Issues in Online Advertising Markets and Applied Econometrics

by

Charles Edouard Gibbons

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Economics in the GRADUATE DIVISION of the UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor Suzanne Scotchmer, Chair
Professor Aaron Edlin
Professor Jasjeet Sekhon

Spring 2012
Issues in Online Advertising Markets and Applied Econometrics

Copyright 2012
by
Charles Edouard Gibbons
Abstract

Issues in Online Advertising Markets and Applied Econometrics

by

Charles Edouard Gibbons
Doctor of Philosophy in Economics
University of California, Berkeley
Professor Suzanne Scotchmer, Chair

This thesis begins by examining the incentives present in online advertising markets. We consider the strategies of online advertising providers, firms, and consumers in the context of ad listings assigned by a generalized second price auction. The first part of the chapter develops a model of consumer responses to ad listings and product offerings from the included firms and uses this behavioral model to derive optimal bidding functions for the firms. We show that the relationship between per-sale margins and product-consumer match probabilities (“relevances”) must meet certain conditions to rationalize this equilibrium for consumers and firms; in particular, we give the conditions for consumers to rationally search from the top of the listing downward. Next, we turn to the incentives facing the ad server to alter the relevances and margins of the firms and the search costs and valuations of the consumer pool. While these incentives align with the desires of consumers, they may conflict with those for firms. We calculate the optimal number of slots for the ad server to offer, which is less than that desired by firms and consumers. We also show that the ad server has an incentive to subsidize its own competitor in the product market. These results have important implications for competition policy, innovation, and online content provision.

Next, in a chapter coauthored with Juan Carlos Suárez Serrato and Michael B. Urbancic, we turn to a topic in applied econometrics: regressions with fixed effects. Though common in the applied literature, it is known that fixed effects regressions with a constant treatment effect generally do not consistently estimate the sample-weighted treatment effect. This chapter demonstrates the extent of the difference between the fixed effect estimate and the sample-weighted effect by replicating nine influential papers from the American Economic Review. We propose a model with fixed effects interactions to identify the sample-weighted treatment effect and derive a test that discriminates between this estimate and the standard fixed effects estimate. For all 9 papers in our replication, at least one set of fixed effects interactions is jointly significant; in 6 of 9 papers, there is a sample-weighted estimate that is statistically different from the standard fixed effects estimate. In 7 of 9 papers, the differences are economically significant (larger than 10%); the average of the largest difference
between the estimators from each paper is over 50% and the median is 19.5%. Our procedure does not markedly increase the variance of the estimators in 7 of 9 papers.

Lastly, in a chapter coauthored with Michael B. Urbancic, we consider the use and interpretation of estimates from instrumental variables regressions. We elucidate a common mistake in the applied literature in comparing results from OLS and IV models. Often, this comparison serves as a test for exogeneity, but this logic is flawed. We offer closed-form and graphical examples illustrating that equality of these estimates does not imply exogeneity and discuss how the Hausman test applied to the IV setting is misguided. We illustrate our point empirically by comparing estimates of the returns to education using a trio of standard instruments. We conclude by offering guidance for the applied researcher.
To my parents, for their unwavering support and for teaching me the meaning of hard work, success, and the American Dream.
Contents

Abstract 1

Acknowledgements vi

1 Ad Server and Firm Strategies in Contextual Advertising Auctions 1

1.1 Introduction ................................................................. 1
1.2 Model ........................................................................ 3
  1.2.1 Framework .............................................................. 3
  1.2.2 Product market outcomes ........................................ 5
  1.2.3 Discussion .............................................................. 8
1.3 Ad Auction Bidding Behavior ........................................ 9
  1.3.1 Ranking of firms in the ad listing .............................. 10
  1.3.2 Deriving equilibrium bids ......................................... 12
  1.3.3 Discussion .............................................................. 14
1.4 Incentives of ad servers to change search structure .......... 16
  1.4.1 Proportional changes in the relevances ..................... 17
  1.4.2 Proportional changes in search costs ......................... 24
  1.4.3 Proportion increase in high value consumers ............. 28
  1.4.4 Proportion increase in margins .................................. 28
1.5 Impact of dispersion of firm characteristics on bids .......... 28
  1.5.1 Dispersion in margins .............................................. 29
  1.5.2 Dispersion in relevances .......................................... 29
1.6 Choosing the optimal number of ads to display .............. 30
1.7 Subsidizing the bid of a firm ......................................... 31
1.8 Extensions and Conclusions .......................................... 34
1.9 Attrition by Low Value Consumers ................................. 36

2 Broken or Fixed Effects? ................................................. 38

2.1 Introduction ................................................................. 38
2.2 Incorporating heterogeneous treatment effects ............... 39
2.3 Interpreting FE estimates using projection results .......... 41
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.1</td>
<td>FE model estimates compared to the SWE</td>
<td>41</td>
</tr>
<tr>
<td>2.3.2</td>
<td>A Test of Equality Between Sample-Weighted and FE Estimates</td>
<td>43</td>
</tr>
<tr>
<td>2.4</td>
<td>A Case Study: Karlan and Zinman (2008)</td>
<td>44</td>
</tr>
<tr>
<td>2.5</td>
<td>Fixed Effects Interactions: An <em>AER</em> Investigation</td>
<td>46</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Replication Results</td>
<td>46</td>
</tr>
<tr>
<td>2.5.2</td>
<td>The interacted and FE models and the variance-bias tradeoff</td>
<td>47</td>
</tr>
<tr>
<td>2.6</td>
<td>Conclusion</td>
<td>53</td>
</tr>
<tr>
<td>2.A</td>
<td>Topics in Fixed Effects Theory</td>
<td>54</td>
</tr>
<tr>
<td>2.A.1</td>
<td>Sufficient Conditions for Estimation of Sample-Weighted Treatment Effects in FE Models</td>
<td>54</td>
</tr>
<tr>
<td>2.A.2</td>
<td>Calculating the Difference Between the Fixed Effects and Weighted Interactions Estimators</td>
<td>55</td>
</tr>
<tr>
<td>2.B</td>
<td><em>GSSUtest.ado</em></td>
<td>58</td>
</tr>
<tr>
<td>2.C</td>
<td><em>AER</em> Replications</td>
<td>58</td>
</tr>
<tr>
<td>2.C.1</td>
<td>Paper Selection</td>
<td>58</td>
</tr>
<tr>
<td>2.C.2</td>
<td>Replication Details</td>
<td>60</td>
</tr>
<tr>
<td>3</td>
<td>LATE for School: Instrumental Variables and the Returns to Education</td>
<td>70</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>70</td>
</tr>
<tr>
<td>3.2</td>
<td>Simple Examples of LATE</td>
<td>72</td>
</tr>
<tr>
<td>3.2.1</td>
<td>A LATE Example: Regression Discontinuity</td>
<td>72</td>
</tr>
<tr>
<td>3.2.2</td>
<td>A Simple Example</td>
<td>73</td>
</tr>
<tr>
<td>3.2.3</td>
<td>A Graphical Comparison of LATE and ATE</td>
<td>74</td>
</tr>
<tr>
<td>3.3</td>
<td>A Formal Comparison of LATE and ATE</td>
<td>75</td>
</tr>
<tr>
<td>3.4</td>
<td>Hausman Specification Tests</td>
<td>82</td>
</tr>
<tr>
<td>3.5</td>
<td>Causal Effects and the Returns to Education</td>
<td>83</td>
</tr>
<tr>
<td>3.5.1</td>
<td>LATE v. ATE in the Returns to Education Literature</td>
<td>84</td>
</tr>
<tr>
<td>3.6</td>
<td>An Empirical Illustration of LATE</td>
<td>85</td>
</tr>
<tr>
<td>3.6.1</td>
<td>IV estimates</td>
<td>85</td>
</tr>
<tr>
<td>3.6.2</td>
<td>Validity of the Three Instruments</td>
<td>88</td>
</tr>
<tr>
<td>3.7</td>
<td>Conclusion</td>
<td>88</td>
</tr>
<tr>
<td>3.A</td>
<td>Data</td>
<td>89</td>
</tr>
<tr>
<td>3.B</td>
<td>Testing for Weak Instruments</td>
<td>90</td>
</tr>
</tbody>
</table>
List of Figures

1.1 Ratio of bid to expected margin .................................................. 14
1.2 Ratio of cost per click to the expected value per click ............................ 15
1.3 Impact of a 20% increase in relevance from $q = 0.2$ .......................... 21
1.4 Impact of changes in relevance from $q = 0.2$ on aggregates .................... 22
1.5 Impact of changes in relevance for consumers .................................... 23
1.6 Impact of a 20% increase in search frequencies from $s = 0.6$ ................. 26
1.7 Impact of changes in search frequencies from $s = 0.6$ on aggregates ........ 27
1.8 Impact of variation in per-sale margin on auction bids and revenue ............ 29
1.9 Impact of variation in per-sale margin on auction bids and revenue .......... 30
1.10 Change in ad server profits from both ad and product sales after privileging its own firm ................................................................. 34
2.1 The relationship between the difference in the estimators and the change in variance among the AER repliations ................................................. 49
3.1 The response curve to treatment with examples of a MTE and LATE equal to the ATE ................................................................. 74

List of Tables

2.1 Karlan and Zinman (2008) treatment effect weighting ............................ 45
2.2 Papers from the AER used in the meta-analysis ................................... 50
2.3 AER replication results ..................................................................... 51
2.4 AER replication results, continued .................................................... 52
2.5 Replication sources .......................................................................... 60
2.6 Fixed effects interactions and regressions by subgroup conducted in the original papers ................................................................. 61
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.7 Detailed replication results</td>
<td>62</td>
</tr>
<tr>
<td>2.8 Detailed replication results, continued</td>
<td>63</td>
</tr>
<tr>
<td>2.9 Detailed replication results, continued</td>
<td>64</td>
</tr>
<tr>
<td>2.10 Detailed replication results, continued</td>
<td>65</td>
</tr>
<tr>
<td>2.11 Detailed replication results for Banerjee and Iyer (2005)</td>
<td>66</td>
</tr>
<tr>
<td>2.12 Detailed replication results for Banerjee and Iyer (2005), continued</td>
<td>67</td>
</tr>
<tr>
<td>2.13 Detailed replication results for Banerjee and Iyer (2005), continued</td>
<td>68</td>
</tr>
<tr>
<td>2.14 Detailed replication results for Banerjee and Iyer (2005), continued</td>
<td>69</td>
</tr>
<tr>
<td>3.1 IV estimates</td>
<td>87</td>
</tr>
<tr>
<td>3.2 First-stage $F$-tests for weak instruments</td>
<td>91</td>
</tr>
<tr>
<td>3.3 Robust IV estimates</td>
<td>92</td>
</tr>
</tbody>
</table>
Acknowledgments

During graduate school, we are required to not only learn the tools of economics, but also the craft of economics. In this light, three professors helped me better understand how I could apply the discipline’s abstract methods to real world problems. Suzanne Scotchmer exposed me to the fascinating and complicated world of digital markets and emphasized the importance of understanding the institutional details peculiar to each issue. Her optimism and encouragement were sources of constant support. Jasjeet Sekhon provided me with the skepticism necessary to be a critical consumer and producer of applied econometric research. His insights showed me the assumptions and limitations of the models that we employ and have made my research more thorough and carefully considered. Gregory Duncan provided a practical lens to view econometric models and sharpened my ability to use these tools to provide the best possible answers to timely questions. My education is far richer because of the many conversations that I have shared with these professors during my time at Berkeley and their perspectives have greatly influenced my approach to economic and statistical questions.

I would also like to recognize Professors Aaron Edlin and Pat Kline for their helpful guidance in developing the chapters of this thesis. For their suggestions, assistance and friendship, I thank my fellow graduate students, including Daniel Gross, John Mondragon, Alex Rothenberg, Juan Carlos Suárez Serrato, and Mike Urbancic. I would like to thank the Searle Center on Law, Regulation, and Economic Growth and the U.S. Federal Trade Commission for inviting me to present my research at their conferences.

My work has been supported, in part, by fellowships and grants from the National Science Foundation’s Integrative Graduate Education and Research Traineeship (IGERT) program in Politics, Economics, Psychology, and Public Policy at UC Berkeley and the Economics Department at UC Berkeley. I would like to note in particular the support provided by the Institute for Humane Studies. Not only did IHS provide generous financial assistance, the organization also challenged me to think more deeply about topics in economic and political philosophy, broadening and deepening my education.
Chapter 1

Ad Server and Firm Strategies in Contextual Advertising Auctions

1.1 Introduction

Advertising is essential in funding online content, from social networking sites to newspaper articles to streaming music and search engines. On all these sites, there is a movement toward contextual ads that are related to keywords found on the page. These ads aim to generate immediate action by consumers, including clicking a link and performing an “action,” such as purchasing a product from the advertiser’s site. By leading to immediate sales, the listings of these ads can impact the market for the advertised product.

The provider of these ads, known as the ad server, may have an incentive to reduce the competitiveness of the product market in order to extract higher advertising revenues from firms that pay a premium to be listed in a less competitive space. Alternatively, the ad server could be encouraged to find ways innovate and increase sales in a competitive market to raise its ad revenue. A close examination of these issues is necessary as interest turns toward competition policy for the online advertising market.

In this chapter, we focus on the incentives facing an ad server to manipulate the primitives of the market in order to generate higher advertising revenue. In particular, we consider the incentives to improve consumer-product match probabilities, reduce search costs, boost firm product margins, and target profitable consumer groups. We also examine how dispersion in the margins and match probabilities among firms impact ad revenues and how the ad server may limit the length of the ad listing to combat dispersion. Lastly, we consider an ad server that is also active in the market for the product being advertised (e.g., Google providing ads for e-mail services, including its own Gmail offering) to determine whether the ad server can increase total profits from combined product and ad sales by artificially placing its own firm at the top of the listing. We determine whether all these incentives align with
those of firms and consumers.

The incentives of all parties agree to reduce search costs. The ad server has an incentive to fulfill consumers’ wishes for better product matches, while this innovation exceeds the level desired by firms. The incentives of the ad server may be contrary to those of consumers, however. We show that the ad server displays a limited number of ads, while more is always better for consumers and for firms in total. Additionally, the ad server has an incentive to subsidize its own competitor in the product market, which changes the sizes and distributions of producer and consumer surpluses. This latter point has important implications for antitrust inquiries into Google in particular.

In answering these questions, we begin by formulating a model of consumer responses to contextual advertising. Consumers need to be matched to a product that they like and only a fraction of consumers like a product offered by a particular firm. For example, a consumer may be searching for a sweater, but may want a different style than the particular one offered by an advertiser. A fraction of consumers are satisfied by a firm’s offering and have a positive valuation for that particular product, while the rest do not like it, giving them 0 valuation. The probability that a consumer likes the product offered by a firm (i.e., has positive valuation for it) is called the firm’s relevance.

