause two observations to have different values of w even though they refer to the same individual. For example, if the identifiers w include an individual's name, then the researcher may standardize w by removing spaces and non-alphabetic characters, replacing common nicknames with full names, or using a phonetic algorithm.

The second step defines a metric for quantifying the similarity between non-numeric identifying variables, such as names, addresses, or geographic coordinates. Examples of such metrics are the Jaro-Winkler distance or Levenshtein "edit" distance for strings. These metrics allow the researcher to declare as a "name match", or "partial name match" any observation pair whose string distance is lower than a pre-specified threshold. Importantly, these metrics may be affected by how the identifying variables are standardized.

For example, Abramitzky, Mill and Perez (2018) note that "Abramtziky" is coded differently than "Abramitzky" using the New York State Identification and Intelligence (NYSIIS) phonetic algorithm, but the Jaro-Winkler distance between these two names is very low (0.02, where the minimum distance is 0, and the maximum distance is 1). Also, "James Tennes" and "James Thomas" have the same NYSIIS code, but the Jaro-Winkler distance between "Tennes" and "Thomas" is 0.4

The third step reduces the computational burden of the matching procedure, which is particularly important in settings where the underlying x- and y- data files are large. This step involves dividing observations into non-overlapping "blocks" based on their values of w. While pairs that are assigned to the same block are considered as potential matches, pairs that are assigned to different blocks are automatically designated as non-matches. One example of a "blocking" rule is to consider as potential matches only those pairs of individuals who have the same recorded sex at birth and who were born in the same hospital. If the variables used to assign observations into blocks are recorded with error, then blocking may increase the probability that an observation is not linked to its true match (e.g. the Type II error rate).

The final step involves deciding which observation pairs to designate as matches. How pairs are designated as matches depends on whether the matching method is deterministic or probabilistic. Because deterministic matching methods are application-specific, the details of the deterministic methods tested in this paper are described in Section 4.4.3 after the data generating process for the Monte Carlo study is introduced. The next section presents additional theoretical background on probabilistic matching methods that is useful for understanding the methods implemented in this paper.

Probabilistic Matching Methods

The theory for probabilistic matching methods comes from Fellegi and Sunter (1969), who present the task of merging data files as a classification problem. According to their theory, the goal is to identify whether a pair of observations (i, j) belongs to the set of matches (M), or non-matches (U). Pairs are evaluated according to K comparison criteria, represented as the vector,

$$\boldsymbol{\gamma}_{\mathbf{ij}} = (\gamma_{ij}^1, \ldots, \gamma_{ij}^k, \ldots, \gamma_{ij}^K)$$

such that the probability of observing a particular configuration of $\boldsymbol{\gamma}_{\mathbf{ij}}$ can be written as:

$$P(\boldsymbol{\gamma}_{\mathbf{ij}}) = P(\boldsymbol{\gamma}_{\mathbf{ij}}|M)p_M + P(\boldsymbol{\gamma}_{\mathbf{ij}}|U)p_U \tag{4.3}$$

where $P(\boldsymbol{\gamma}_{\mathbf{ij}}|M)$ and $P(\boldsymbol{\gamma}_{\mathbf{ij}}|U)$ are the probabilities of observing the pattern $\boldsymbol{\gamma}_{\mathbf{ij}}$ conditional on the record pair (i, j) belonging to M or U, and p_M and $p_U = 1 - p_M$ are the marginal probabilities of observing a matched or unmatched pair.

Applying Bayes' Rule, the probability that (i, j) is a match conditional on observing $\boldsymbol{\gamma}_{\mathbf{ij}}$, is

$$P(M|\boldsymbol{\gamma}_{\mathbf{ij}}) = \frac{p_M P(\boldsymbol{\gamma}_{\mathbf{ij}}|M)}{P(\boldsymbol{\gamma}_{\mathbf{ij}})} \tag{4.4}$$

and the probability that (i, j) not a match, conditional on $\boldsymbol{\gamma}_{\mathbf{ij}}$, is

$$P(U|\boldsymbol{\gamma}_{\mathbf{ij}}) = \frac{p_U P(\boldsymbol{\gamma}_{\mathbf{ij}}|U)}{P(\boldsymbol{\gamma}_{\mathbf{ij}})} \tag{4.5}$$

Thus, if the parameters of the mixture distribution in (4.3) are estimated, they can be used to estimate the probability that any two observations refer to the same individual us-

ing the formulas (4.4) and (4.5). These probabilities can in turn be used to designate pairs as matches, or to quantify uncertainty about the matched data set.

Because membership to M or U is not observed, most probabilistic matching methods use the expectation-maximization (EM) algorithm from Dempster et al. (1977) to simultaneously estimate the parameters in (4.3) and classify record pairs as matches or non-matches (see, for example, Lahiri and Larsen (2005)). Additionally, most methods assume that the comparison fields γ_{ij}^k are independent across k, conditional on match status, such that

$$P(\gamma_{ij}|C) = \prod_{k=1}^{K} P(\gamma_{ij}^k|C)^{\gamma_{ij}^k}(1 - Pr(\gamma_{ij}^k|C))^{1-\gamma_{ij}^k} \qquad C \in \{M, U\}. \tag{4.6}$$

The conditional independence assumption significantly reduces the number of parameters necessary to describe $P(\boldsymbol{\gamma_{ij}}|M)$ and $P(\boldsymbol{\gamma_{ij}}|U)$. Another alternative is to model the the mixture probabilities using log-linear models (see Larsen and Rubin (2001)), however the probabilistic method tested in this paper maintains the conditional independence assumption because this is what most empirical applications do.

4.3.2 Estimation Methods for Analyzing Linked Data

In this section, I review two methods for estimating linear regression models using data sets where observations may be linked to possible matches.

Scheuren-Winkler Estimator

The first estimation method comes from Scheuren and Winkler (1993), who build upon the work of Neter, Maynes and Ramanathan (1965). The method assumes that the matched data set consists of observations $(x_i, z_i)_{i=1}^N$, where each x_i is linked with a single

outcome z_i that may or may not correspond to the true outcome, y_i. Specifically,

$$z_i = \begin{cases} y_i & \text{with probability } q_{ii} \\ y_j & \text{with probability } q_{ij} \text{ for } j \neq i,\ j = 1, \ldots, N_y \end{cases}$$

and $\sum_{j=1}^{N_y} q_{ij} = 1,\ i = 1, \ldots, N$, where N_y is the size of the y-data file and N is the size of the matched data set. The OLS estimator that uses z_i as the dependent variable is,

$$\widehat{\beta}_N = (X'X)^{-1}X'z \tag{4.7}$$

and is biased if $q_{ii} \neq 1$ for some i.

Scheuren and Winkler (1993) derive a formula for the bias, conditional on the observed values of y,

$$\text{bias}(\widehat{\beta}_N|y) = E[(\widehat{\beta}_N - \beta)|y] = (X'X)^{-1}X'B \tag{4.8}$$

where $B = (B_1, \ldots, B_n)'$ and $B_i = (q_{ii} - 1)y_i + \sum_{j \neq i} q_{ij}y_j$. Letting $q_i = (q_{i1}, \ldots, q_{iN_y})'$, then each element $B_i = q_i'y - y_i$ is the difference between a weighted average of all of the non-matches y_j and the true outcome y_i. Thus, if B can be estimated consistently, a consistent estimator can be constructed:

$$\widehat{\beta}_{SW} = \widehat{\beta}_N - (X'X)^{-1}X'\widehat{B}^{TR} \tag{4.9}$$

which I henceforth refer to as the SW estimator.