For a consumer to make a purchase, the product must not only be relevant, but also its price must be below the consumer’s valuation. We first consider a model in which a common market price prevails across all firms. Each firm knows its relevance as well as its cost. Based upon these factors, the firms bid for placement in the advertising list via a generalized second price auction. We incorporate our model of consumer behavior into the Varian (2007) framework to consider how heterogeneous margins and relevances of the firms impact bidding strategies.

Under the special case that all the firms have the same cost, firms sort by relevance, as in the model of Athey and Ellison (2009). Since better matches are placed at the top, consumers are most likely to find a relevant product quickly and consumer surplus is maximized, as is producer surplus. Alternatively, if firms all have the same relevance, they sort in decreasing order of cost, another ordering that maximizes total surplus. If both costs and relevance vary across firms, the ordering depends upon how these characteristics covary and we derive the maximum covariance between these factors that can exist in an equilibrium with consumers that search from the top down.

The ranking of firms, then, is endogenous to the model. We require this feature in order to understand the incentives that the ad server faces to improve or alter these rankings. Our model builds upon other models of ordered consumer search that assume exogenous rankings (see Arbatskaya, 2007; Armstrong, Vickers and Zhou, 2009; Xu, Chen and Whinston, 2011). The work of Xu, Chen and Whinston (2011) endogenizes firm rank, but the model of consumer behavior here is most similar to the work of Chen and He (2006) and Athey and Ellison (2009), models that also endogeneize firm rank and incorporate the idea of relevance. All but the latter of these papers focus on firm pricing strategies that
depend upon position; Athey and Ellison (2009) focus on optimal auction design in this context. Our chapter instead focuses on the incentives for the ad server to innovate and to change the structure of the market, as discussed above.

Unlike these papers, we allow consumers to have different valuations for the product, conditional upon finding it relevant. This allows us to consider possible selection effects that impact the profitability of the pool of consumers that visit each site. For example, consumers with high valuations may quickly find products priced below their valuation, allowing them to make a purchase and to leave the market, while low valuation consumers continue onward. Alternatively, high valuation consumers have the most to gain from finding a relevant product and may continue searching longer than low-valuation consumers. These patterns may influence the profitability of being in a particular slot in the ad listing, an issue that has not been studied in the literature.

This examination of consumer behavior and bidding strategies serves as a foundation for future applications. Very little is known about competition in ad serving and the ability of contextual advertising revenue to adequately fund online content. Innovation is crucial in this relatively young industry and we must study the returns to innovation accruing to ad servers to determine the adequacy of technological advancement in this area. Additionally, if the ad server has the ability and incentive to influence competition in the product markets themselves, this can raise important antitrust concerns. The model and results of this chapter can serve as a useful framework to examine issues in innovation, antitrust, and business strategy.

1.2 Model

The first step in analyzing the optimal bidding strategy of firms and the resulting incentives facing ad servers is to formulate a model of consumer responses to ad listings. We develop such a model in this section.

1.2.1 Framework

We begin by creating a model of consumer behavior in perusing the advertising listing and in purchasing a product offered by one of the firms. There is a unit mass of consumers, indexed by \(i \in [0, 1]\). The firms are indexed \(j \in \{1, \ldots, J\}\). Consumers view a listing of \(M\) advertisements and consider each product being advertised sequentially starting at the top of the list.

Consumers have a sort of lexicographic preferences that are firm-specific. A product is either relevant for consumer \(i\), yielding a positive valuation for that product \(v_i\), distributed with cumulative distribution function \(F\), or it fails to meet his needs and the consumer has

\[\footnote{In a later section, we consider whether this search pattern is an equilibrium response to firm decisions.} \]
Chapter 1. Ad Server and Firm Strategies in Contextual Advertising Auctions

no value for it at all. The needs of each consumer are met stochastically with probability \( q_j \) by firm \( j \). That is, the valuation of the product offered by firm \( j \) to consumer \( i \) is

\[
v_{ij} = \begin{cases} 
  v_i & \text{with probability } q_j \\
  0 & \text{with probability } 1 - q_j.
\end{cases}
\]

We refer to the probability of firm \( j \) offering a relevant product to a consumer as the *relevance* of firm \( j \) \( q_j \). While the relevance varies across firms, it is the same for all consumers facing a given firm. This model of preferences differentiates the firms and provides a rationale for multiple firms to each to receive a positive market share. Product differentiation can also be helpful for avoiding the Diamond (1971) and Bertrand paradoxes in which firms respond to consumer search with perfectly monopolistic or competitive pricing strategies, thereby eliminating the need for consumers to search at all (for a similar justification for product differentiation in search models, see Anderson and Renault, 1999).

A consumer searches by visiting site \( j \), which is placed in slot \( m \) of the ad listing, then determines whether that product is relevant to him. If so, he compares the price of the product \( p_j \) to his valuation. If the price is below his valuation, he makes the purchase and the search ends. If the product’s price exceeds his valuation or it does not meet his needs, he continues searching with probability \( s_m \). Note that \( s_m \) is a conditional probability: given that a consumer visited site \( m \) and did not make a purchase, \( s_m \) gives the probability that he continues searching. For completeness, define \( s_0 \) as the probability that a consumer visiting a site displaying the ads looks at the ad listing at all. These search probabilities are exogenous and independent of the valuation held by the consumer.

Let us summarize the quantities that define the structure of the model. First, we have the consumer-specific valuations \( v_i \) for consumer \( i \). Next, we have firm-specific costs \( c_j \), prices \( p_j \), and relevances to consumers \( q_j \) for firm \( j \). Lastly, we have slot-specific search continuation probabilities \( s_m \) for slot \( m \). All these quantities are exogenous. In the case of prices, we begin by considering the case in which all firms charge the same price \( p \). Here, prices are set outside the model and there is a single price that prevails in the market.\(^3\)

\(^2\)In this model, consumers do not search for the best deal; instead, they simply find a product that meets their needs at a sufficiently low price and make the purchase. If they do not find such a product, some fraction continue searching. Unlike Athey and Ellison (2009), the probability of searching forward is exogenous in this model.

\(^3\)One way to understand this assumption is that firms charge the same price to consumers that visit their sites via the advertising list, to consumers that visit the sites directly without any search aids, and perhaps even to consumers that find the firms’ products in brick-and-mortar retailers. Other authors have endogenized price into models of search, assuming that firms price discriminate based upon the particular ordering of a particular listing (see Chen and He 2006, Arbatskaya 2007, Armstrong, Vickers and Zhou 2009, Xu, Chen and Whinston 2011a,b). These works do not discuss whether such discrimination can be maintained; for example, consumers could find a relevant product via searching, but then visit the site directly by typing in its address manually to find a better deal.
Distinguishing between firms and slots at first appears cumbersome. Suppose that the index of firm $j \in \{1, \ldots, J\}$ corresponds to the rank of $q_j(p - c_j)$ in decreasing order; that is, firms are indexed in decreasing order of their full expected margins. We show in Section 1.3.1 that, in the equilibrium of interest in this chapter, there is a direct mapping from this ordering to the slots in which the firms appear: the first $M$ firms according to this ranking appear on the list in the slots that corresponds to their ranks, while the remaining firms do not appear on the list. Then, rather than cite “firm $j$ in slot $m$,” we can simply refer to firm $j$, as the slot $m$ must equal rank $j$ in equilibrium.

1.2.2 Product market outcomes

An important quantity of interest in online advertising is the click-through rate (CTR) for the ad in slot $j$, defined as the probability of a consumer clicking on slot $j$’s ad. Recall that firm $j$ appears in slot $j$ in equilibrium. The CTR $r_j$ for firm $j$ is the probability that the consumer enters the list, does not purchase from firms 1 to $j - 1$, and continues on to site $j$.

In the case of the first site in the ad listing, the CTR is simply the proportion of the unit mass of consumers that search the list at all:

$$r_1 = s_0.$$

Next, consider the firm in the second slot. A consumer arrives at this site because either

- The product of firm 1 did not meet his needs or
- Though the product of firm 1 did meet his needs, it was too expensive (i.e., the market price is above his valuation)

and he decided to continue searching. Since prices are the same across firms, the consumers in the group described by the second bullet above never makes a purchase. The CTR is

$$r_2 = s_0s_1(1 - q_1) + s_0s_1q_1F(p) = s_0s_1[(1 - q_1) + F(p)q_1].$$

4We assume that there are no ties.

5It may seem odd that the model permits these consumers to fruitless search forward. This is to enable us to apply this model to a market with varying, perhaps endogenous, prices. We could consider the opposite scenario whereby low value consumers realize, after visiting the first site, that they can never find a product cheap enough for them to purchase and quit searching immediately. This is attrition by low value consumers, the opposite of the attrition by high value consumers described here. In Appendix 1.A we show that the framework developed in this section, notably the use of adjustment factors, is sufficiently broad that it can encompass attrition by low value consumers as well. In fact, with appropriate definition of the adjustment factor, the qualitative results in the chapter do not change with this alternate conception of consumer behavior.
We can generalize this expression to site \( j \leq M \):

\[
    r_j = \prod_{k=0}^{j-1} s_k \prod_{k=1}^{j-1} (1 - q_k) + F(p) \left[ 1 - \prod_{k=1}^{j-1} (1 - q_k) \right].
\]  

(1.1)

As there are only \( M \) sites listed, the CTR for firms \( M + 1, \ldots, J \) is 0. The two sets of terms inside the brackets above describe two groups of consumers: those that have yet to find a relevant product when reaching site \( j \) and those that had found a relevant product at some firm prior to firm \( j \), but that have low valuations. Both are multiplied by a product of search frequencies to account for consumers that end their search without making a purchase.

The CTR is decreasing in list rank. This occurs for two reasons. One, some consumers find a suitable product, make a purchase, and quit searching \((q_k > 0)\). Two, only a fraction of consumers continue searching down the list \((s_k \leq 1)\). A falling CTR is a well-known feature of ad listings and it is important that our model reflect this important empirical reality.

The CTR measures the size of the market that a firm faces. Now, consider the overall demand that each firm receives. In our model, consumers that purchase from firm \( j \) entered the list, were not satisfied by any of the previous \( j - 1 \) firms, searched all the way to firm \( j \), found a relevant product at firm \( j \), and have a valuation above the price. Putting these pieces together, the demand for firm 1 is

\[
    D_1(p) = s_0 q_1 [1 - F(p)]
\]

and the demand for firm \( 1 < j \leq M \) is

\[
    D_j(p) = [1 - F(p)] s_0 q_j \prod_{k=1}^{j-1} s_k (1 - q_k).
\]

A firm not on the list, \( j > M \), does not face any demand. This is the gross sales that a firm makes and it varies across firms.

We see that demand is falling as we move down the list for the same reasons that the CTR was decreasing: some consumers are satisfied by previous firms and some consumers stop searching altogether. This naturally leads us to ask whether the ratio of overall demand to the CTR—the demand per click—is decreasing as well. As we see below, demand per click is also important in determining the value of a click to a firm, which is used in Section 1.3.
to calculate the optimal bid in the placement auction. The demand per click is

\[
D_j(p) = \frac{[1 - F(p)]s_0 \prod_{k=1}^{j-1} s_k (1 - q_k) q_j}{\prod_{k=0}^{j-1} s_k \left[ \prod_{k=1}^{j-1} (1 - q_k) + F(p) \left[ 1 - \prod_{k=1}^{j-1} (1 - q_k) \right] \right]}
\]

\[= a_j q_j,\]

where

\[
a_j = \frac{[1 - F(p)] \prod_{k=1}^{j-1} (1 - q_k)}{\prod_{k=1}^{j-1} (1 - q_k) + F(p) \left[ 1 - \prod_{k=1}^{j-1} (1 - q_k) \right]}.\]  

The value \(a_j \in (0, 1]\) is a slot-specific adjustment factor. Note that \(a_j = 1\) for all slots if \(F(p) = 0\); that is, if there are no low-value consumers. Otherwise, the adjustment factor is decreasing down the list.

The adjustment factor accounts for two features of the model. First, while firm \(j\) offers a product relevant to \(q_j\) consumers, a portion of these consumers have valuations lower than the price. Second, the distribution of valuations changes by the slot: consumers with valuations above the market price make purchases and quit searching, while low-valuation consumers continue searching. Thus, the profitability of the pool of consumers visiting each site decreases down the list. This implies that, at a fixed price, firms further down the list make fewer sales than they would expect based upon their relevances alone. From the perspective of the firm, too many consumers (i.e., the low valuation ones) continue searching. This is called attrition by high value consumers.

The adjustment factor accounts for the changing profitability of the pool of consumers visiting the site in a particular slot. This framework is general enough such that adjustment factors can be calculated in other contexts when profitability is changing down the list. Rather than permit low value consumers to continue searching knowing that they can never find a product with a price low enough for them to make a purchase, we could have these consumers leave the market entirely after visiting the first listing. Appropriately defined adjustment factors can be used in this situation without changing the overall structure of the model or the auction in the following section (see Appendix 1.A). As another example, it is plausible that high value consumers continue searching longer than low value consumers, increasing the profitability of slots lower on the list. Again, adjustment factors could be defined to capture this behavior.

As detailed in the next section, firms bid for slots in the ad listing, paying the ad server for each click received. This bid is going to depend upon the value of a click to the firm. We need to normalize the total profit received by the number of clicks received to
calculate the expected value per click:

\[
\frac{(p - c_j) D_j(p)}{r_j} = m_j a_j q_j,
\]

(1.3)

where \( m_j \) is the margin for firm \( j \).

Indexing the CTR and adjustment factor according to firm \( j \) obscures the calculation of these values. Looking to Equations 1.1 and 1.2, we see that \( r_j \) and \( a_j \) do not depend upon the firm \( j \)'s own relevance \( q_j \), but rather, they depend upon the relevances of the preceding firms. It is more accurate to say that the CTR and adjustment factor are specific to slot \( j \), regardless of the firm that takes that position, given that firms \( 1, \ldots, j - 1 \) remain the same.

The prime reason for defining the adjustment factor is to break the expected value per click into a portion that is the same for the firm no matter where it is listed (\( m_j q_j \)) and a portion that depends upon its actual slot \( a_j \). As mentioned earlier, the indexing of firms corresponds to their equilibrium slot placement, but proof of this fact in Section 1.3.1 requires a distinction between firm-specific and slot-specific quantities. If firm \( j \) is placed into slot \( k \) out of equilibrium, then the expected value per click would be \( m_j a_k q_j \) and it would receive a CTR of \( r_k \).

### 1.2.3 Discussion

Previous work in the ad auction literature assumes that the value that a firm places on being at a particular ranking can be separated into a CTR effect and a firm-specific value effect. CTRs are assumed to decrease monotonically down a list, but a firm has the same value per click of being in any slot. Though a lower-ranked firm may receive fewer clicks, each click has the same value to that firm whether the firm was in the first slot or the last. In these models, consumers are identical and thus there can be no selection in the group that continues searching. If there is attrition by high-valued consumers, this structure is called into question.

One paper that does incorporate heterogeneous valuations is [Chen and He (2006)](#). Their framework combines consumers with differing valuations, but identical search costs that increase with the number of sites visited and endogenize pricing decisions by firms. They do not consider the potential for selection effects in the distribution of consumer valuations down the list. When [Chen and He (2006)](#) consider the firms’ pricing decisions in their Equation 1, they assert that all firms face the same pricing decision, yielding no price dispersion, but they do not consider that firms may face different demand conditions depending upon their ranks and, as a result, the firms’ maximization decisions will vary. In particular, firms further down the list face fewer high-value consumers and may be inclined to cut prices due to attrition.

---

6 Adding to the complexity, the out of equilibrium values for the adjustment factor \( a_k \) and CTR \( r_k \) would be different from their equilibrium values if one or more preceding firms, \( 1, \ldots, j - 1 \) also changed position.
of high-value consumers. Our base model does not endogenize pricing decisions, hence, we do not evaluate this strategy here.

1.3 Ad Auction Bidding Behavior

Contextual ads are sold using a Generalized Second Price (GSP) auction. A firm places a bid to be included in the ad listing based upon keywords that appear in the substantive content (search queries, articles, reviews, etc) on the page. In the simpler case developed by Overture for Yahoo, firms are assigned slots in decreasing order of their bids. A firm pays the bid of the next ranked firm each time that its own ad is clicked. Many prominent papers have focused on this framework (see, e.g., Edelman, Ostrovsky and Schwartz 2007; Varian 2007; Athey and Ellison 2009).

In Google’s auction, firms are ranked by the product of their bids and their “quality scores” and a firm pays the product of the bid and quality score of the next firm down the list on a per-click basis. Quality scores aim to estimate the expected CTR for a firm’s ad. For example, for the keyword “airplane,” suppose that both Boeing and a toy airplane manufacturer would like to have their ads listed. Boeing may be willing to pay more for a listing because, if a click turns into a sale, the firm earns greater profit relative to the profit earned on the sale of a toy plane. Few viewers are interested in purchasing jumbo jets, however, so the Boeing ad receives few clicks, earning Google little revenue. Google could earn greater revenues by putting a firm with a lower bid but higher firm-specific CTR at the top of the list than a firm that bids high, but receives few clicks.

Advertisers can change their bids frequently; this might lead us to model the auction as an infinitely repeated game. The Folk Theorem, however, asserts that these games have many equilibria, rendering analysis extremely difficult. Resultingly, most work has focused on the single-shot version of the auction game to identify an equilibrium.

An early effort in the literature, Edelman, Ostrovsky and Schwartz (2007) places the GSP auction in the context of established auction designs, including the second price auction, Vickery-Clarke-Groves (VCG) mechanism, and the ascending English auction. They show that the GSP auction is not equivalent to the VCG mechanism. Unlike VCG, this auction does not have an equilibrium in dominant strategies and truth-telling is not an equilibrium. Under a set of restrictions, one of the equilibria that arises provides the same payoffs as under the dominant strategy VCG equilibrium. Edelman, Ostrovsky and Schwartz (2007) call these equilibria “locally envy-free equilibria.” Varian (2007) independently identifies the same equilibria and calls them “symmetric Nash equilibria.” The ad intermediary is better off at any other locally envy free equilibrium other than the one equivalent to the VCG equilibrium, while advertisers are worse off.

The quality score of a firm is no longer this transparent. Many papers consider the optimal weights to maximize auction revenue, but we do not consider this literature here.
Most of the existing literature on advertising auctions has focused on the elements of optimal auction design. Alternative mechanisms have been offered that provide higher profits to ad intermediaries or more efficient assignments of ad slots. Other papers extend the standard GSP framework by incorporating the quality scores found in Google auctions or other weighting schemes and reserve prices. This chapter focuses on the properties of the standard auction mechanism, but incorporates the consumer behavior underpinning click-through rates. While the structure of the auction is undoubtedly important for firms and the ad server, we ignore these complexities and use the simplified version of the auction developed by Yahoo/Overture in our analysis.

1.3.1 Ranking of firms in the ad listing

We begin by incorporating our model for CTR into the approach of Varian (2007), specifically, a one-shot, simultaneous move, complete information game. Of the $J$ firms in the market, $M$ appear on the ad list. The CTR for firms $M + 1, \ldots, J$ is 0, while a firm on the list in slot $j$ experiences a CTR $r_j$ following Equation 1.1. Varian (2007) assumes that the CTR is exogenous and decreasing down the list; in the preceding section, we provide a behavioral foundation for this assumption.

A firm is charged on a per-click basis at a price equal to the bid of the firm one slot down on the ad list. Recall from Equation 1.3 that the expected value per click for firm $j$ in slot $j$ is $m_j a_j q_j$. And, based upon the discussion on page 8, the expected value per click of firm $j$ in the off-equilibrium slot $k$ is $m_j a_k q_k$ with a CTR of $r_k$. In the symmetric Nash equilibria of Varian (2007), the expected profits in firm $j$’s equilibrium slot must be weakly higher than those it receives in any other slot $k$:

$$r_j (m_j a_j q_j - b_{j+1}) \geq r_k (m_j a_k q_k - b_{k+1}).$$  \hspace{1cm} (1.4)

Note that the CTR and the slot-specific adjustment factor change with the slot for a given firm, but the relevance of the firm and its margin do not. The firm faces the following trade-

---

8We do not consider the case of “unsold pages,” where there are fewer willing bidders than slots. Additionally, we assume that the highest $M + 1$ firms all bid above the reserve price of the auction.

9Bear in mind that firms lower on the list have higher indices—firm $j$ is one slot above firm $j + 1$.

10The CTR for slot $k$ depends upon the relevances of the firms $1, \ldots, k - 1$. As a result, the CTR for slot $k$ is different when different firms are in the preceding slots. If firm $j$ moves up to slot $k$, then the ordering of firms $1, \ldots, k - 1$, and thus $r_k$ and $a_k$ remain unchanged. If firm $j$ moves down the list to slot $k$, this changes the firms that appear in slots $1, \ldots, k - 1$. If firm $j$ moves down the list, then $r_k$ would be an out-of-equilibrium CTR. We will claim that, in equilibrium, higher-ranked firms must have weakly higher relevances. If a relatively high relevance firm $j$ moves down the list to slot $k$, fewer consumers find a relevant product from firms $1, \ldots, k - 1$. This implies that more consumers search forward and these out-of-equilibrium CTRs for slots $j, \ldots, k$ are higher than their equilibrium values. The following equation fails to distinguish between in- and out-of-equilibrium CTRs. Since the latter are weakly higher, this inequality remains valid.
off: Accepting a lower slot on the page requires a smaller payment for the slot. However, the firm receives fewer clicks in this space and faces a less profitable pool of consumers.

Consider the equilibrium conditions for firm \( j \) moving to slot \( k \) and firm \( k \) moving to slot \( j \):

\[
m_j q_j (a_j r_j - a_k r_k) \geq r_j b_{j+1} - r_k b_{k+1} \quad \text{and} \quad -m_k q_k (a_j r_j - a_k r_k) \geq -r_j b_{j+1} + r_k b_{k+1}.
\]

Adding these inequalities together gives

\[
(m_j q_j - m_k q_k)(a_j r_j - a_k r_k) \geq 0. \tag{1.5}
\]

Recall from Section 1.2.2 that the CTR \( r \) and the adjustment factor \( a \) are both decreasing down the list. This expression reveals that the relevance \( q \) times the margin \( m \) must move in the same direction, namely, decreasing down the list; the full expected margin is decreasing down the list. If there are no ties in the full expected margin \( m_j q_j \), ensuring that this ranking is unique.

Varying margins, constant relevance

An interesting special case is when \( q_j = q \) for all \( j \). Here, firms sort in decreasing order of margins. All firms charge the same price \( p \) and have the same relevance; consumers are indifferent to the order of firms that they search. In the case of indifference, assume that consumers still search from the top down. While the ordering of the firms has no impact on consumer surplus, producer surplus is largest when firms sort in increasing order of costs—that is, decreasing order of margin. This is precisely the result given by the auction, hence, total surplus is maximized.

Varying relevances, constant margins

At the other extreme, suppose that firms all have the same costs and thus the same margin, but have different relevances. The equilibrium condition reveals that the firms sort in decreasing order of relevance. Consumers prefer to visit the sites most likely to offer a relevant product. Given the bidding strategies of the firms, this would imply that consumers should search starting from the top of the list, confirming this outcome as an equilibrium. Since consumers visit a limited number of sites in order, the greatest number of sales occur when the most relevant firms are listed at the top; this ranking also maximizes both consumer and producer surpluses.
Varying margins and relevances

Of course, the intermediate cases are perhaps the most interesting and most difficult to characterize. Considering the expected ordering of firms, we ask how a firm’s cost is correlated with its relevance. If these factors are negatively correlated, then low cost firms have high relevances and the intuition developed in the two preceding subsections do not conflict; we expect the low cost, high relevance firms to be at the top and the high cost, low relevance firms to be at the bottom.

We can go further by considering the case that the cost of firm $j$ is $c + \alpha q_j$. We could impart a causal story: it is more or less costly (depending upon the sign of $\alpha$) to produce a product that a high proportion of people like. Or we could consider the model as one of association, used only to highlight existing correlations between relevance and cost. Our equilibrium condition becomes

$$[(p - c)(q_j - q_k) - \alpha (q_j^2 - q_k^2)] (a_j r_j - a_k r_k) \geq 0.$$ 

Section 1.2.2 demonstrates that, in this equilibrium, both $r$ and $a$ are decreasing down the list, giving $a_j r_j - a_k r_k$ a positive sign. If the CTR is falling down the list, as it does when consumers search from top to bottom, then firms sort in decreasing order of relevance if

$$\frac{p - c}{q_j + q_k} \geq \alpha$$

(1.6)

and sort in increasing order of relevance otherwise. Note that the lefthand side of this expression is positive. The CTR is only decreasing down the list if consumers have an incentive to search downward; this is the case if relevance is weakly decreasing down the list. Hence, if $\alpha$ satisfies Equation (1.6), then this equilibrium exists. Intuitively, this condition states that the relevance and cost of a firm can covary positively, to a point, and still sort in decreasing order of relevance. Firms with smaller per-sale margins have higher expected margins due to their higher relevance.

1.3.2 Deriving equilibrium bids

To this point, we have considered how firms must order themselves relative to their costs and relevances in our equilibrium. Next, we find the bids that are consistent with this equilibrium. To find the bids chosen by the firms, we return to Condition 1.4. Varian (2007) shows that, if this condition holds for a firm moving up one slot or down one slot (i.e., from $j$ to $j - 1$ or to $j + 1$), then it holds for a move to any slot or a move off the list entirely. Using the fact that firm $j$ does not want to move to slot $j + 1$ and that firm $j + 1$ does not
want to move to slot \( j \), we find that
\[
m_j q_j \left( a_{j-1} - \frac{r_j}{r_{j-1}} a_j \right) + \frac{r_j}{r_{j-1}} b_{j+1} \leq b_j \leq m_j q_{j-1} \left( a_{j-1} - \frac{r_j}{r_{j-1}} a_j \right) + \frac{r_j}{r_{j-1}} b_{j+1}. \tag{1.7}
\]
These bounds can be solved recursively by recalling that \( r_j = 0 \) for the firms not listed, firms \( j > M \), yielding
\[
\frac{1}{r_{j-1}} \sum_{j+1 \leq k \leq M+1} m_k q_k (a_{k-1} r_{k-1} - a_k r_k) \leq b_j \leq \frac{1}{r_{j-1}} \sum_{j+1 \leq k \leq M+1} m_{k-1} q_{k-1} (a_{k-1} r_{k-1} - a_k r_k). \tag{1.8}
\]
Firm \( j \) can bid any value in this range without changing its slot or the slot of other firms.

We can learn many things from these recursive bounds. First, firms have positive profits. To see this, return to Condition 1.4 and set \( k = M + 1 \), the first firm not listed. Here, \( r_k = 0 \), implying that \( m_j a_j q_j \geq b_{j+1} \). Hence, the net profit from being in slot \( j \), \( r_j (m_j a_j q_j - b_{j+1}) \geq 0 \). The lower bound of Equation 1.8 is less than or equal to the expected margin from slot \( j \), \( m_j a_j q_j \)—firms may shade their bids. It is possible that the upper bound of Equation 1.8 is above the expected margin, implying that a firm may bid above its valuation. The logic here is that the firm in slot \( j \) must bid high enough so that the firm just above it in slot \( j - 1 \) does not have an incentive to switch to slot \( j \) and sacrifice clicks to increase per-click profit. Nonetheless, the expected margin from the slot must be positive; remember, the firm does not pay its own bid, but rather the bid of the firm below it in the listing.

We can rearrange this equation to get the ad revenue raised from firm \( j \):
\[
\sum_{j+1 \leq k \leq M+1} m_k q_k (a_{k-1} r_{k-1} - a_k r_k) \leq r_j b_{j+1} \leq \sum_{j+1 \leq k \leq M+1} m_{k-1} q_{k-1} (a_{k-1} r_{k-1} - a_k r_k).
\]
Total ad revenue has a lower bound of
\[
\sum_{2 \leq k \leq M+1} (k - 1) m_k q_k (a_{k-1} r_{k-1} - a_k r_k)
\]
and an upper bound of
\[
\sum_{2 \leq k \leq M+1} (k - 1) m_{k-1} q_{k-1} (a_{k-1} r_{k-1} - a_k r_k).
\]

The pattern of bid shading is not obvious from Equation 1.8. We calculate the upper and lower bounds for the bid of the firm in each slot for a particular set of parameter values. For concreteness, let each firm have the same relevance \( q = 0.2 \), with the proportion of low-

\[\text{This procedure actually gives the bounds for } b_{j+1}; \text{ appropriate reindexing gives the result shown.} \]
value consumers set to $F(p) = 0$ and all search frequencies set to $s = 1$. The margins of the firms vary from 0.9 in slot 2 to 0.1 in slot 10 in increments of 0.1.

Figure 1.1 plots the ratio of the bid of firm $j$ to the expected margin $m_j a_j q_j = m_j q_j$ in this case (as there are no low-value consumers). The pattern in the shading of the bids is not monotone; shading is highest for firms in the middle of the listing, higher for top-ranked firms, and non-existent for the final firm (as was obvious from Equation 1.8). The range of possible bids gets larger moving down the list.

Figure 1.1: Ratio of bid to expected margin

Of course, firms do not pay their own bids; they pay the bids of the next firm down the list. We may wonder how the ratio of the cost of being in slot $j$, which is the bid of firm $j + 1$, to the expected margin of firm $j$ varies down the list. Using the same parameter values as Figure 1.1, Figure 1.2 shows these results. Again, the pattern here is not monotone. Instead, this ratio falls through slot 7, then increases sharply. For both figures, the pattern depends upon the chosen parameter values, but these illustrations reveal that overall patterns are difficult to characterize.