Scheuren and Winkler (1993) propose estimating $\widehat{B}$ using the first and second largest elements of q_i, and their corresponding values y_j to compute

$$\widehat{B}_i^{TR} = (q_{ij_1} - 1)y_{ij_1} + q_{ij_2}y_{ij_2} \tag{4.10}$$

for each i. In principle, $\widehat{B}^{TR}$ can incorporate all of elements in q_i, however doing so may be computationally burdensome. The authors show that if q_{ij1} is sufficiently large for all i, then using just the two largest values of q_i works well in practice.

Anderson, Honoré and Lleras-Muney (2022) Estimator

The second estimation method is an inefficient, but convenient, estimator based on the ideas in Anderson, Honoré and Lleras-Muney (2022). The authors, in a previous version of the paper, consider estimating θ_0 that satisfies the model

$$E_0 \left[m \left(y_i, x_i; \theta_0 \right) \right] = 0 \tag{4.11}$$

where y_i and x_i are vectors or scalars of data for an individual i, the function $m(\cdot)$ is known, and the expectation is taken with respect to the joint distribution $f_0(y, x)$. The data consist of observations $\left(x_i, w_i, \{y_{i\ell}\}_{\ell=1}^{L_i} \right)_{i=1}^{N}$, where the x_i and $y_{i\ell}$ are recorded in distinct data sets and matched according to the identifier w_i. They assume that $L_i > 1$ for some i, so that the identity of the outcome that generates the relationship in (4.11) is unknown.

Under Assumptions 1-3, (4.11) can be rewritten as

$$E_0[m(y_i, x_i; \theta)] = E \left[\sum_{\ell=1}^{L_i} m(y_{i\ell}, x_i; \theta) \right] - E \left[(L_i - 1) g(w_i, L_i; x_i; \theta) \right]$$

where $g(w_i, L_i, x_i; \theta) = E \left[m(y_i, x_i; \theta) | \; w_i, L_i \right]$. Thus, if g is known or can be estimated consistently, then a sample analog of (4.11) can be constructed, and estimation proceeds via Generalized Method of Moments (GMM). I henceforth refer to this method as the AHL estimator.

As noted in Anderson et al. (2022), if the model to be estimated is a linear regression model as in (4.1), then the AHL estimator is the OLS estimator for the transformed

model,

$$\sum_{\ell=1}^{L_i} y_{i\ell} - (L_i - 1)\hat{g}(w_i, L_i) = x_i'\beta + u_i \tag{4.12}$$

and $\hat{g}(w_i, L_i)$ is a (possibly nonparametric) estimator of $E[y_{i\ell}|w_i, L_i]$, $u_i = \varepsilon_i + \sum_{\ell=1}^{L_i} \nu_{i\ell}$, and $\nu_{i\ell} = y_{i\ell} - \hat{g}(w_i, L_i)$.

If, additionally, the errors are homoskedastic in the sense that $E[\varepsilon_i^2|x_i, w_i, L_i] = \sigma_\varepsilon^2$ and $E[\nu_{i\ell}^2|x_i, w_i, L_i] = \sigma_\nu^2$, then the efficient AHL estimator is weighted least squares,

$$\hat{\beta}^{AHL} = \left(\sum_{i=1}^{N} \frac{x_i x_i'}{\sigma(X_i)}\right)^{-1} \left(\sum_{i=1}^{N} \frac{x_i}{\sigma(X_i)} \left(\sum_{\ell=1}^{L_i} y_{i\ell} - (L_i - 1)g(w_i, L_i)\right)\right) \tag{4.13}$$

where $\sigma(X_i) = \sigma_\varepsilon^2 + (L_i - 1)\sigma_\nu^2$.

In practice, $\hat{\beta}^{AHL}$ is estimated in multiple steps. First, g is estimated via parametric or nonparametric methods using all of the y-values in the matched data set, or the larger y-data file if it is available. The residuals from this step are also used to estimate σ_ν^2. Second, σ_ϵ^2 is estimated using the residuals from the OLS estimator applied to the subset of observations that have a unique match. Finally, the estimator in (4.13) is computed using the estimates $\hat{g}$, $\widehat{\sigma}_\epsilon^2$ and $\widehat{\sigma}_\nu^2$.

4.4 Monte Carlo

In this section, I describe the details of a Monte Carlo study to evaluate the combined performance of different methods for linking data sets and estimating linear regression models. Each replication consists of (i) generating unique x- and y-data files, (ii) linking the x- and y-data files to obtain a matched data that has the form of (4.2), and (iii) estimating the parameters of the linear regression model in (4.1) using each of the matched data sets. The next section provides an overview of the study, and the subsequent sections provide additional details on each replication step.

4.4.1 Overview of the Design

Each replication involves generating a unique x- and y- data file according to the procedure in Section 4.4.2 below. The data are generated such that some observations will not be linked to their true match and some observations will be linked to multiple matches.

Each x- and y- data file pair is matched a total of four times to produce four distinct matched data sets. I use two deterministic and two probabilistic matching methods, where the methods within each category differ according to whether they produce either one or multiple matches per observation. The four matching methods and their outputs are summarized in Table 4.1.

For the matched data sets that contain a unique match for each observation, I compute the two benchmark estimators:

1. OLS (Single): the OLS estimator that uses only the matched data set
2. OLS (True): the OLS estimator that uses the correct match for each observation in the matched data set

For the matched data sets that contain potentially multiple matches per observation, I compute:

1. OLS (Single): the OLS estimator that uses only the subset of observations that are linked to a unique outcome
2. OLS (All): the OLS estimator that treats each matched (x, y) pair as a distinct observation, such that observations with multiple matches are counted multiple times
3. OLS (True): the OLS estimator that uses the correct match for each observation in the matched data set
4. the SW estimator in (4.9)

5. the AHL estimator in (4.13)

The above process is repeated for 1,000 replications.

4.4.2 Data Generation Step

I begin by constructing a "ground truth" data set with 1,000 observations of $(x_{1i}, x_{2i}, y_i, w_i)$, where x_{1i} and x_{2i} are mutually independent, i.i.d draws from Bernoulli(0.5) and Normal(0,2) distributions, respectively. The y_i values are generated according to

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad \varepsilon_i \mid x_{1i}, x_{2i} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2) \tag{4.14}$$

where ε_i are independent draws from a Normal(0,2) distribution. I set $(\beta_0, \beta_1, \beta_2, \sigma^2) = (2, 0.5, 1, 2)$, so that estimating the correctly specified linear regression model yields an R^2 value of approximately 0.50.

The identifying variables, w_i, include a first name, last name, and birth year. In total, there are 960 unique first and last name combinations, so multiple observations will be assigned the same first and last name. The birth years are drawn at random from a uniform distribution over the set of integers between 1900 and 1925. The resulting data set resembles the top panel of Figure 4.1.

Next, I split the ground truth data set into the x- and y-data files, as in the bottom panel of Figure 4.1. The x-data file contains values of (x_{i1}, x_{i2}) for 500 observations selected at random from the ground truth data set. The identifiers in the x-data file are equal to the original w_i, plus some type of typographical error. These errors include deleting characters (e.g. "Anderson" becomes "Andersn"), exchanging vowels (e.g. "Rachel" becomes "Rachal"), and swapping English phonetic equivalents (e.g. "Ellie" becomes "Elie"). The probability of introducing a certain type of typographical error matches the probability that such an error appeared in the 1940 Census data in Abramitzky et al.

(2019a)[1] Additionally, half of the observations in the x-data file have a birth year that equals their original birth year plus a random draw from a Normal(0, 2.5) distribution and rounding it to the nearest integer.

The y-data file includes all 1,000 values of y_i from the ground truth data, along with the original identifiers w_i. The result of this choice is that some observations in the x-data file will be linked to multiple values of y.