1.3.3 Discussion

Varian (2007) arrives at these results by assuming complete information. He offers several justifications for this assumption. First, Google reports view and click rates on an hourly basis to bidders and, if bidders experiment with different bidding strategies, they can infer
Figure 1.2: Ratio of cost per click to the expected value per click

many of these quantities fairly quickly. Additionally, Google offers a “Traffic Estimator” that predicts the number of clicks and total costs for different bid-keyword combinations. Lastly, private, experienced search engine optimizers can offer clients assistance with bidding strategies.

Athey and Ellison (2009) do not assume perfect information in their model. Instead, they choose to study an incomplete information game. Specifically, each firm $j$’s relevance, $q_j$, is private information, while all margins $m_j = 1$. Additionally, rather than consider a simultaneous-move game, they consider a multistage game where each slot is sold separately from bottom to top. This change lets the authors condition on bids for lower slots in solving for higher bids; we sense why this is important by recalling the equilibrium conditions of the complete information model given by Equation 1.8.

Most papers in the position auction literature take the CTRs to be separate, exogenous parameters. Here and in Athey and Ellison (2009), the CTRs are derived using behavioral assumptions about consumers and are functions of the firms placed above a given slot, i.e., the CTR for firm $j$ depends upon the relevances of firms 1 through $j - 1$. This implies that the value that a firm obtains from being in a given slot depends upon those

\footnote{Varian (2007) offers a version of his model in an incomplete information setting as an appendix to his paper. In his case, however, the CTRs remain common information and exogenous, while Athey and Ellison (2009) turn to an incomplete information context precisely because the CTRs are functions of the unknown relevances.}
ranked above it. As Athey and Ellison (2009) note, this fact implies that the auction is no longer based upon private values.\footnote{Additionally, the GSP is no longer the most efficient mechanism for ranking ads (Aggarwal et al., 2008). We do not look for the optimal auction design in this chapter.} The common values aspect is easily separated from the private value of appearing in the list and this wrinkle is not difficult to handle.

The authors use a perfect Bayesian equilibrium solution concept to solve the game. Using this procedure, the authors provide a bid function that is qualitatively the same as that of Varian (2007) and thus the behavioral model of our chapter could easily be placed within their framework without fundamentally altering the results that follow.

### 1.4 Incentives of ad servers to change search structure

There are three quantities that define the consumer side of the market: the tastes of consumers given by the match probabilities \( q \), search costs that are implicit in the search frequencies \( s \), and the valuations for the product \( v \). A fourth quantity in the model that describes the supply side is the per-sale margin \( m \). These factors are not immutable, however; altering these quantities may change the profits and overall welfare generated in the product market. In this section, we explore the incentives than an ad server has to alter these quantities and how these changes impact firms and consumers.

The prime incentive for the ad server to produce such changes stems from changes in advertising revenue. Equation (1.8) reveals that the lower bound for the advertising revenue generated after a change in the market structure would be

\[
\sum_{2 \leq k \leq M+1} (k-1)[m_kq_k + \Delta(m_kq_k)][(a_{k-1}r_{k-1} - a_k r_k) + \Delta(a_{k-1}r_{k-1} - a_k r_k)].
\]

The change in revenue, then, is given by

\[
\sum_{2 \leq k \leq M+1} (k-1)[\Delta(m_kq_k)[(a_{k-1}r_{k-1} - a_k r_k) + \Delta(a_{k-1}r_{k-1} - a_k r_k)] + m_kq_k\Delta(a_{k-1}r_{k-1} - a_k r_k)];
\]

the increase in the margin times a number that is a function of the new CTRs and adjustment factors plus the old margin times the change in the CTRs and adjustment factors. The first piece captures the increased value-per-click to a firm after the change. The latter captures whether the number of clicks has changed.

We find \( a_{k-1}r_{k-1} - a_k r_k \) by noting that, by definition, \( a_k r_k = \frac{D_k(p)}{q_k} \); an analogous
result is found for firm $k - 1$. The difference between these two quantities is

$$[1 - F(p)]s_0 \prod_{p=1}^{k-2} s_p (1 - q_p) [1 - s_{k-1}(1 - q_{k-1})]. \quad (1.10)$$

To find $\Delta(a_{k-1}r_{k-1} - a_k r_k)$, we can analyze changes in Expression [1.10].

Of course, these expressions are difficult to characterize for general changes in the relevances across firms. Instead, we look at comparative statics for proportional changes in the deep parameters of the model, changes that impact each firm by the same proportion. This permits us to develop intuition by simplifying the expressions above and by permitting us to conduct concrete simulations of the policy changes.

1.4.1 Proportional changes in the relevances

Suppose that the ad server has the ability to boost all firms’ relevance by a certain percentage. This could occur by achieving a better matching algorithm, by using information known about a particular user, or, rather than increasing the relevances of given firms, by having bigger pool of advertisers, thereby yielding more high quality matches.

Intuition from the model

Consider this change in the context of Expression [1.9]. Since $q_k$ goes up, $\Delta(m_k q_k)$ is positive and the first component of the sum is positive. For expository purposes, let all firms have the same relevance $q$. Then, Expression [1.10] becomes

$$[1 - F(p)](1 - q)^{k-2} \prod_{p=0}^{k-2} s_p [1 - s_{k-1}(1 - q)].$$

Taking the derivative with respect to $q$ yields

$$[1 - F(p)](k - 2)(1 - q)^{k-3} \prod_{p=0}^{k-2} s_p [s_{k-1}(k - 1)(1 - q) - (k - 2)].$$

A necessary condition for the quantity in question and thus ad revenue generated by a particular slot to be increasing is\footnote{Recall that bids are calculated starting at $k = 2$; division by 0 is not an issue here.}

$$s_{k-1}(1 - q) > \frac{k - 2}{k - 1}.$$
This inequality does not hold in general; it holds only for firms above some threshold value of \( k \) that ensures that this inequality is satisfied.

Firms receive a higher margin per click because a consumer is more likely to find a relevant product on its site. This implies, however, that more consumers are satisfied high on the list and do not visit lower-ranked sites. While the expected margin may be higher, the pool of consumers is smaller. These conditions act as opposing forces in changing the ad revenue generated by a firm. We expect the bids of high-ranked firms to increase more after the change in the relevances compared to lower-ranked firms. However, as bids are solved recursively, drops in the bids of lower-ranked firms temper increases in higher-ranked firms. As firm 1 does not experience any drop in its CTR or adjustment factor, we expect it to exhibit the greatest change in advertising revenue generated.\(^{15}\) Expression 1.10 thought of more simply, is the difference in CTRs between firms \( j-1 \) and \( j \).\(^{16}\) This result states that, after a proportional increase in the relevance of each firm, the change in ad revenue generated by firms high on the list is larger than that for firms low on the list.

**Simulation of the change**

While these calculations give us some intuition for the impact of a change in relevances on ad revenues, let us consider a numerical example. Largely irrelevant to these calculations are the search frequencies \( s_k \) and the proportion of low-value consumers \( F(p) \); set the former all to 1 and the latter to 0 for simplicity. Assume that all firms have the same relevance of 0.2. We consider an increase in this value by 20%.

Figure 1.3 gives the impact of this change on ad revenues, bids, and gross and net (of advertising costs) firm profit. First, we note that, in this case, the CTR drops by a factor of \( \frac{1−1.2×0.2}{1−0.2} \) for site \( k \). After the 20% increase in relevance, firms bid at least 20% more.

The highest increases in bids come from firms in the middle. High ranked firms do not experience a large change in their CTRs. Middle ranked firms have large drops in their CTRs and need to bid higher to avoid slipping down the list and experiencing even greater changes. Firms low on the list had low CTRs anyhow and, while the drop may be relatively larger than for other slots, the absolute drop is smaller and these firms do not have as strong an incentive to bid to avoid it.

Ad revenue is a product of the CTR and the bid. Higher bids more than offset the reduced CTR for firms 1 through 6, increasing the ad revenue generated by these firms. For the last 3, ad revenue decreases. Total ad revenue increased by 21%.

Turning to the perspective of the firms, firm profits increase for the first 4 firms, but fall for the remainder; for the latter group, the higher match probability (and thus expected margin) is offset by fewer clicks. For the high-ranked firms, higher match probabilities lead

---

\(^{15}\) One issue not yet discussed is that, if an ad server can increase the relevance of its ads, then it may attract a larger pool of consumers to its site, increasing the size of the market for all firms.

\(^{16}\) This occurs when \( F(p) = 0 \); everyone is a high-value consumer.
to more sales and more revenue, but these increased sales revenues are offset by the necessity of offering higher bids. Indeed, only the first 2 firms have higher profits net of advertising costs after the increase in relevances. Total firm net profits across all the firms actually fell by 2.2%.

This implies that, while the very top-ranked firms do benefit from the increased probability of a firm-consumer match, the ad server is able to siphon a substantial share of the increase in profits. The ad server is able to take a large enough share of the increase in profits such that the bottom 7 lose profits net of advertising costs and the firms overall experience a reduction in net profits.

Intuitively, when the ad server increases the probability of a match between a consumer, this increases the probability and thus the number of sales; the sales revenue pie grows. Because the ad server created value, the firms are willing to pay for it—the equilibrium condition commands that they do so. But, as the size of the pie grows, the size of the slice garnered by the ad server does not grow by the same proportion. In this case, the slice received by the ad server grows by a larger proportion than the pie overall, leaving less for the firms themselves. This lack of proportionality does not have to be so, as we explore in the next simulation.

For this particular increase in the relevances, the ad server earns higher revenues, while firms’ net revenues fall. This is not necessarily the case. Figure 1.4 show the total ad revenue, ad elasticity, and total firm gross and net profits across changes in the base relevance of 0.2 by factors of 0.5 to 1.5. “Total” refers to measures summed across all firms. By “elasticity,” we mean the proportion change in ad revenues divided by the proportion change in relevance.

Revenues for the ad server and profits for the firms are both increasing with the relevance. The ad revenue elasticity and total firm net revenues have maximum values, however. The ad revenue elasticity is maximized at a proportion increase of 1.2, an increase from 0.2 to 0.24. This is higher than the point where firm net profits are maximized, at a relevance of 0.19. These plots reinforce that firms in aggregate may be hurt by overall increases in match probabilities.

Next, in Figure 1.5, we turn to the consumer side of the market. Since we assume that prices are constant across firms and unchanging with $q$, consumer welfare is higher if a consumer has a better chance of finding a match and thus making a purchase. As a measure of welfare, this implies that consumers are better off if sales are higher. In Figure 1.5a, we see how sales change by firm in the context of the analysis leading to Figure 1.3. We see that the first 4 firms increase their sales, while the remaining firms have lower sales. Sales increase because the match probability increases, but fall for lower-ranked firms because fewer consumers reach these firms without already being satisfied (i.e., the CTR is lower). Overall

---

17 If improving matches has linear cost with no fixed costs, then this would be the elasticity of revenue with respect to costs. This cost function for improving matches by the ad server is highly unlikely, but this calculation provides useful insights nonetheless.
sales are increasing in $q$, as seen in Figure 1.5b, an analogue to the panels of Figure 1.4. Consumers are unambiguously better off if the ad server increases the relevances.
Figure 1.3: Impact of a 20% increase in relevance from $q = 0.2$
Figure 1.4: Impact of changes in relevance from $q = 0.2$ on aggregates
Chapter 1. Ad Server and Firm Strategies in Contextual Advertising Auctions

Figure 1.5: Impact of changes in relevance for consumers

(a) Sales by firm, $q = 0.2$ increased by 20%

(b) Total sales for a range of proportional changes in $q = 0.2$
1.4.2 Proportional changes in search costs

The ad server may also be able to reduce search costs. Practically, this may mean caching pages for faster loading, subsidizing high-speed internet access, or making consumers more proficient searchers. Unlike in the case of increasing relevance, this change does not alter firms’ expected margins. Instead, it just increases the size of the customer base visiting each site. We imagine that such a change should leave both firms and the ad server better off.

Intuition from the model

Again, return to Expression 1.9. The full margin \( m_k q_k \) does not change, leaving only Expression 1.10 to consider. For simplicity, let all the search frequencies be the same, i.e., \( s_j = s \) for all \( j \). This expression becomes

\[
[1 - F(p)]s^{k-1} \prod_{p=1}^{k-2} (1 - q_p)[1 - s(1 - q_{k-1})].
\]

Taking the derivative with respect to \( s \) gives

\[
[1 - F(p)]s^{k-2} \prod_{p=1}^{k-2} (1 - q_p)((k - 1) - ks(1 - q_{k-1})].
\]

Hence, a sufficient condition for the quantity to be increasing in search frequencies is

\[
s(1 - q_k) < \frac{k - 1}{k}.
\]

This inequality only holds for firms low on the list, firms below a threshold value of \( k \) such that this inequality holds for the given relevances of the firms on the list. It is perhaps surprising that ad revenues do not unambiguously increase for all firms. This is because, while the firm receives a higher CTR, the cost in terms of lost sales to moving a slot down the list is less severe (as a higher proportion of consumers now visit this site) and bids are shaded more.

Simulation of the change

Following an analogous analysis to that for altering the relevance, we consider an example with \( q = 0.2 \) and a base search frequency of 0.6, both of which are the same across firms. We consider increasing the search frequency by 20%. The results are given in Figure 1.6.

Bids decrease for all firms except the first excluded firm. Recall that this firm bids its true valuation per click for being included in the list; since this has not changed, neither has
its bid. Reduced bids are more than offset by higher CTRs, as evidenced by the fact that ad revenue from every site increases—by dramatic proportions for many slots. Site 1 has the smallest increase in ad revenue, a change of 14%, smaller than the change in visitors (20%). All other firms increase the ad revenues that they generate by a larger percentage than the change in search frequencies. This is sensible, as changes in search frequency compound and the proportion increase in the size of the consumer group after the change gets larger down the list. Firm net profits increase by a larger percentage than gross profits. Unlike the case of increasing relevance, firms keep a large share of the gains from increasing search frequencies.

We can explore these properties in aggregate across a variety of changes in search frequencies; we show the resulting patterns in Figure 1.7. These plots show that constant increases in search frequencies benefit both the ad server and the firms. Net revenue is the most responsive of all, suggesting that gains in search frequencies mostly benefit the firms. A larger proportion of consumers make purchases when a larger proportion search forward and more purchases leads to higher consumer surplus. We find that all parties benefit from reduced search costs or, more precisely here, higher search frequencies.
Figure 1.6: Impact of a 20% increase in search frequencies from $s = 0.6$
Chapter 1. Ad Server and Firm Strategies in Contextual Advertising Auctions

Figure 1.7: Impact of changes in search frequencies from $s = 0.6$ on aggregates
1.4.3 Proportion increase in high value consumers

The ad server may be able to increase the profitability of the consumers that visit its site. It may be able to target high valuation demographics in a variety of ways, such as providing targeted content or advertising to this select group. We see how changes in $1 - F(p)$, the proportion of high value consumers, change the revenues accruing to the ad server and the firms.

Looking to Equation 1.8 and using the fact that $a_k r_k = \frac{D_k(p)}{q_k}$, we see that $1 - F(p)$ factors out of the sum. Hence, a proportional change in the probability of high value consumers leads to a change of the same proportion in ad revenue. Firm profits is given by $D_k(p)m_k$. Here, too, demand is directly proportional to $1 - F(p)$.

Gross firm profits and ad revenues increase by the same proportion as and thus net profits, too, grows by the same proportion. If a larger fraction of the consumers are high valuation types, a larger proportion make purchases. All parties are improved if the proportion of high value types increase.

Note that a higher fraction of consumers make purchases, reducing the CTRs for lower ranked firms. And yet these firms are better off because the clicks that they do receive are more valuable and attrition by high value consumers is attenuated.

1.4.4 Proportion increase in margins

A final variable to consider is the margins of the firms. The result is quite similar to that found in the preceding subsection. Equation 1.8 clearly shows that a proportional increase in margins (by lower costs; price remains fixed) leads to the same proportion increase in ad revenue. Gross firm revenue increases by the same proportion, implying that net revenue increases by this proportion as well. Firms and the ad server are better off. Consumers are not paying higher prices and the same fraction make purchases as before, so they are indifferent to the change.

1.5 Impact of dispersion of firm characteristics on bids

The lower bound of Expression 1.7 demonstrates that all firms may shade their bids, except for the first excluded firm. The lower bound reveals that the bid is nearly a weighted average of the value of being in slot $j$ to firm $j$ and the bid of firm $j + 1$. If all the other firms’ expected margins are close to that of the first excluded firm and this firm bids its true value of being in the final slot on the list, the magnitude of the shading is likely reduced. We can consider how dispersion in margins and relevances across firms impacts the proportion of firm revenue that the ad server can extract through bid revenue.
1.5.1 Dispersion in margins

First, consider firms that all have the same relevance of 0.2, but have margins that vary. Figure 1.8 considers a range of variances for these margins. Margins are distributed uniformly with mean 0.5 and bounds determined by the standard deviation of the distribution of the margins. The first panel of the figure confirms our conjecture above: the less dispersion in the margins, the higher the share of firm revenue that is transferred to the ad server. As margins become more dispersed, bid shading becomes more extreme and ad revenues fall. The variation in bids relative to the variation in expected margins (here, relevance times the per-sale margin) exhibits no clear pattern and is varies little itself.

Figure 1.8: Impact of variation in per-sale margin on auction bids and revenue

1.5.2 Dispersion in relevances

We might imagine that firms aim to produce products are broadly enjoyed by many consumers. The rise of the internet, however, has created new incentives and opportunities for firms to produce niche products that a relatively small segment of the population loves, while the rest does not (Bar-Isaac, Caruana and Cuñat 2009). In our model, niche products would have lower relevances than broadly appreciated products and the advent of “long tail” niche products can lead to dispersion in the relevances in a market.
We can examine the impact of variation in the relevance of firms with constant per-sale margins of 0.5. Figure 1.9 shows the results of this analysis. Bid shading increases as relevances become more dispersed, just as in the case of dispersion in per-sale margins. The magnitude of this change is much smaller, however (compare the scale of the y axis in Figure 1.8a to that of Figure 1.9a). Also, the dispersion of bids relative to the dispersion of expected margins increases as relevance becomes more dispersed. This reflects the fact that shading responds by only a small amount to changes in dispersion of the relevances.

Figure 1.9: Impact of variation in per-sale margin on auction bids and revenue

1.6 Choosing the optimal number of ads to display

The ad server will choose an optimal number of ads to display. Each additional slot increases the number of clicks in the list, leading to more payments from firms, a benefit of increasing the length of the list. But the bids of all other firms change in response to the addition; indeed, they fall. The ad server chooses the optimal number of slots by balancing these factors.
Chapter 1. Ad Server and Firm Strategies in Contextual Advertising Auctions

Recall that the lower bound of the total ad revenue is

\[ \sum_{2 \leq k \leq M+1} (k-1)m_k q_k (a_{k-1} r_{k-1} - a_k r_k). \]

When the list is \( M \) firms long, \( r_{M+1} = 0 \). Extending the list to \( M+1 \) firms, of course, results in a positive \( r_{M+1} \). Plus, a new term is added to this sum equal to \((M+1)m_{M+2}q_{M+2}a_{M+1}r_{M+1}\), which uses the fact that the CTR for firm \( M+2 \) is 0. Hence, the change in total revenue with the addition of a new firm is

\[ -Mm_{M+1}q_{M+1}a_{M+1}r_{M+1} + (M+1)m_{M+2}q_{M+2}a_{M+1}r_{M+1}, \]

which is negative \( M \) times the profit for firm \( M+1 \) when it is included on the list plus \( M+1 \) times the profit for firm \( M+2 \) if it was included in slot \( M+1 \) instead. This change is positive if

\[ \frac{m_{M+2}q_{M+2}}{m_{M+1}q_{M+1}} \geq \frac{M}{M+1}. \]

One way to think about this condition is that the ad server chooses a set of firms so that the dispersion in total expected margins is relatively small. This accords with the findings of Section 1.5—dispersion is bad for ad server revenues.

Interestingly, none of the changes that we consider in Section 1.4 change the optimal length of the list. The length of the list does not change with proportional changes in relevances or margins and the length does not depend at all upon the proportion of low-value consumers or search costs. Hence, we might expect the length of ad listings to be relatively stable despite changes in the structure of the market. In Section 1.4, we considered 10 firms, each with a relevance of 0.2 and margins ranging from 0.1 to 1.0 in increments of 0.1. With these parameters, the optimal list length is not 9, as we considered, but instead would be 5.

1.7 Subsidizing the bid of a firm

To this point, we have considered the firms selling the advertised product to have separate interests from the ad server. Instead, suppose that the ad server also has a division that sells the product being advertised. For example, Google displays ads, but it also offers e-mail, mapping, and music services. One complaint lodged against Google is that it artificially boosts its own products to the top of advertising lists. In this section, we consider the strength of the incentives for the ad server to privilege its own products in the listing.

The obvious benefit to the ad server of its product being placed at the top of the list is that it gets additional sales. Let the ad server’s product be firm \( j \), following the labeling
established throughout this chapter. This firm would be in slot \( j \) earning a profit of
\[
m_jD_j(p) = m_j[1 - F(p)]s_0q_j \prod_{k=1}^{j-1} s_k(1 - q_k).
\]

If, instead, firm \( j \) was placed into the first slot, it would receive a higher profit equal to
\[
m_jD_1(p) = m_j[1 - F(p)]s_0q_j.
\]
The difference is
\[
m_j[1 - F(p)]s_0q_j \left[ 1 - \prod_{k=1}^{j-1} s_k(1 - q_k) \right].
\]
This is increasing in the number of slots the firm is placed down the list, and, holding its slot fixed, the margin \( m_j \) and relevance \( q_j \). It is also increasing in the relevances of the preceding firms. Boosting the firm to the top of the list generates more profit from sales of the product.

The cost of this maneuver is the reduction in bids by firms further down the list. The lower bound for ad revenue generated by the listing generated without privileging the ad server’s firm is
\[
\sum_{2 \leq k \leq j} (k - 1)m_kq_k(a_{k-1}r_{k-1} - a_kr_k) + \sum_{j+1 \leq k \leq M+1} (k - 2)m_kq_k(a_{k-1}r_{k-1} - a_kr_k).
\]
This formulation does not include ad revenue stemming from the ad server’s own listing, as we did not count this as a cost above (this explains the change to the multiple \( k - 2 \) when summing the ad revenue generated by the bottom \( j \) firms). If the ad server’s firm moved to the top of the list, the bids of the firms \( j + 1 \) to \( M + 1 \) would be unchanged as the firms preceding each of these would remain the same, though in a different order. Only the bids and thus the revenue generated by firms 1 to \( j - 1 \) would change. Total revenue after firm \( j \)’s move to the top is
\[
\sum_{2 \leq k \leq j} (k - 2)m_kq_k(a_k^*r_k^* - a_{k+1}^*r_k^*) + \sum_{j+1 \leq k \leq M+1} (k - 2)m_kq_k(a_{k-1}r_{k-1} - a_kr_k),
\]
where the asterisks indicate the out-of-equilibrium (relative to the auction that does not privilege the ad server’s own firm) values of \( a \) and \( r \) under the new ranking. Changes to the first sum reveal that firm 2’s bid had depended upon \( a_1r_1 \) and \( a_2r_2 \), but now firm 2 goes to slot 3 and thus its bid depends upon \( a_2r_2 \) and \( a_3r_3 \), as an example. Also, as the new top firm no longer generates revenue, \( k - 1 \) becomes \( k - 2 \).
The change in ad revenue, using the definitions of the parameters, is

\[
\sum_{2 \leq k \leq j} (k - 2)m_kq_k(a_k r_k - a_{k+1} r_k - a_{k-1} r_{k-1} - a_k r_k) - \sum_{2 \leq k \leq j} m_kq_k(a_{k-1} r_{k-1} - a_k r_k).
\]

This change is clearly negative. Holding the other values fixed, this change is decreasing in the initial slot assignment to the ad server’s firm, namely, \(j\). Holding that slot assignment fixed, it is decreasing in \(q_j\); i.e., it is becoming costlier. This is because the ad server’s firm in the top slot lets fewer consumers through, reducing downstream profits. It is decreasing in the full margin of the other firms \(m_kq_k\); the more profitable the bumped firms, the more costly it is to move them to a lower slot. The question, then, is whether these costs are offset by higher profits from sales of the product being advertised.

Again, we can turn to an explicit example to calculate the change in profit from an ad server privileging its own firm. As in the other examples, consider search frequencies all equal to 1 and no low-value consumers. Also, fix every firm’s margin to be 1 and set a range of relevances from 0.05 to 0.50 in increments of 0.05. Figure 1.10 shows the proportion change in total profits for the ad server from selling both ad space and the product being advertised by moving its firm to the top of the list relative to following the results of the auction without assistance given to the ad server’s own firm. The \(x\) axis gives the unassisted ranking of the ad server’s firm.

Increasing the rank of an already highly-ranked firm has small cost in terms of lost bids, but gains from increasing the size of the pool of consumers are also small. For low-ranked firms, the cost of lost bids becomes more substantial and begins to overwhelm gains from higher sales. Only by moving to the top a firm that was previously excluded from the list (here, firm 10) does the proportion change in profit start to increase again. This increase, too, should wain, however. An ad server has greatest incentive to subsidize its own firm by moving it to the top of the list when that firm would otherwise be neither high- or low-ranked or is just excluded from the listing.

Since this example uses search frequencies that are 1 all the way down the list, consumers continue to search until they find a suitable product. The margin of all firms is the same. Additionally, the relevances on the list do not change, only their order does (except when the excluded firm is added to the list and a formerly listed firm is removed). These facts together imply that total sales do not change and neither does consumer surplus; moving
the favored firm to the top only changes the distribution of the revenues among the firms. This would not be the case if the search frequencies were not 1. If consumers did not search the full list, moving a less relevant firm to the top would reduce the total number of sales and consumer surplus, changing producer surplus and its distribution. If firms had different margins, then changing the order would also change the efficiency in the allocation of production. Internal subsidization of a firm by the ad server can have important implications for the sizes and distributions of consumer and producer surpluses.

Note that this analysis assumes that consumers continue searching from the first listing downward even after the ad server’s firm is artificially raised to the top of the listing. We assume that consumers are not aware that they are viewing a less relevant ad and would perhaps instead be better off starting with the listing in the second position. This latter behavior would mitigate the incentive of the ad server to raise its firm to the top of the ad listing.

### 1.8 Extensions and Conclusions

The revenue raised in contextual advertising auctions has become essential to funding online content, from blogs to news to search engines. Innovation continues in this area to improve the relevance of the ads shown to consumers and to reduce search costs, each generating more
product sales and higher ad revenue. While these innovations benefit consumers, ad servers may be able to manipulate the market for the products being advertised, potentially harming producers and consumers. We must understand the incentives for an ad server to effect the product market as we consider the role of competition policy in online advertising markets. To assess these issues, we incorporate consumer behavior with the bidding strategies of firms to calculate the revenue generated by contextual advertising auctions.

We begin by developing a model of consumer responses to ad listings and products being offered at the listed sites. Based upon these responses, we find the optimal firm bidding strategies. We show how these strategies depend upon per-sale margins and the probability of a consumer liking the product in question, known as the relevance of the firm. We characterize how the margin and relevance can covary while consumers still find it rational to search from the top of the advertising list downward.

Given these strategies, we consider the incentives facing the ad server. We find that it has an incentive to decrease search costs, increase firm margins (holding prices fixed), and cultivate a more valuable pool of consumers, actions that benefit itself, firms, and consumers alike. Consumers also desire improvements in match probabilities. The ad server has an incentive to develop such improvements only to a point, while firms want even less improvement in matching algorithms. The ad server seeks thick markets that generate top firms with little dispersion in margins and relevances, as this reduces the ability of firms to shade their bids. Firms, of course, desire more shading. Ad servers limit the length of the listing to mitigate shading, while firms in total, along with consumers, prefer to have more listings. Ad servers have an incentive to subsidize internal divisions that provide the product being advertised, changing the sizes and distributions of producer and consumer surpluses. We see that the preferences of consumers align with the incentives of the ad server in some cases, while they align with the incentives of firms against the ad server in others.

There are several important extensions that can be made using the model in this chapter. The first would be to endogenize the search process of consumers by deriving an optimal stopping rule. Endogenized search decisions can impact the profitability of each slot and alter the incentives facing an ad server in other ways as well. Next, we might consider firms that offer the product at different prices and derive the conditions for it to remain optimal for consumers to continue searching from the top of the list to the bottom given the strategies of the firms.

Other papers have attempted to endogenize pricing decisions in the ordered search model (see, e.g., [Chen and He, 2006; Arbatskaya, 2007; Armstrong, Vickers and Zhou, 2009; Xu, Chen and Whinston, 2011a,b]). This assumes that a firm price discriminates based upon how consumers found its product; it charges a different price to consumers that clicked on its link listed first in one ad listing than to those that clicked on its link listed third in another listing and a different price still to those consumers that visited the site directly without searching. While price dispersion has been characterized across online firms for homogeneous goods (see, e.g., [Baye, Morgan and Scholten, 2006]), it has not been demonstrated for a single
firm across means of finding the product (e.g., different search engines, different ad listings, direct visits, etc). An interesting empirical exercise would be to look for evidence of this sort of price discrimination.