4.4.3 Matching Step

Deterministic Methods

The two deterministic matching methods used in this paper are based upon those used by Abramitzky, Boustan and Eriksson (2012, 2014, 2019b), and Ferrie (1996). The first method matches each observation in the x-data file to at most one observation in the y-data file, and discards any observations that result in multiple matches. The second method designates as a match *any* pair of observations that have the same phonetically spelled first and last names, and whose birth years fall within a 2-year band, which can result in linking a single observation to multiple matches.

The basic algorithm that I use is as follows,

1. Use the New York State Identification and Intelligence (NYSIIS) phonetic algorithm to obtain phonetically-spelled versions of the names in the x- and y-data files.

2. Restrict the sample to people in the x-data file with unique first name, last name, birth year, and x_i combinations.

3. For each record i in the x-data file, search for a record j in the y-data file whose phonetically spelled first and last names and birth year match exactly.

[1] For example, 7% of observations have misreported first names and 17% of observations have misreported last names.

(a) If there is a *unique* match, designate (i, j) as a match, and stop searching for additional possible matches.

(b) If there are multiple possible matches in the y-data file, discard the observation i.

(c) If there are no observations in the y-data file that match i's exact year of birth, search for a match within ± 1 year of i's reported birth year; and, if this is unsuccessful, search for a match within ± 2 years. If i matches multiple observations at any point, or if none of these attempts produces an exact name match, then discard the observation.

4. Repeat Steps 2 and 3 for each record in the y-data file, searching for matches in the x-data file.

5. Return the matched data set equal to the intersection of the two sets of matches produced by Steps 3 and 4.

The version that allows for multiple matches in exactly the same except that Step 3 is replaced with

3.* Designate as a match any observation in the y-data file that matches i's phonetically spelled first and last name exactly, and whose birth year falls within ± 2 years of i's birth year.

Importantly, these methods only consider as potential matches those pairs of observations that are born within two years of one another. Because I introduce random error into about half of observations' recorded birth years, I expect this restriction to increase the methods' Type II error rates by about 6 percent. This is because half of observations' birth years include a random draw from a Normal(0, 2.5) distribution.

Probabilistic Matching Methods

The probabilistic matching methods analyzed in this paper are implemented using the fastLink package in `R` created by Enamorado, Fifield and Imai (2019). The fastLink package estimates the posterior match probabilities in (4.4) and (4.5) for each pair of observations in the x- and y-data files using the expectation-maximization algorithm. The fastLink package allows the user to specify a lower bound on a pair's posterior match probability for designating it at a match. I set the lower bound equal to 0.7, because the default value of 0.85 does not produce any matches. The fastLink algorithm also allows the user to specify whether unique matches are desired, so I use this option to obtain one matched data set with a unique match per observation[2] and one matched data set that accepts as matches all pairs that have a posterior match probability above 0.7. If an observation is matched to multiple values of y, I normalize its posterior match probabilities to sum to one so they can be used to compute the SW estimator.

The fastLink package uses Jaro-Winkler distances to calculate the similarity between strings (Jaro, 1989; Winkler, 2006). The Jaro-Winkler distance is the default string metric for many automated linking procedures, so this choice best reflects how researchers use probabilistic methods in practice.

4.4.4 Estimation Step

I compute the estimators for each data set listed in Section 4.1.

I compute the SW estimator using the formulas in Section 4.3.2. I set the probabilities q_i equal to the normalized posterior match probabilities when they are available (e.g. for the matched data set linked via probabilistic methods); otherwise, I treat all of an observation's matches as equally likely and randomly select two of the matches to compute (4.10).

[2]To produce unique matches, the fastLink package uses a linear sum assignment program that maximizes the sum of the posterior matching probabilities across all matches, subject to the constraint that no observations in either data file can be matched multiple times.

I compute the AHL estimator using (4.13), by setting g equal to the mean of all of the y's in the matched data file. In the first stage, I estimate $\widehat{\widetilde{\beta}}$ via OLS applied to the transformed model in (4.12). This allows me to estimate σ_ε^2 using the residuals from that regression. Additionally, I estimate σ_ν^2 using the matched values of y. Finally, I compute the AHL estimator in (4.13) using the weights $\hat{\sigma}(X_i) = \hat{\sigma_\varepsilon}^2 + (L_i - 1)\hat{\sigma}_\nu^2$.

4.5 Results

First, I evaluate the performance of the different matching methods to assess the quality of the matched data sets that they produce. Then, I evaluate the results from the estimation step.

4.5.1 Results from Matching Step

The matching methods are evaluated using the metrics reported in Table 4.2. The first column reports the average "match rate" for each method, defined as the proportion of observations in the x-data file that are linked to at least one observation in the y-data file, averaged across the Monte Carlo replications. For deterministic methods, allowing for multiple matches increases the match rate from 71 to 79 percent, which means that the sample size increases by about 8 percent. For probabilistic methods, the average match rate is 74 percent, regardless of whether multiple matches are allowed. By plotting the distribution of each method's match rates in Figure 4.3, we see that the probabilistic methods produce the same match rate across all replications. This is likely because the same posterior probability threshold for designating pairs as matches was used for both methods. Thus, allowing for multiple matches increased the number of observations in the x-data file that were linked to an outcome when the deterministic matching method was used, but not when the probabilistic method was used.

Figure 4.3 also shows that match rates vary within each method by about 10 percentage points across replications. The deterministic method that allows for multiple matches consistently matches more observations than any other procedure, and the deterministic method that produces a unique match for each observation produces match rates below 70 percent about half of the time.

The second column of Table 4.2 displays the average number of unique (x, y) pairs that were designated as matches by each method. For the methods that produce multiple matches, each unique (x, y) pair counts as a distinct match, so the number of matches may exceed the number of observations in the x-data file. For example, the deterministic method that allows for multiple matches produced an average of 505 matches per replication, even though the x-data files include only 500 observations.

Unsurprisingly, the number of matches was higher for methods that allow multiple matches per observation. These procedures also had significantly higher Type I error rates (the proportion of false matches) and lower Type II error rates (the proportion of matches that are not identified) than the methods that matched each observation to a unique outcome. This means that allowing for multiple matches increased the number of correct matches that were identified, but also the number of false matches.

Among the methods that produced unique matches, the probabilistic method produced slightly more matches than the deterministic method; however, the quality of those matches was significantly worse. The average Type I error rate was significantly higher for the probabilistic method compared to the deterministic method (11 percent vs. 3 percent), and the Type II error rate was slightly higher (35 percent vs. 31 percent). Thus, the probabilistic methods in this paper matched more observations, but many of the additional matches were incorrect.

The last column in Table 4.2 reports the average proportion of observations that were linked to their correct match. This includes the case where an observation is linked to

multiple matches, as long as the correct match was included among the set of matches. This is an important metric to study, because the AHL estimator assumes that an observation's correct match is included among its potential matches. Both deterministic methods performed very well according to this metric, as they included the correct match over 97 percent of the time. For the probabilistic method, allowing for multiple matches significantly improved the likelihood that an observation was linked to its correct match.

Finally, I explore in greater the detail the benefit of allowing additional matches in Table 4.3. The table reports the proportion of observations that were linked to $L = 1, 2, \ldots, 6+$, potential matches, as well as the proportion of observations that were linked to their correct match, conditional on the number of matches. The values in the table do not sum to one across values of L because they are the average value across the 1,000 replications. Table 4.3 shows that both methods matched over half of observations to a unique match, and about one third of observations to two matches, on average. Also, less than 10 percent of observations were linked to four or more matches. The deterministic method was more likely to assign multiple matches per observation than the probabilistic method.

4.5.2 Results from Estimation Step

To begin, I compare how the quality of the matched data set affects the accuracy of the benchmark OLS estimators by plotting the distribution of the estimates produced using each type of matched data in Figure 4.2. The three versions of the OLS estimator that are tested are: "OLS (All)", which treats each of the matched (x, y) pairs as a unique observation, such that values of x are repeated as explanatory variables as many times as they are matched to a unique value of y. "OLS (True)" is the OLS estimator that uses only the subset of correct matches contained in a matched data set, and "OLS (Single)"

is the OLS estimator that uses only the set of observations that were matched to a single outcome variable.

For the matched data sets that contain a unique outcome per observation, OLS (All) is the same as OLS (Single) by construction, and so the focus is on comparing OLS (True) and OLS (Single). Figure 4.2 shows that the parameter estimates for β_2 were more accurate when the data matched via the deterministic method. This is because, as noted above, the deterministic method produced higher quality matches than the probabilistic method.

When multiple matches were allowed, the deterministic method still performed slightly better than the probabilistic method, however the difference between the methods was much smaller. Interestingly, there is more overlap between the distributions for OLS (True) and OLS (Single) for these methods, compared to the matching methods that enforce a unique match per individual. This pattern suggests that, when multiple matches are allowed, observations that are linked to a unique match are likely to be higher quality matches. Finally, the estimates for OLS (All) are biased toward zero, because the inclusion of false matches creates measurement error.

Next, I evaluate the performance of the AHL and SW estimators. The distribution of the estimates are plotted in Figure 4.4. The AHL estimates are more accurate than the SW estimates for both data sets, as seen by the fact that the distribution of the AHL estimates is centered around the true value of β_0, but the distribution of the SW estimates is not. Additionally, the distribution of the SW estimates is more dispersed than that of the AHL estimator for all parameter values.

Table 4.4 reports the median absolute deviation of the estimators, where each column refers to a different estimator and each group of rows refers to each type of matched data set.[3] The AHL estimator had a lower median absolute deviation than the OLS estimator

[3]For data that contain unique matches, the AHL and SW estimators cannot be computed, and OLS (All) is the same as OLS (Single), so the corresponding cells in the table are empty.

that used only the subset of observations that were assigned to a unique match, regardless of which matching method was used. Additionally, the median absolute deviation of the AHL estimator was about the same or smaller than the OLS estimator that used the matched data sets with unique matches. These results are consistent with the simulation results in Anderson et al. (2022), which show that the efficient AHL estimator performs well under a variety of data generating processes. Furthermore, the AHL estimator implemented in this paper is a convenient but less efficient version of the estimator in (Anderson et al., 2022), and so the fact that it performs well is highly promising.

4.6 Conclusion

This paper used a Monte Carlo study to evaluate the combined performance of methods for matching and estimating linear regression models when the variables of interest are recorded in separate data sets that must be merged prior to analysis. The results show that the application-specific, deterministic algorithms for linking data generated higher quality matches than the probabilistic matching methods tested in this paper. This difference affected the accuracy of the Ordinary Least Squares estimator, but only if the data were matched using a method that enforces unique matches. When observations were allowed to be linked to multiple, possible matches, both the deterministic and probabilistic matching method produced data sets that yielded similar estimates. Furthermore, the estimator that accommodates multiple matches (from Anderson et al. (2022)) outperformed both the OLS estimator computed using the subset of observations that were linked to a unique match and the OLS estimator computed using the data produced by the deterministic matching procedure that enforces unique matches. These findings demonstrate that it is possible to obtain more precise estimates by allowing for multiple matches per observation, provided the appropriate estimation method is used. Further work is necessary to

investigate whether these patterns generalize to other data generating processes, econometric models, and empirical settings.

4.7 Figures and Tables

Figure 4.1: Creation of Synthetic Datasets

ID	y	x_1	x_2	First Name	Last Name	Birthday
1	y_1	$x_{1,1}$	$x_{2,1}$	Tyler	Ashenfelter	1915-05-13
2	y_2	$x_{1,2}$	$x_{2,2}$	Brandon	Christensen	1904-06-27
				$\vdots$		
195	y_{195}	$x_{1,195}$	$x_{2,195}$	Samantha	Andersen	1914-08-18
196	y_{196}	$x_{1,196}$	$x_{2,196}$	Victoria	Andersen	1918-11-25
				$\vdots$		
1000	y_{500}	$x_{1,500}$	$x_{2,500}$	Vicky	Anderson	1915-04-14

$\downarrow$

x-Datafile

ID	x	Name	Birthday
2	$(x_{1,2}, x_{2,2})$	Branden Christenson	1905-06-27
		...	
195	$(x_{1,195}, x_{2,195})$	Samantha Anderson	1914-08-21
198	$(x_{1,198}, x_{2,198})$	Jon Smyth	1918-12-20
		...	
1000	$(x_{1,1000}, x_{2,1000})$	Vic Andersn	1915-04-14

y-Datafile

ID	y	Name	Birthday
1	y_1	Tyler Ashenfelter	1915-05-13
2	y_2	Brandon Christensen	1904-06-27
		...	
195	y_{195}	Samantha Anderson	1914-08-18
		...	
1000	y_{1000}	Vicky Anderson	1915-04-14