This model can also be extended to a situation where the ad server shares its profits with a content provider—Google sharing its profits from placing ads alongside an independent blog, for example. We can ask how these revenues are shared and how sharing alters the incentives for the ad server to pursue the actions discussed in this chapter. Going further, we can ask how competition among ad servers to secure the space alongside independent content changes the incentives facing the ad servers and how it impacts the quality and quantity of online content provision, all important questions as the internet continues to grow.

1.A Attrition by Low Value Consumers

In Section 1.2.2 we assume that consumers with valuations less than the market price $p$ continue searching forward despite the fact that they can never find a product that is relevant and priced below their valuation. This assumption may seem odd. It does not impact the qualitative results of the chapter, however.

Suppose instead that low value consumers realize that they are unable to ever find a suitable product after visiting the first site; that is, low value consumers stop searching after visiting the first site, knowing that, no matter how long they search, they will never make a purchase. This is the opposite case of Section 1.2.2 rather than experience attrition by high value consumers, here we find attrition by low value consumers.

This does not change the CTR, demand, or adjustment factor facing the first firm. For the second firm, the CTR is

$$r_2 = s_0 s_1 (1 - F(p))(1 - q_1);$$

the only consumers that continue onward are those that enter the list, have high valuations, did not find a relevant product at the first site, and continued on to the second site. Demand is

$$D_2(p) = s_0 s_1 (1 - F(p))(1 - q_1)q_2,$$

the proportion of consumers that both visit site 2 and find a relevant product. The expected margin per click is

$$\frac{m_2 D_2(p)}{r_2} = m_2 q_2.$$ 

Defining the adjustment factor analogously here as in Section 1.2.2 as the expected margin per click divided by the full margin of $m_2 q_2$ gives $a_2 = 1$. Indeed, $a_j = 1$ for all $1 < j \leq M$. Recall that the adjustment factor for firm 1 is $a_1 = 1 - F(p)$.

We can reapply the exact same results of the equilibrium analysis of Section 1.3.1 to
this case. The main condition given by Equation 1.5 is that
\[(m_jq_j - m_kq_k)(a_jr_j - a_kr_k) \geq 0.\]

In that section, we concluded that the full expected margins \(m_q\) were decreasing down the list because both \(a\) and \(r\) are decreasing down the list. In the where low value consumers drop out of the market after visiting the first site, the adjustment factor is increasing from site 1 to site 2: \(1 - F(p)\) to 1.

Nevertheless, the product \(ar\) is decreasing. See that \(a_1r_1 = s_0(1 - F(p))\), while \(a_2r_2 = s_0(1 - F(p))s_1(1 - q_1); a_1r_1 > a_2r_2\). Since \(a_j = 1\) for \(1 < j \leq M\), we have \(a_jr_j = r_j\).

The CTR is \(r_j = [1 - F(p)]s_0\prod_{k=1}^{j-1} s_k(1 - q_k)\), which is decreasing down the list, as consumers make purchases or quit searching.

The notion of an adjustment factor that adjusts the full expected margin by changes in the profitability of the consumer group visiting a particular site is flexible enough to incorporate many different behaviors of consumers. Hence, even in the case of attrition by low value consumers, the results of the auction given in Section 1.3 still apply and the qualitative results of the remainder of the chapter remain.
Chapter 2

Broken or Fixed Effects?

with Juan Carlos Suárez Serrato and Michael B. Urbancic
Department of Economics
University of California, Berkeley

2.1 Introduction

Fixed effects are a common means to “control for” unobservable differences related to particular qualities of the observations under investigation; examples include age, year, or location in cross-sectional studies or individual or firm effects in panel data. While fixed effects permit different mean outcomes between groups conditional upon covariates, the estimates of treatment effects are required to be the same; in more colloquial terms, the intercepts of the conditional expectations may differ, but not the slopes. An established result is that fixed effects regressions average the group-specific slopes proportional to both the conditional variance of treatment and the proportion of the sample in each group.\footnote{See, e.g., Angrist and Krueger (1999); Wooldridge (2005a); Angrist and Pischke (2009).} Researchers may believe that assuming a fixed effects model provides a convenient approximation of the sample-weighted effect and that models that incorporate group-specific effects yield estimates with significantly larger variances. In contrast to these beliefs, our replications of nine influential papers reveals large differences between these estimates without large increase in variances.

This chapter empirically demonstrates large differences between the estimate from a fixed effects model and an average of treatment effects weighted only by the sample frequency of each group, our desired estimand. To identify this parameter, we interact the treatment variable with the fixed effects to identify a separate effect for each group and to average these estimates weighted by sample frequencies. Our approach can be applied to a broad array of questions in applied microeconomics. We demonstrate the generality of our point by
examining nine papers from the *American Economic Review* between 2004 and 2009. We choose these papers because they are among the most highly cited articles from this period in the *AER* and are widely considered as important pieces in their fields.

The replication exercise demonstrates that, across a variety of units and groups of analysis, there are economically and statistically significant differences between the fixed effect estimate and the sample-weighted estimate. We employ the specification test that we develop to show that 6 of the 9 papers that we consider have sample-weighted estimates that are statistically different from the standard fixed effects estimates. Additionally, 7 of the 9 papers have estimates that differ in an economically significant way (taken here to mean differences of at least 10%). Averaging the largest deviance for each paper gives over 50% difference in the estimated treatment effect. We also show that our procedure does not markedly increase the variance of the estimator in 7 of 9 papers. While some of these papers do include interactions or run separate regressions for different groups, we show that there may be other statistically and substantively important interactions that might offer more informative estimates.

Our chapter begins by situating our approach in the literature in Section 2.2. In Section 2.3, we precisely define the parameter of interest in the presence of heterogeneity and show that FE models in this context are inconsistent estimators for the sample-weighted average except in special cases. We derive a test that distinguishes between the sample-weighted average and the FE estimate. To illustrate these results through an empirical example, in Section 2.4, we use a simplified model from Karlan and Zinman (2008) to compare the weighting scheme from the FE model to a sample-weighted approach and study the implications for the final estimate. We demonstrate the generality of these points in Section 2.5 in which we replicate eight other influential papers. We conclude in Section 2.6 by offering guidance to the applied researcher.

### 2.2 Incorporating heterogeneous treatment effects

In the presence of heterogeneous treatment effects across groups in the sample, the FE estimator gives an average of these effects. These weights depend not only on the frequency of the groups, but also upon sample variances within the groups. Angrist and Krueger (1999) compare the results from regression and matching estimators, demonstrating that the

---

2 For a discussion of how these papers were chosen, see Appendix 2.C.1. An earlier draft of this chapter had a stronger emphasis on the returns to education literature and included an analysis of the results of Acemoglu and Angrist (2000).

3 Thanks to a recent policy decision by the editorial board of the *AER*, it is possible to access the data and programs used in recently published articles and to replicate the results of these studies. We only analyze the data that the authors provide openly on the EconLit website. Though some of these papers include both OLS and instrumental variables approaches, we consider the implications of our approach for the OLS specifications to focus on the weighting scheme applied in this procedure.
effects of a dichotomous treatment are averaged using different weights in each procedure. Closest to our derivation below, Wooldridge (2005a) finds sufficient conditions for FE models to produce sample-weighted averages in correlated random coefficient models. Our analysis builds upon this derivation for the case of fixed coefficients and offers a different interpretation of the necessary conditions for this result. Additionally, while these papers provide a strong theoretical reason to believe that FE estimators do not provide sample-weighted estimates, we illustrate the empirical importance of this distinction using a broad array of microeconometric questions.

There has long been an interest in coefficient heterogeneity across cross-sectional groups. A notable early piece is Chow (1960). Here, he runs regressions separately by group, which is the most flexible way of permitting heterogeneity across these groups for a given model, and compares the predictive power of the separate regressions to that of the pooled regression, forming a test for differences in slopes and intercepts. We begin with a test in the same spirit, but we only test for different treatment effects and use a test robust to heteroskedasticity by using a Wald test. Our suggested means of incorporating heterogeneous treatment effects is through interaction terms, a less flexible, but more parsimonious solution. Many studies, including many of those that we replicate in this chapter, run separate regressions by group precisely because of the presence of treatment effect heterogeneity. Less common is the interacted model that we propose. Notable exceptions include Heckman and Hotz (1989), who consider the specific case of individual-specific time trends, which they call the random growth rate model. Papke (1994) and Friedberg (1998) also use the random growth model and find that the results of their studies are greatly influenced by trends that vary across geographic districts.

These examples, however, use interactions on predictors to avert omitted variables bias or to improve the fit of their models. In a different approach, Lochner and Moretti (2011) consider non-linearities in treatment effects, but do not estimate heterogeneous treatment effects across groups as we do here. In contrast to these works, the point of our analysis is that models that do not account for heterogeneous effects may provide inconsistent estimates of average effects.

We extend this literature in three ways. First, while Wooldridge (2005a) gives the sufficient conditions for a fixed effects model to deliver the sample-weighted treatment effect, we offer an alternative exposition and show what estimate is given by a FE model when this assumption fails. We focus on treatment effect heterogeneity and illustrate how it can be characterized and incorporated into a model in a parsimonious manner. Next, we derive a test that can distinguish between sample-weighted estimates derived from an interacted model and FE estimates. Our most important contribution is to show that these models are broadly empirically relevant in the applied economics literature.

4See also Angrist and Pischke (2009).
2.3 Interpreting FE estimates using projection results

In this section, we consider a specific model of heterogeneous treatment effects. Intuition might lead us to believe that, in the presence of heterogeneous treatment effects, FE estimates are sample-weighted averages of the group-level effects, the implicit parameter of interest. Instead, it has been established that, though the estimates are weighted combinations of group effects, they are not weighted by the size of the group; instead, these weights depend upon sample variances. We illustrate this point by applying the Frisch-Waugh-Lovell theorem to the fixed effects model.

2.3.1 FE model estimates compared to the SWE

Suppose that a researcher estimates a fixed-effects model using data arising from a process with heterogeneous treatment effects given by

\[ y_{ig} = \alpha_g + w_i \gamma + x_i \beta_g + \nu_i \]

\[ y = Z_{INT} \theta_{INT} + \nu, \]

where the effect of interest, \( \beta_g \), is group-specific. In this model, \( x_i \) is treatment, \( I_g \) is a vector of group fixed effects, and \( w_i \) is a vector of additional covariates.\(^5\) Though it may be instructive to consider the heterogeneity in these effects across groups, researchers often want a single summary of the treatment effect. A natural candidate would be the sample-weighted treatment effect, as explored in Wooldridge (2005b), as an example.

**Definition 1 (Sample-weighted treatment effect).** The sample-weighted treatment effect for the model in Equation 2.1 is

\[ \bar{\beta} = \sum_g \hat{\Pr}(g) \beta_g, \]

where \( \hat{\Pr}(g) = \frac{N_g}{N} \), \( N \) is the total number of observations in the sample and \( N_g \) is the number of observations belonging to fixed effect group \( g \in 1, \ldots, G \).

**Definition 2 (Sample-weighted coefficient estimates).** The sample-weighted coefficient estimates from an interacted model with regression coefficients \( \hat{\theta}_{INT} \) are

\[ \hat{\theta}_{SWE} = W \hat{\theta}_{INT} \equiv [I_K F_0] \hat{\theta}_{INT}, \]

\(^5\) Though there are \( G \) groups, there are \( G - 1 \) fixed effects included in the model for identification purposes. Assume that group \( G \) is the excluded group.
where \( I_K \) is a \( K \)-dimensional identity matrix, with \( K \) being the number of covariates not involving treatment, and

\[
F_0 = \frac{1}{N} \begin{bmatrix}
0 & \ldots & \ldots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & \ldots & \ldots & 0 \\
N_1 & N_2 & \ldots & N_{G-1}
\end{bmatrix}.
\]

Suppose that the researcher estimates a FE model that contains a single treatment effect parameter,

\[
y_{ig} = a_g + w_i c + x_i b + u_i \\
y = A_{FE} \theta_{FE} + x b + u;
\]

here, \( A_{FE} \) contains the fixed effects and covariates other than treatment. Following the Frisch-Waugh-Lovell theorem, we can find the coefficient estimate \( \hat{b} \) by multiplying both sides of this expression by the annihilator matrix \( M_A = I - (A' A)^{-1} A' \), giving

\[
M_A y = M_A x b + M_A u \\
\Rightarrow \hat{b} = (x' M_A x)^{-1} x' M_A y = \frac{\text{Cov} (\tilde{x}_i, y)}{\text{Var} (\tilde{x}_i)},
\]

where \( \tilde{x}_i \) is the projected value of treatment for observation \( i \).

The FE model above posits that the effect of treatment across groups is homogeneous. The OLS estimator \( \hat{b} \) is a consistent estimator of the sample-weighted effect only in special cases. Instead of a sample-weighted estimate, the FE estimator gives

\[
\sum_{g \in G} \hat{\text{Pr}}(g) \hat{\beta}_g \left( \frac{\text{Var} (\tilde{x}_i | g)}{\text{Var} (\tilde{x}_i)} \right),
\]

See Appendix 2.A.1 for a derivation of this result. We see that the FE and SWE are the same when the treatment effects are homogeneous or the variance of the projected treatment is the same across all groups. Otherwise, the FE estimator overweights groups that have larger variance of treatment conditional upon other covariates and underweights groups with smaller conditional variances.

From Equation 2.2, we see that, while FE models do provide a weighted combination of group effects, these effects are not weighted by sample frequencies. Instead, these weights depend upon sample variances, thereby producing estimates that are less informative for policy analysis. The weighting scheme employed by FE models provides a more efficient estimate of the treatment effect in the absence of heterogeneous treatment effects.
Chapter 2. Broken or Fixed Effects?

In the presence of heterogeneity, however, it does not produce an estimate that is readily interpretable or comparable across studies.

If the FE model is the true data-generating process, then there are homogeneous treatment effects. Hence, estimates arising from an analysis using only subgroup of our sample should be identical to those obtained by examining the entire sample with fixed effects included. This implies that the estimate of the treatment effect is invariant to the distribution of the groups in the sample. If the FE model does not hold, then the FE estimate \( \hat{b} \) is a function of the sample covariances; this statistic may change across samples or in subsamples. As a result, estimates are sample-dependent and not comparable across subsamples or studies.

**Proposition 1** (Sufficient condition for consistent estimation of sample-weighted treatment effects). The fixed effects model consistently estimates the sample-weighted average in the presence of heterogeneous treatment effects if the variance of treatment conditional on all other covariates is the same across all groups; i.e., \( \text{Var}(\tilde{x}_i | g) = \text{Var}(\tilde{x}_i) \forall g \). (see Appendix 2.A.1).

Thus, a regression on data from a perfectly randomized experiment where treatment has the same variance across groups yields the sample-weighted treatment effect. Such perfection is likely unattainable in observational or experimental settings, however. Indeed, in Section 2.5, we replicate a randomized experiment in Karlan and Zinman (2008). In that experiment, treatment (an interest rate on a microloan in South Africa) is randomized within different fixed effects groups (the risk category of the borrower), but the ranges of the (multi-valued) treatment are not the same across groups and, as a result, neither are the variances. In this case, we find that the sample-weighted treatment effect differs from the FE estimate by 61%. We use this case study to quantitatively illustrate the proposition above in Section 2.4.