Figure 4.2: Monte Carlo Distribution of Benchmark OLS Estimators

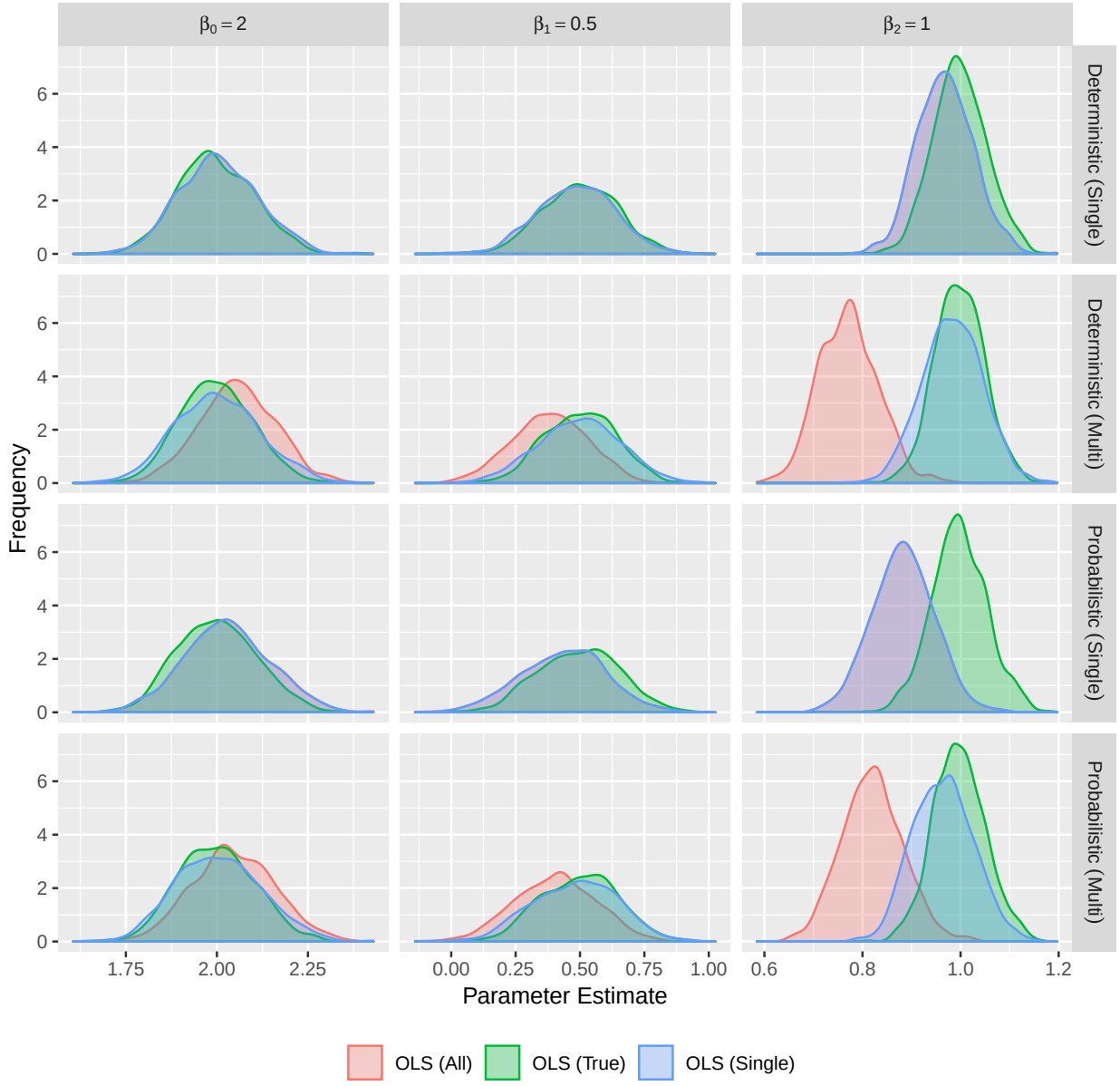
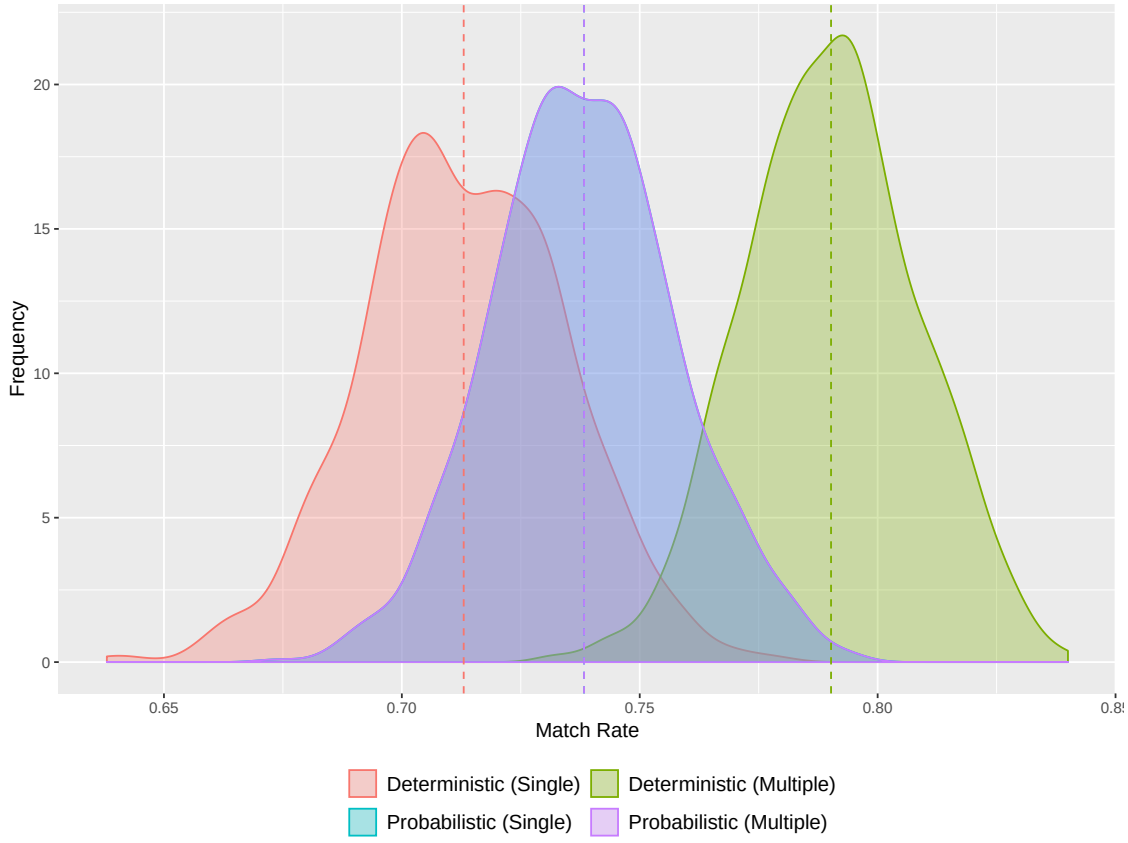


Figure 4.3: Match Rates by Linking Procedure



*Based on 1,000 simulations. Vertical line indicates the sample mean.

Figure 4.4: Monte Carlo Distribution of the SW and AHL Estimators

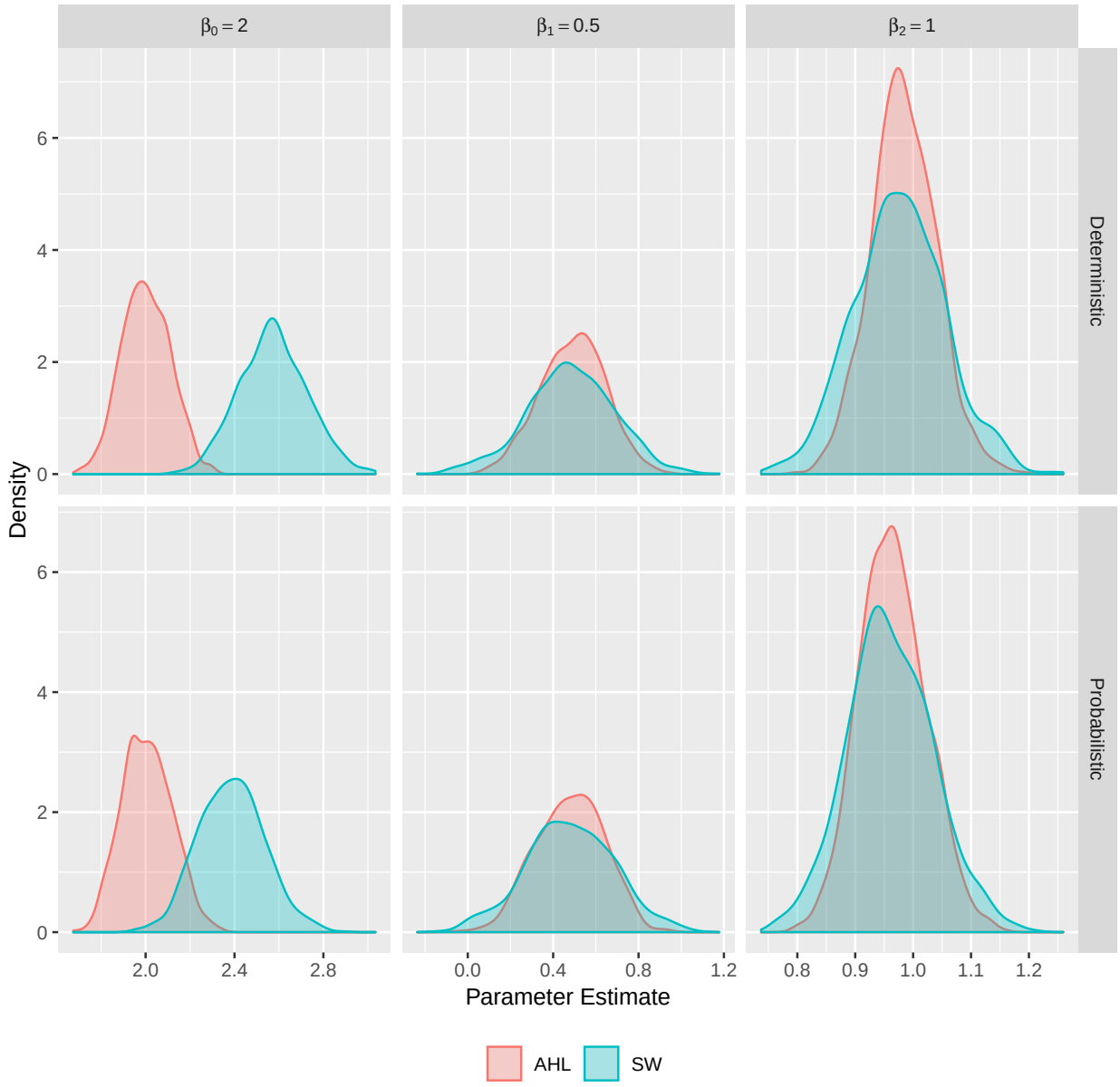


Table 4.1: Overview of Matching Methods

Method	Unique Match	Matched Dataset
Deterministic	Yes	$(x_i, y_i)_{i=1}^{N}$
Deterministic	No	$\left(x_i, \{y_{i\ell}\}_{\ell=1}^{L_i}\right)_{i=1}^{N}$
Probabilistic	Yes	$(x_i, y_i, \pi_i)_{i=1}^{N}$
Probabilistic	No	$\left(x_i, \{y_{i\ell}\}_{\ell=1}^{L_i}, \{\pi_{i\ell}\}_{\ell=1}^{L_i}\right)_{i=1}^{N}$

Table 4.2: Summary of matching algorithm performance

Method	Match Rate	# Matches	Type I	Type II	P(Contains True)
Deterministic (Multiple)	0.79 (0.02)	505.08 (17.30)	0.23 (0.02)	0.22 (0.02)	0.99 (0.01)
Deterministic (Single)	0.71 (0.02)	356.50 (10.60)	0.03 (0.01)	0.31 (0.02)	0.97 (0.01)
Probabilistic (Multiple)	0.74 (0.02)	435.65 (14.94)	0.18 (0.02)	0.28 (0.02)	0.97 (0.01)
Probabilistic (Single)	0.74 (0.02)	369.15 (9.65)	0.11 (0.02)	0.35 (0.02)	0.89 (0.02)

Note: Based on 1,000 simulations. Standard deviations are reported in parentheses.

Table 4.3: Performance of multiple match methods by value of L_i

L	1	2	3	4	5	6+
Deterministic						
Pr(Contains True)	0.99 (0.01)	0.99 (0.01)	0.99 (0.02)	0.99 (0.07)	0.99 (0.10)	1.00 (0.00)
Pr(L=ℓ)	0.52 (0.15)	0.35 (0.16)	0.11 (0.11)	0.03 (0.03)	0.02 (0.03)	0.02 (0.01)
Probabilistic						
Pr(Contains True)	0.97 (0.01)	0.98 (0.02)	0.98 (0.06)	0.98 (0.12)	0.99 (0.05)	1.00 (0.00)
Pr(L=ℓ)	0.59 (0.21)	0.33 (0.22)	0.07 (0.11)	0.02 (0.05)	0.03 (0.06)	0.01 (0.01)

Note: Based on 1,000 simulations. Standard deviations are reported in parentheses.

Table 4.4: Median Absolute Deviations for Estimators

Parameter	AHL	SW	OLS (All)	OLS (True)	OLS (Single)
Deterministic (Single)					
β_0				0.109	0.113
β_1				0.152	0.150
β_2				0.054	0.057
Probabilistic (Single)					
β_0				0.112	0.115
β_1				0.162	0.163
β_2				0.056	0.063
Deterministic (Multiple)					
β_0	0.115	0.155	0.103	0.100	0.120
β_1	0.155	0.201	0.149	0.148	0.159
β_2	0.055	0.077	0.064	0.050	0.061
Probabilistic (Multiple)					
β_0	0.116	0.150	0.112	0.106	0.123
β_1	0.168	0.204	0.165	0.158	0.175
β_2	0.058	0.074	0.062	0.055	0.064

Bibliography

Abito, Jose Miguel and Yuval Salant, "The Effect of Product Misperception on Economic Outcomes: Evidence from the Extended Warranty Market," *Review of Economic Studies*, 2019, *86* (6), 2285–2318.

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Perez, "Automated Linking of Historical Data," *NBER Working Paper*, 2019.

_ , Leah Platt Boustan, and Katherine Eriksson, "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration," *American Economic Review*, 5 2012, *102* (5), 1832–1856.

_ , _ , and _ , "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration," *Journal of Political Economy*, 2014, *122* (3), 467–506.

_ , _ , and _ , "To the New World and Back Again: Return Migrants in the Age of Mass Migration," *ILR Review*, 2019, *72* (2), 300–322.

_ , Roy Mill, and Santiago Perez, "Linking Individuals Across Historical Sources: a Fully Automated Approach," Technical Report 24324, National Bureau of Economic Research 2 2018.

Aizer, Anna, Shari Eli, Joseph Ferrie, and Adriana Lleras-Muney, "The Long-Run Impact of Cash Transfers to Poor Families," *American Economic Review*, 4 2016, *106* (4), 935–971.

Anderson, Rachel, Bo Honoré, and Adriana Lleras-Muney, "Estimation and inference using imperfectly matched data," *Working paper*, 8 2022.

Auffhammer, Maximilian and Edward Rubin, "Natural Gas Price Elasticities and Optimal Cost Recovery Under Consumer Heterogeneity: Evidence from 300 million natural gas bills," *National Bureau of Economic Research*, 2018.

Averch, Harvey and Leland L Johnson, "American Economic Association Behavior of the Firm Under Regulatory Constraint," *Source: The American Economic Review*, 1962, *52* (5), 1052–1069.

Bailey, Martha, Connor Cole, Morgan Henderson, and Catherine Massey, "How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data," Technical Report 24019, National Bureau of Economic Research 11 2017.

Baldwin, Susan M, "Are Consumers Benefiting from Competition? An Analysis of the Individual Residential Electric Supply Market in Massachusetts," 2018.

Brown, David P. and Andrew Eckert, "The effect of default rates on retail competition and pricing decisions of competitive retailers: The case of Alberta," *Energy Policy*, 2018, *118* (December 2017), 298–311.

Burger, Scott P., Jesse D. Jenkins, Carlos Batlle, and Ignacio J. Pérez-Arriaga, "Restructuring revisited part 1: Competition in electricity distribution systems," *Energy Journal*, 2019, *40* (3), 31–54.

Bushnell, James, Erin T Mansur, and Kevin Novan, "Review of the Economics Literature on US Electricity Restructuring," Technical Report 2017.

Byrne, David P. and Leslie A. Martin, "Consumer search and income inequality," *International Journal of Industrial Organization*, 2021.

_ , _ , and Jia Sheen Nah, "Price Discrimination by Negotiation: A Field Experiment in Retail Electricity," *SSRN Electronic Journal*, 3 2022.

Chamberlain, Gary, "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 3 1987, *34* (3), 305–334.

Chetty, Raj, John N Friedman, and Jonah E Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," Technical Report 19424, National Bureau of Economic Research 9 2013.

_ , _ , Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan, "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star," *The Quarterly Journal of Economics*, 2011, *126* (4), 1593–1660.

Chretien, Larry, Daria Mark, Eugenia Gibbons, Yaima Braga, and Kate Marcellino, "Green Municipal Aggregation in Massachusetts," 2020, (February), 46.