### 2.3.2 A Test of Equality Between Sample-Weighted and FE Estimates

Even if the included interactions are statistically significant, it could be that their sample-weighted average is not statistically different from the standard FE model that excludes these interactions. We derive a specification test to discriminate between the FE estimate and the sample-weighted average.

**Proposition 2** (Specification Test of the differences between the FE estimates and the
Chapter 2. Broken or Fixed Effects?

sample-weighted average). The test of the following null hypothesis

\[ H_0 : \text{plim} (\hat{\theta}_{\text{SWE}} - \hat{\theta}_{\text{FE}}) = 0 \]
\[ H_a : \text{plim} (\hat{\theta}_{\text{SWE}} - \hat{\theta}_{\text{FE}}) \neq 0, \]

can be conducted by noting that the Wald test statistic

\[ H = \left( \hat{\theta}_{\text{SWE}} - \hat{\theta}_{\text{FE}} \right)' \left( N^{-1} \text{Var} \left[ \hat{\theta}_{\text{SWE}} - \hat{\theta}_{\text{FE}} \right] \right)^{-1} \left( \hat{\theta}_{\text{SWE}} - \hat{\theta}_{\text{FE}} \right) \]

has an asymptotic \( \chi^2(q) \) distribution under \( H_0 \), where \( q = \text{rank} \left( \hat{\theta}_{\text{SWE}} - \hat{\theta}_{\text{FE}} \right) \); \( H_0 \) is rejected at level \( \alpha \) when \( H > \chi^2_\alpha(q) \). Robust estimation of this test statistic is addressed in Appendix 2.A.2. This test is implemented by the Stata command GSSUtest discussed in Appendix 2.B.

This test compares all coefficients in both models. Other tests can also be conducted using \( \left( \hat{\theta}_{\text{SWE}} - \hat{\theta}_{\text{FE}} \right) \) and \( \text{Var} \left[ \hat{\theta}_{\text{SWE}} - \hat{\theta}_{\text{FE}} \right] \) by imposing the necessary restrictions on \( H \). For example, we provide \( t \) tests of the single null hypothesis that the estimate of the treatment effect from a FE model differs from the sample-weighted average in our meta-study in Section 2.5.

2.4 A Case Study: Karlan and Zinman (2008)

In this section, we provide a detailed case study of one of our selected AER papers. This example illuminates the exposition of Section 2.3.1 and further clarifies the relationship between the FE and sample-weighted estimates.

We show in Section 2.3.1 that if an experiment is perfectly randomized, then the FE estimate should equal the sample-weighted average. More specifically, all covariates need to be precisely uncorrelated with treatment within each group and the variance of treatment must be the same across all groups (see Equation 2.2). Among our AER replications, we have one experiment that we can consider more closely. Karlan and Zinman (2008) randomized the interest rate offered for a microloan across a population of South Africans. They look to identify the credit elasticity among this group.

In the case of Karlan and Zinman (2008), the authors include two sets of covariates other than the treatment: the financial risk of the borrower and the mailer wave of the experiment when the borrower participated. The distributions of treatment and risk level are nearly uncorrelated with the mailer wave, hence, we ignore these fixed effects in this section only for expository purposes. But, to offer interest rates commensurate with prevailing market rates, the authors needed to charge higher rates to higher risk individuals. Recall
Chapter 2. Broken or Fixed Effects?

that differing means in treatment do not drive the difference between the FE and SWE estimates, but rather differences in variances.

The authors offer not only higher rates to higher risk borrowers, but also offer a greater range of rates to this group; the variance of treatment differs across the groups. As a result, the FE estimate will not be equal to the SWE if the responsiveness to interest rates varies across risk groups.

The FE weights are given in column 2 of Table 2.1. These are the variances of treatment by group multiplied by the sample frequency of that group. Using these weights and the group effect estimated from an interacted model, given in column 4 of Table 2.1, we can calculate the FE estimate; this estimate is given in the bottom row of the table.

We can compare the weights from a FE model to the sample frequencies used to calculate a sample-weighted average; these weights appear in column 3 of Table 2.1. We see that high risk individuals are overweighted in the FE model and the low and medium risk individuals are underweighted. This accords with the design of the study—high risk borrowers had a wider range of interest rate offers and this relatively high variance in treatment leads to overweighing in the FE estimate.

Differences in weighting scheme are only important if the treatment effect is heterogeneous. We find that high-risk borrowers are much less responsive to the interest rate than low-risk borrowers. Because high-risk individuals are overweighted and have a smaller (in absolute value) treatment effect, the FE estimate underestimates the responsiveness of individuals to the interest rate by nearly 70%.

Table 2.1: Karlan and Zinman (2008) treatment effect weighting

<table>
<thead>
<tr>
<th>Risk group</th>
<th>FE weight</th>
<th>Sample freq.</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.045</td>
<td>0.125</td>
<td>-32.4</td>
</tr>
<tr>
<td>Medium</td>
<td>0.061</td>
<td>0.092</td>
<td>-9.9</td>
</tr>
<tr>
<td>High</td>
<td>0.894</td>
<td>0.783</td>
<td>-2.7</td>
</tr>
<tr>
<td>Average</td>
<td>-4.450</td>
<td>-7.050</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Note that the FE analogue here, -4.450, does not precisely equal the actual FE estimate of -4.37 due to correlation between mailer wave fixed effects and the interest rate (i.e., treatment).

6 Again, we assume that mailer wave is uncorrelated with treatment and drop it from the model that the authors actually employ. This is a reasonable assumption for these data. Hence, the variance conditional upon all covariates is just the variance of treatment by group.

7 The estimate that we calculate is not precisely equal to the FE estimate given in the paper. This is because we did not include the mailer wave fixed effects, explaining the difference between cited differences of 61 and 70%.
2.5 Fixed Effects Interactions: An AER Investigation

We have seen that, even in randomized experiments, FE models generally do not provide the sample-weighted estimate in the presence of heterogeneous treatment effects. To produce the SWE, we propose using an interacted model, following Equation [2.1] where the treatment effects are summarized by averaging the interacted effects weighted by the sample frequency of each group. To examine the differences between FE models and our approach more broadly, we turn to highly cited papers published in the American Economic Review between 2004 and 2009. We choose this publication due to its influence and the quality of its papers and consider recent years in order to capitalize upon the AER editorial board’s decision to require posting of data and other replication details to the EconLit online repository. The papers that we choose are well known in their respective fields and serve as prime examples of respected empirical work.

We find the nine most cited papers that use fixed effects in an OLS model as part of their primary specification and meet additional requirements, which serve to limit our scope to papers in applied microeconomics with a clear effect of interest. These papers are listed in Table 2.2 along with the outcomes, effects of interest, and fixed effects considered here. A complete description of the process that we follow to identify these papers can be found in Appendix 2.C.1 and a more detailed description of the regressions that we consider is given in Appendix 2.C.2.

2.5.1 Replication Results

To consider the importance of interactions in these papers, we first test the joint significance of the coefficients on the interactions between the effect of interest and the fixed effects using a standard Wald test. Then, we test whether a sample-weighted average arising from the interacted model differs from the estimate of the FE model. We develop a command called GSSUtest to perform these tests in Stata.

Our results appear in Tables 2.3 and 2.4. This table provides the *p*-values for Wald tests of joint significance of the interaction terms and the single test of the difference between the sample-weighted treatment effect and the fixed effect estimate and the percent difference between the treatment effects. Additional detail is provided in Tables 2.7 through 2.14.

Every paper that we consider has at least one set of fixed effects interactions that is significant at the 5% level. Some authors correctly separate regressions to account for these issues. For example, Lochner and Moretti (2004) are correct in separating their regressions by race, an alternative to adding interaction terms. Card, Dobkin and Maestas (2008) are

---

8 See Appendix 2.A.2 for details on this test.
9 See Appendix 2.B. The authors have posted a copy of this code online for researchers interested in implementing this test.
the most aggressive in the use of separate regressions, dividing the sample into educationby-race categories; the results suggest that this is merited. The use of separate regressions and interaction terms by all the authors is detailed in Table 2.6. For most papers, there is a need to include fixed effects interactions in the analysis and we recommend that authors explore this possibility.

Having demonstrated that fixed effects interactions are important covariates in these models using joint Wald tests, we now demonstrate that their inclusion produces sample-weighted averages that are statistically and economically different from estimates arising from the standard FE model. We define economically significant as a difference between the two estimates of more than ten percent of the FE estimate.

Seven of these papers have differences that are economically significant, exceeding ten percent upwards to over three hundred percent; averaging the largest difference for each paper gives over a 50% difference in the estimated treatment effect with a median of 19.5%. Six of the nine papers have a set of interactions that produce a sample-weighted average that is individually statistically different from the FE estimate at the 5% level. We note that our ability to distinguish between these two estimates is related to the power of the original analysis. These results are similar to those found by Graham and Powell (2010) in their case study on heterogeneous treatment effects. It is crucial that policy makers calibrate the estimates that they obtain from the sample to their population of interest in order to obtain accurate and informative economic assessments. Fixed effects interactions provide a way of obtaining estimates relevant for policy analysis.

2.5.2 The interacted and FE models and the variance-bias tradeoff

Our implementation of the interacted model incorporates group-specific treatment effects into a standard fixed effects regression. The choice between the standard FE model and the interacted version, then, can be viewed as the choice between short and long versions of a regression. The preceding discussion focuses on the bias of FE estimators relative to the SWE in a world of treatment effect heterogeneity. But, we are concerned with the variance of our estimators as well.

Suppose that the variance of our estimates is lower in the FE model relative to the interacted model. Goldberger (1991) provides rationales for short, potentially biased, regression over a long regression that has higher variance using the variance-bias tradeoff framework. We consider these rationales in the context of FE and interacted models using the empirical evidence found in our meta-study. They are:

\footnote{We may be worried about multiple testing issues here. A conservative Bonferroni correction states that, for a set of \( n \) hypotheses, we can reject the joint null that all \( n \) null hypotheses are true with size \( \alpha \) if we can reject any hypothesis individually at the \( \frac{\alpha}{n} \) level. Since we obtain \( p \)-values on the order of 0.000, we can reject the joint null that all the sample-weighted averages equal the FE estimates.}
• The researcher believes that $\theta_{INT,2} = 0$; i.e., treatment effects are homogeneous and thus the coefficients on the interactions are expected to be zero. Fortunately, this assumption can be tested using a joint significance test of the coefficients on the interaction variables. These interactions are significant in a vast majority of the cases that we consider, rendering this an inappropriate justification for choosing the FE model.

• The researcher believes that $\theta_{INT,2} \neq 0$, but might accept an imperfect approximation $\theta_{FE}$ with smaller standard errors. This choice depends upon the magnitude of the difference between the estimators. We find that the difference between the FE estimate and a sample-weighted average exceeds 10% in eight of the nine papers that we consider and averaging the largest deviations from each paper gives a difference of 50% between the treatment effects; the difference between the estimators is often substantial and consequential for policy analysis.

To evaluate the variance-bias tradeoff in our replications, we can examine the relationship between the largest absolute difference for each paper and compare that to the percent difference in standard error of the treatment effect between the two models; Figure 2.1 shows this relationship.[11] We see that, for seven of the papers, the variance does not substantively increase when calculating the SWE from an interacted model; indeed, it decreases for four of these papers. Hence, for these papers, it is not necessary to accept an imperfect estimate in order to achieve reduced standard errors.

[11] If the difference in the standard errors is positive, the SWE from the interacted model has a larger standard error. For Griffith, Harrison and Van Reenen (2006), the absolute difference is 324% and the percent change in standard errors is 630%; we exclude this outlier from the plot.
Figure 2.1: The relationship between the difference in the estimators and the change in variance among the AER replications
### Table 2.2: Papers from the *AER* used in the meta-analysis

<table>
<thead>
<tr>
<th>Citation</th>
<th>Outcome</th>
<th>Effect of interest</th>
<th>Fixed effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banerjee and Iyer (2005)</td>
<td>Fertilizer use</td>
<td>Prop. non-landlord land</td>
<td>Coastal dummy, year</td>
</tr>
<tr>
<td></td>
<td>Proportion irrigated</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Proportion other cereals</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Proportion rice</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Proportion wheat</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Proportion white rice</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rice yield (log)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wheat yield (log)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bedard and Deschênes (2006)</td>
<td>Smoking dummy</td>
<td>Veteran status</td>
<td>Age, education, race, region</td>
</tr>
<tr>
<td>Card et al. (2008)</td>
<td>Saw doctor dummy</td>
<td>Age over 65 dummy</td>
<td>Ethnicity, gender, region, year, education level</td>
</tr>
<tr>
<td></td>
<td>Was hospitalized dummy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Karlan and Zinman (2008)</td>
<td>Loan size</td>
<td>Interest rate (log)</td>
<td>Mailer wave, risk category</td>
</tr>
<tr>
<td>Lochner and Moretti (2004)</td>
<td>Imprisonment</td>
<td>Education</td>
<td>Race, age, year</td>
</tr>
<tr>
<td>Meghir and Palm (2005)</td>
<td>Wage (log; change in)</td>
<td>Education reform</td>
<td>High ability dummy, high father’s education dummy, sex, year</td>
</tr>
<tr>
<td>Oreopoulos (2006)</td>
<td>Wage (log)</td>
<td>Education</td>
<td>Age, Northern Ireland</td>
</tr>
<tr>
<td></td>
<td>Operating returns</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Additional details on our replications are found in Appendix 2.C.
Table 2.3: *AER* replication results