Collins, William J and Marianne H Wanamaker, "Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data," Technical Report 19124, National Bureau of Economic Research 6 2013.

Dempster, A P, N M Laird, and D B Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1977, *39* (1), 1–38.

Deryugina, Tatyana, Alexander MacKay, and Julian Reif, "The long-run dynamics of electricity demand: Evidence from municipal aggregation," *American Economic Journal: Applied Economics*, 2020, *12* (1), 86–114.

Distributed Energy Financial Group LLC, "Annual Baseline Assessment of Choice in Canada and the United States," Technical Report July, Distributed Energy Financial Group LLC 2015.

Dobbie, Will and Roland G Fryer, "Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone," *American Economic Journal: Applied Economics*, 2011, *3* (3), 158–187.

Dormady, Noah, Matthew Hoyt, Alfredo Roa-Henriquez, and William Welch, "Who pays for retail electric deregulation? Evidence of cross-subsidization from complete bill data," *Energy Journal*, 2019, *40* (2), 161–194.

EIA, "Status of State Electric Industry Restructuring Activity as of February 2003," 2003, (February).

_ , "Annual Electric Power Industry Report, Form EIA-861 Detailed Data Files," 2014.

Enamorado, Ted, Benjamin Fifield, and Kosuke Imai, "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records," *American Political Science Review*, 2019, *113* (2), 353?371.

Fellegi, I P and A B Sunter, "A Theory for Record Linkage," *Journal of the American Statistical Association*, 1969, *64*, 1183–1210.

Ferrie, Joseph and Karen Rolf, "Socioeconomic status in childhood and health after age 70: A new longitudinal analysis for the U.S., 1895G2005," *Explorations in Economic History*, 2011, *48* (4), 445–460.

Ferrie, Joseph P, "A New Sample of Males Linked from the Public Use Microdata Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules," *Historical Methods*, 1996, *29* (4), 141–156.

Giulietti, Monica, Jesus Otero, and Michael Waterson, "Pricing behaviour under competition in the UK electricity supply industry," *New Series*, 2010, *62* (3), 478–503.

Hansen, Lars Peter, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 1982, *50* (4), 1029–1054.

Hortaçsu, Ali, Seyed Ali Madanizadeh, and Steven L. Puller, "Power to choose? An analysis of consumer inertia in the residential electricity market," *American Economic Journal: Economic Policy*, 2017, *9* (4), 192–226.

Hsu, David, "choice aggregation and its spread to other states Version spread to other states," 2022.

Jaro, Matthew A, "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, 1989, *84* (406), 414–420.

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz, "Experimental Analysis of Neighborhood Effects," *Econometrica*, 2007, *75* (1), 83–119.

_ , Jens Ludwig, and Lawrence F Katz, "Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment," *The Quarterly Journal of Economics*, 2005, *120* (1), 87–130.

Lahiri, P and Michael D Larsen, "Regression Analysis with Linked Data," *Journal of the American Statistical Association*, 2005, *100* (469), 222–230.

Larsen, Michael D and Donald B Rubin, "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association*, 2001, *96* (453), 32?41.

Littlechild, Stephen, "The regulation of retail competition in US residential electricity markets," Technical Report 2018.

MacKay, Alexander and Ignacia Mercadal, "Deregulation, Market Power, and Prices: Evidence from the Electricity Sector," *SSRN Electronic Journal*, 2021.

Neter, John, E Scott Maynes, and R Ramanathan, "The Effect of Mismatching on the Measurement of Response Errors," *Journal of the American Statistical Association*, 1965, *60* (312), 1005–1027.

Newey, Whitney K., "Consistency of two-step sample selection estimators despite misspecification of distribution," *Economics Letters*, 5 1999, *63* (2), 129–132.

Newey, Whitney K and Daniel McFadden, "Large Sample Estimation and Hypothesis Testing," in Robert F Engle and Daniel L McFadden, eds., *Handbook of Econometrics*, number 4. In 'Handbooks in Economics,.', Amsterdam, London and New York: Elsevier, {N}orth-{H}olland, 1994, pp. 2111–2245.

Office of Consumer Counsel, "OCC FACT SHEET: In Support of Increased Consumer Protection for the Residential Retail Choice Electric Supply Market," Technical Report, State of Connecticut Office of Consumer Counsel 2020.

O'Shaughnessy, Eric, Jenny Heeter, Julien Gattaciecca, Jenny Sauer, Kelly Trumbull, and Emily Chen, "Empowered communities: The rise of community choice aggregation in the United States," *Energy Policy*, 9 2019, *132*, 1110–1119.

SAS, "SAS 9.4 Functions and CALL Routines: Reference, Fifth Edition," Technical Report, SAS Institute Inc. 2019.

Scheuren, Fritz and William Winkler, "Regression analysis of data files that are computer matched," *Survey Methodology*, 1993, *19.*

Singer, Stephen, "Why is Connecticut No. 1 in the cost of electricity among the 48 lower states? Here are 7 reasons," 4 2021.

Su, Xuejuan, "Have customers benefited from electricity retail competition?," *Journal of Regulatory Economics*, 2015, *47* (2), 146–182.

Swadley, Adam and Mine Yücel, "Did residential electricity rates fall after retail competition? A dynamic panel analysis," *Energy Policy*, 12 2011, *39* (12), 7702–7711.

Tsai, Chen-Hao and Yi-Lin Tsai, "Competitive Retail Electricity Market Under Continuous Price Regulation," *SSRN Electronic Journal*, 2017.

Upton, Gregory, Kenneth M. Rosen, and Brittany Tarufelli, "Electricity Market Restructuring and Retail Rates," *SSRN Electronic Journal*, 3 2021.

U.S. Energy Information Administration, "Electric Power Annual 2020," Technical Report 2020.

Wang, Xueting, "Long term contracts in retail electricity," 2021.

Winkler, William E, "Overview of record linkage and current research directions," Technical Report, U.S. Census Bureau 2006.

Bibliography

Abito, Jose Miguel and Yuval Salant, "The Effect of Product Misperception on Economic Outcomes: Evidence from the Extended Warranty Market," *Review of Economic Studies*, 2019, *86* (6), 2285–2318.

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Perez, "Automated Linking of Historical Data," *NBER Working Paper*, 2019.

_ **, Leah Platt Boustan, and Katherine Eriksson**, "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration," *American Economic Review*, 5 2012, *102* (5), 1832–1856.

_ **,** _ **, and** _ , "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration," *Journal of Political Economy*, 2014, *122* (3), 467–506.

_ **,** _ **, and** _ , "To the New World and Back Again: Return Migrants in the Age of Mass Migration," *ILR Review*, 2019, *72* (2), 300–322.

_ **, Roy Mill, and Santiago Perez**, "Linking Individuals Across Historical Sources: a Fully Automated Approach," Technical Report 24324, National Bureau of Economic Research 2 2018.

Aizer, Anna, Shari Eli, Joseph Ferrie, and Adriana Lleras-Muney, "The Long-Run Impact of Cash Transfers to Poor Families," *American Economic Review*, 4 2016, *106* (4), 935–971.

Anderson, Rachel, Bo Honoré, and Adriana Lleras-Muney, "Estimation and inference using imperfectly matched data," *Working paper*, 8 2022.

Auffhammer, Maximilian and Edward Rubin, "Natural Gas Price Elasticities and Optimal Cost Recovery Under Consumer Heterogeneity: Evidence from 300 million natural gas bills," *National Bureau of Economic Research*, 2018.

Averch, Harvey and Leland L Johnson, "American Economic Association Behavior of the Firm Under Regulatory Constraint," *Source: The American Economic Review*, 1962, *52* (5), 1052–1069.

Bailey, Martha, Connor Cole, Morgan Henderson, and Catherine Massey, "How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data," Technical Report 24019, National Bureau of Economic Research 11 2017.

Baldwin, Susan M, "Are Consumers Benefiting from Competition? An Analysis of the Individual Residential Electric Supply Market in Massachusetts," 2018.

Brown, David P. and Andrew Eckert, "The effect of default rates on retail competition and pricing decisions of competitive retailers: The case of Alberta," *Energy Policy*, 2018, *118* (December 2017), 298–311.

Burger, Scott P., Jesse D. Jenkins, Carlos Batlle, and Ignacio J. Pérez-Arriaga, "Restructuring revisited part 1: Competition in electricity distribution systems," *Energy Journal*, 2019, *40* (3), 31–54.

Bushnell, James, Erin T Mansur, and Kevin Novan, "Review of the Economics Literature on US Electricity Restructuring," Technical Report 2017.

Byrne, David P. and Leslie A. Martin, "Consumer search and income inequality," *International Journal of Industrial Organization*, 2021.

_ , _ , and Jia Sheen Nah, "Price Discrimination by Negotiation: A Field Experiment in Retail Electricity," *SSRN Electronic Journal*, 3 2022.

Chamberlain, Gary, "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 3 1987, *34* (3), 305–334.

Chetty, Raj, John N Friedman, and Jonah E Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," Technical Report 19424, National Bureau of Economic Research 9 2013.

_ , _ , Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan, "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star," *The Quarterly Journal of Economics*, 2011, *126* (4), 1593–1660.

Chretien, Larry, Daria Mark, Eugenia Gibbons, Yaima Braga, and Kate Marcellino, "Green Municipal Aggregation in Massachusetts," 2020, (February), 46.

Collins, William J and Marianne H Wanamaker, "Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data," Technical Report 19124, National Bureau of Economic Research 6 2013.

Dempster, A P, N M Laird, and D B Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1977, *39* (1), 1–38.

Deryugina, Tatyana, Alexander MacKay, and Julian Reif, "The long-run dynamics of electricity demand: Evidence from municipal aggregation," *American Economic Journal: Applied Economics*, 2020, *12* (1), 86–114.

Distributed Energy Financial Group LLC, "Annual Baseline Assessment of Choice in Canada and the United States," Technical Report July, Distributed Energy Financial Group LLC 2015.

Dobbie, Will and Roland G Fryer, "Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone," *American Economic Journal: Applied Economics*, 2011, *3* (3), 158–187.

Dormady, Noah, Matthew Hoyt, Alfredo Roa-Henriquez, and William Welch, "Who pays for retail electric deregulation? Evidence of cross-subsidization from complete bill data," *Energy Journal*, 2019, *40* (2), 161–194.

EIA, "Status of State Electric Industry Restructuring Activity as of February 2003," 2003, (February).

_ , "Annual Electric Power Industry Report, Form EIA-861 Detailed Data Files," 2014.

Enamorado, Ted, Benjamin Fifield, and Kosuke Imai, "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records," *American Political Science Review*, 2019, *113* (2), 353?371.

Fellegi, I P and A B Sunter, "A Theory for Record Linkage," *Journal of the American Statistical Association*, 1969, *64*, 1183–1210.

Ferrie, Joseph and Karen Rolf, "Socioeconomic status in childhood and health after age 70: A new longitudinal analysis for the U.S., 1895G2005," *Explorations in Economic History*, 2011, *48* (4), 445–460.

Ferrie, Joseph P, "A New Sample of Males Linked from the Public Use Microdata Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules," *Historical Methods*, 1996, *29* (4), 141–156.

Giulietti, Monica, Jesus Otero, and Michael Waterson, "Pricing behaviour under competition in the UK electricity supply industry," *New Series*, 2010, *62* (3), 478–503.

Hansen, Lars Peter, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 1982, *50* (4), 1029–1054.

Hortaçsu, Ali, Seyed Ali Madanizadeh, and Steven L. Puller, "Power to choose? An analysis of consumer inertia in the residential electricity market," *American Economic Journal: Economic Policy*, 2017, *9* (4), 192–226.

Hsu, David, "choice aggregation and its spread to other states Version spread to other states," 2022.

Jaro, Matthew A, "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, 1989, *84* (406), 414–420.

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz, "Experimental Analysis of Neighborhood Effects," *Econometrica*, 2007, *75* (1), 83–119.

_ , Jens Ludwig, and Lawrence F Katz, "Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment," *The Quarterly Journal of Economics*, 2005, *120* (1), 87–130.

Lahiri, P and Michael D Larsen, "Regression Analysis with Linked Data," *Journal of the American Statistical Association*, 2005, *100* (469), 222–230.

Larsen, Michael D and Donald B Rubin, "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association*, 2001, *96* (453), 32?41.

Littlechild, Stephen, "The regulation of retail competition in US residential electricity markets," Technical Report 2018.

MacKay, Alexander and Ignacia Mercadal, "Deregulation, Market Power, and Prices: Evidence from the Electricity Sector," *SSRN Electronic Journal*, 2021.

Neter, John, E Scott Maynes, and R Ramanathan, "The Effect of Mismatching on the Measurement of Response Errors," *Journal of the American Statistical Association*, 1965, *60* (312), 1005–1027.

Newey, Whitney K., "Consistency of two-step sample selection estimators despite misspecification of distribution," *Economics Letters*, 5 1999, *63* (2), 129–132.

Newey, Whitney K and Daniel McFadden, "Large Sample Estimation and Hypothesis Testing," in Robert F Engle and Daniel L McFadden, eds., *Handbook of Econometrics*, number 4. In 'Handbooks in Economics,.', Amsterdam, London and New York: Elsevier, {N}orth-{H}olland, 1994, pp. 2111–2245.

Office of Consumer Counsel, "OCC FACT SHEET: In Support of Increased Consumer Protection for the Residential Retail Choice Electric Supply Market," Technical Report, State of Connecticut Office of Consumer Counsel 2020.

O'Shaughnessy, Eric, Jenny Heeter, Julien Gattaciecca, Jenny Sauer, Kelly Trumbull, and Emily Chen, "Empowered communities: The rise of community choice aggregation in the United States," *Energy Policy*, 9 2019, *132*, 1110–1119.

SAS, "SAS 9.4 Functions and CALL Routines: Reference, Fifth Edition," Technical Report, SAS Institute Inc. 2019.

Scheuren, Fritz and William Winkler, "Regression analysis of data files that are computer matched," *Survey Methodology*, 1993, *19*.

Singer, Stephen, "Why is Connecticut No. 1 in the cost of electricity among the 48 lower states? Here are 7 reasons," 4 2021.

Su, Xuejuan, "Have customers benefited from electricity retail competition?," *Journal of Regulatory Economics*, 2015, *47* (2), 146–182.

Swadley, Adam and Mine Yücel, "Did residential electricity rates fall after retail competition? A dynamic panel analysis," *Energy Policy*, 12 2011, *39* (12), 7702–7711.

Tsai, Chen-Hao and Yi-Lin Tsai, "Competitive Retail Electricity Market Under Continuous Price Regulation," *SSRN Electronic Journal*, 2017.

Upton, Gregory, Kenneth M. Rosen, and Brittany Tarufelli, "Electricity Market Restructuring and Retail Rates," *SSRN Electronic Journal*, 3 2021.

U.S. Energy Information Administration, "Electric Power Annual 2020," Technical Report 2020.

Wang, Xueting, "Long term contracts in retail electricity," 2021.

Winkler, William E, "Overview of record linkage and current research directions," Technical Report, U.S. Census Bureau 2006.