<table>
<thead>
<tr>
<th>Citation</th>
<th>Fixed effect</th>
<th>Joint test</th>
<th>Diff. test</th>
<th>% diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banerjee and Iyer (2005)</td>
<td>Coastal</td>
<td>0.231</td>
<td>0.827</td>
<td>-1.1</td>
</tr>
<tr>
<td>(Prop. irrigated)</td>
<td>Soil — black</td>
<td>0.387</td>
<td>0.482</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>Soil — red</td>
<td>0.080</td>
<td>0.172</td>
<td>19.5†</td>
</tr>
<tr>
<td></td>
<td>Soil — other</td>
<td>0.555</td>
<td>0.649</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>0.000**</td>
<td>0.901</td>
<td>0.0</td>
</tr>
<tr>
<td>Bedard and Deschénes (2006)</td>
<td>Age</td>
<td>0.944</td>
<td>0.914</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>Education</td>
<td>0.002**</td>
<td>0.374</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>Race</td>
<td>0.080</td>
<td>0.089</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Region</td>
<td>0.701</td>
<td>0.218</td>
<td>0.2</td>
</tr>
<tr>
<td>Card et al. (2008)</td>
<td>Ethnicity† (saw doctor)</td>
<td>0.000**</td>
<td>0.044*</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>0.000**</td>
<td>0.665</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>Region</td>
<td>0.156</td>
<td>0.882</td>
<td>-0.1</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>0.067</td>
<td>0.004**</td>
<td>-23.0†</td>
</tr>
<tr>
<td></td>
<td>Education (whites)†</td>
<td>0.004**</td>
<td>0.002**</td>
<td>-12.5†</td>
</tr>
<tr>
<td></td>
<td>Education (non-whites)‡</td>
<td>0.771</td>
<td>0.323</td>
<td>-1.3</td>
</tr>
<tr>
<td></td>
<td>Ethnicity (hospitalized)‡</td>
<td>0.000**</td>
<td>0.459</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>0.000**</td>
<td>0.012*</td>
<td>-1.3</td>
</tr>
<tr>
<td></td>
<td>Region</td>
<td>0.015*</td>
<td>0.732</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>0.778</td>
<td>0.722</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>Education (whites)‡</td>
<td>0.003**</td>
<td>0.048*</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>Education (non-whites)‡</td>
<td>0.746</td>
<td>0.295</td>
<td>5.7</td>
</tr>
<tr>
<td>Griffith et al. (2006)</td>
<td>Industry</td>
<td>0.000**</td>
<td>0.016*</td>
<td>-324.3†</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>0.040*</td>
<td>0.050*</td>
<td>6.5</td>
</tr>
</tbody>
</table>

Notes: Column 3 gives the *p*-value for the test of the joint significance of the interaction terms using a Wald test. Column 4 gives the *p*-value for a *t* test of the difference between the sample-weighted estimate and the FE estimate. Column 5 gives the percent difference between these two estimates. A single star indicates significance at the 5% level; two stars indicate significance at the 1% level. A dagger indicates a difference of more than 10% between the two estimates. A double dagger indicates whether the author considers heterogeneity among these groups. Results for two outcomes of interest are reported for Card et al. (2008); those outcomes are indicators for whether the individual saw a doctor or was hospitalized in the previous year.
Table 2.4: *AER* replication results, continued

<table>
<thead>
<tr>
<th>Citation</th>
<th>Fixed effect</th>
<th>Joint test</th>
<th>Diff. test</th>
<th>% diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karlan and Zinman (2008)</td>
<td>Mailer wave</td>
<td>0.330</td>
<td>0.837</td>
<td>-1.1</td>
</tr>
<tr>
<td></td>
<td>Risk category</td>
<td>0.016*</td>
<td>0.010*</td>
<td>61.3†</td>
</tr>
<tr>
<td>Lochner and Moretti (2004)</td>
<td>Race‡ (all)</td>
<td>0.000**</td>
<td>0.000*</td>
<td>-0.9</td>
</tr>
<tr>
<td></td>
<td>Age (blacks)</td>
<td>0.000**</td>
<td>0.000**</td>
<td>33.4†</td>
</tr>
<tr>
<td></td>
<td>Year (blacks)</td>
<td>0.000**</td>
<td>0.000**</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>Age (whites)</td>
<td>0.000**</td>
<td>0.000**</td>
<td>30.9†</td>
</tr>
<tr>
<td></td>
<td>Year (whites)</td>
<td>0.002**</td>
<td>0.286**</td>
<td>0.22</td>
</tr>
<tr>
<td>Meghir and Palme (2005)</td>
<td>High father’s education‡</td>
<td>0.000**</td>
<td>0.244</td>
<td>18.5†</td>
</tr>
<tr>
<td></td>
<td>Sex‡</td>
<td>0.527</td>
<td>0.747</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>0.000**</td>
<td>0.013*</td>
<td>0.5</td>
</tr>
<tr>
<td>Oreopoulos (2006)</td>
<td>N.Ireland‡</td>
<td>0.000**</td>
<td>0.000**</td>
<td>4.4</td>
</tr>
<tr>
<td></td>
<td>Age (GB)</td>
<td>0.000**</td>
<td>0.360</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>Age (NI)</td>
<td>0.000**</td>
<td>0.150</td>
<td>-2.7</td>
</tr>
<tr>
<td></td>
<td>Age (NI &amp; GB)</td>
<td>0.000**</td>
<td>0.634</td>
<td>0.6</td>
</tr>
<tr>
<td>Pérez-González (2006)</td>
<td>Family ownership (MB)</td>
<td>0.223</td>
<td>0.243</td>
<td>18.0†</td>
</tr>
<tr>
<td></td>
<td>Family ownership (OR)</td>
<td>0.483</td>
<td>0.489</td>
<td>10.4†</td>
</tr>
<tr>
<td></td>
<td>Year (MB)</td>
<td>0.002**</td>
<td>0.329</td>
<td>-11.4†</td>
</tr>
<tr>
<td></td>
<td>Year (OR)</td>
<td>0.010**</td>
<td>0.829</td>
<td>-2.4</td>
</tr>
</tbody>
</table>

Notes: Column 3 gives the *p*-value for the test of the joint significance of the interaction terms using a Wald test. Column 4 gives the *p*-value for a *t* test of the difference between the sample-weighted estimate and the FE estimate. Column 5 gives the percent difference between these two estimates. A single star indicates significance at the 5% level; two stars indicate significance at the 1% level. A dagger indicates a difference of more than 10% between the two estimates. A double dagger indicates whether the author considers heterogeneity among these groups.
2.6 Conclusion

This chapter contributes to the applied econometrics literature by illustrating a common issue in the application of fixed effects. Fixed effects are commonly employed to “control for” differences between groups. In the presence of heterogeneous treatment effects, researchers may intuitively believe that their estimates are sample-weighted averages of the group treatment effects. Though this is generally the parameter of interest, it is generally not the parameter that is identified by standard fixed effects models. We demonstrate this point using econometric theory and characterize its relevance to empirical applications.

Using an application of the Frisch-Waugh-Lovell theorem, we show that fixed effects models do not estimate the sample-weighted average treatment effect. We offer a sufficient condition for this difference to be 0 asymptotically and give an intuitive explanation of what is estimated if this condition is not met. We provide statistical tools to assess the importance of interaction terms, including a statistical test for the difference between the fixed effects estimate and the sample-weighted average from an interacted model. By employing these techniques, researchers can find estimates that are easier to interpret, that can be compared across academic studies, and that are more relevant for policy analysis.

While the sample-weighted average may be the most informative single statistic of the treatment effect for a sample, even it may not be the most relevant result for policy analysis. By identifying different effects for each subgroup, researchers can characterize patterns of treatment effect heterogeneity, permitting them to conduct more appropriate policy analysis and produce results that are comparable across academic studies. This process also generates a more flexible functional form that can better approximate the true data generating process.

Results from a replication exercise show that fixed effects interactions are significant in every paper that we consider across a variety of effects of interest and outcomes. The sample-weighted estimate is statistically different from the fixed effects estimate in six papers of the nine papers that we consider and substantively different in seven; using the largest difference for each paper, the average difference across replications is over 50%. Our results also show that we can achieve our desired estimand without accepting an increase in variance. Finally, while authors often include interactions or run regressions separately for different subpopulations, incorporating these heterogeneous effects into a meaningful summary of mean effects would provide a better characterization of the data generating process without a substantial increase in variance.
2. A  Topics in Fixed Effects Theory

2. A. 1  Sufficient Conditions for Estimation of Sample-Weighted Treatment Effects in FE Models

Suppose that a researcher estimates a fixed-effects model

\[ y_{ig} = \alpha_g + w_i \gamma + x_i b + e_i \]

\[ \equiv a_i \delta + x_i b + e_i, \]

where \( a_i \) contains the fixed effects and covariates other than treatment, \( x_i \). Stacking these equations across all observations \( i \) gives

\[ y = A \delta + xb + e. \]

Following the Frisch-Waugh-Lovell theorem, we can find the coefficient \( b \) by multiplying both sides of this expression by the annihilator matrix \( M_A = I - (A'A)^{-1} A' \), giving

\[ M_A y = M_A xb + M_A u \]

\[ \Rightarrow \hat{b} = (x'M_A x)^{-1} x'M_A y = \frac{\text{Cov} (\tilde{x}_i, y)}{\text{Var} (\tilde{x}_i)}, \]

where \( \tilde{x}_i \) is the projected value of treatment for observation \( i \). Define the group-specific effect as

\[ \hat{\beta}_g = \frac{\text{Cov} (\tilde{x}_i, y \mid g)}{\text{Var} (\tilde{x}_i \mid g)}. \]

We can decompose the estimate of \( \hat{b} \) following

\[ \hat{b} = \frac{\text{Cov} (\tilde{x}_i, y_i)}{\text{Var} (\tilde{x}_i)} \]

\[ = \sum_{g \in G} \Pr (g) \frac{\text{Cov} (\tilde{x}_i, y_i \mid g)}{\text{Var} (\tilde{x}_i)} \]

\[ = \sum_{g \in G} \Pr (g) \hat{\beta}_g \frac{\text{Var} (\tilde{x}_i \mid g)}{\text{Var} (\tilde{x}_i)} \]

\[ = \sum_{g \in G} \Pr (g) \hat{\beta}_g \left( \frac{\text{Var} (\tilde{x}_i \mid g)}{\text{Var} (\tilde{x}_i)} \right). \]
The second equality follows because we are considering a specific type of covariate—binary fixed effects. Thus, it is clear that the estimate of the treatment effect arising from the fixed effects model is not simply a frequency-weighted average of the group-specific effects. This is only the case if the conditional variances of the treatment within each group are the same.

The bias of the FE model in estimating the sample-weighted average, $\bar{\beta}$, has the following limit:

$$\text{plim } (\bar{\beta} - \hat{b}) = \sum_g \left( \Pr(I_g = 1) - \frac{\Pr(I_g = 1) \text{Var}(x|I_g = 1)}{\text{Var}(x)} \right) \beta_g$$

$$= \sum_g \Pr(I_g = 1) \left( 1 - \frac{\text{Var}(x|I_g = 1)}{\text{Var}(x)} \right) \beta_g.$$

Again, this difference is 0 if $\text{Var}(\tilde{x}_i | g) = \text{Var}(\tilde{x}_i) \forall g$.

### 2.A.2 Calculating the Difference Between the Fixed Effects and Weighted Interactions Estimators

We may wonder whether the difference between the FE model estimate of the treatment effect is statistically significantly different from a sample-weighted estimate of the treatment effect arising from the interacted model. Define the fixed-effects model (FE) as

$$\begin{align*}
y_{ig} &= a_g + w_i c + x_i b + u_i \\
y &= Z_{FE} \theta_{FE} + u
\end{align*}$$

and the interacted model as

$$\begin{align*}
y_{ig} &= \alpha_g + w_i \gamma + x_i \beta_g + \nu_i \\
y &= Z_{INT} \theta_{INT} + \nu,
\end{align*}$$

where $i$ indexes the individual unit from 1 to $N$, $g$ indexes group membership from 1 to $G$, and $\theta_{FE}' = [a_1, \cdots, a_G, c', b']$, and $\theta_{INT}' = [\alpha_1, \cdots, \alpha_G, \gamma', \beta_1', \cdots, \beta_G']$. The crucial difference between these two models is that the interacted model allows the coefficient on $x_i$ to vary across groups.

The test that we propose considers whether the sample-weighted average of $\beta_g$ in the interacted model equals $b$ from the FE model. We derive the distribution of the test statistic through joint estimation of the models using a Method of Moments (MM) approach. We first derive the joint distribution of the estimators, then we develop a specification test for our particular hypothesis.
For these models, the sets of moment conditions are given by:

\[
\sum_{i=1}^{N} h_{FE,i} \left( \hat{\theta}_{FE} \right) \equiv \sum_{i=1}^{N} z_{FE,i} \left( y_{ig} - z_{FE,i} \hat{\theta}_{FE} \right) = 0 \quad \text{and} \\
\sum_{i=1}^{N} h_{INT,i} \left( \hat{\theta}_{INT} \right) \equiv \sum_{i=1}^{N} z_{INT,i} \left( y_{ig} - z_{INT,i} \hat{\theta}_{INT} \right) = 0.
\]

Stacking these equations into \( \sum_{i=1}^{N} h_i \left( \hat{\delta} \right) = 0 \), where \( \hat{\delta}' = \left[ \hat{\theta}'_{FE}, \hat{\theta}'_{INT} \right] \) and \( \delta'_0 = [\theta'_{FE}, \theta'_{INT}] \), and applying standard MM arguments (see, e.g., [Cameron and Trivedi, 2005]), it follows that \( \hat{\delta} \) has the property that

\[
\sqrt{N} \left( \hat{\delta} - \delta_0 \right) \xrightarrow{d} N' \left( 0, G_0^{-1} S_0 (G_0')^{-1} \right),
\]

where

\[
G_0 = \text{plim} \frac{1}{N} \sum_{i=1}^{N} \left[ \frac{\partial h_i}{\partial \delta} \right]_{\delta=\delta_0} \quad \text{and} \quad S_0 = \text{plim} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \left[ h_i h'_j \right]_{\delta=\delta_0}.
\]

Note that, by partitioning the matrix \( G_0 = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \) and using the fact that

\[
\frac{\partial h_{i,FE}}{\partial \theta_{INT}} = 0 \quad \text{and} \quad \frac{\partial h_{i,INT}}{\partial \theta_{FE}} = 0,
\]

it follows that \( G_{21} = G_{12} = 0 \).

As is standard (once again, see [Cameron and Trivedi, 2005]), we estimate \( G_0 \) via

\[
\hat{G} = \frac{1}{N} \sum_{i=1}^{N} \left[ \frac{\partial h_i}{\partial \delta} \right]_{\delta=\hat{\delta}}.
\]

To estimate \( S_0 \) we consider two cases. First, assuming independence over \( i \), an estimator robust to heteroskedasticity is

\[
\hat{S}_R = \frac{1}{N} \sum_{i=1}^{N} h_i \left( \hat{\delta} \right) h_i \left( \hat{\delta} \right)'.
\]
A second estimator that incorporates clustered errors is
\[ \hat{S}_C = \frac{1}{N} \sum_{c=1}^{C} \sum_{i=1}^{N_c} \sum_{j=1}^{N_c} n_{ic} \left( \hat{\delta}_i \right) n_{jc} \left( \hat{\delta}_j \right)'. \]

Thus, robust and clustered estimators of the variance of \( \hat{\delta} \) are
\[ \hat{\text{Var}}[\hat{\delta}] = \hat{G}^{-1} \hat{S}_e \left( \hat{G}' \right)^{-1} \] for \( e = R, C \) respectively.

Now we turn to the specific hypothesis that we would like to consider; namely, that the sample-weighted averages of the estimates from the interacted model are equal to the FE estimates. Specifically, our hypothesis is
\[ H_0 : \text{plim} \left( W \hat{\theta}_{\text{INT}} - \hat{\theta}_{\text{FE}} \right) = 0 \]
\[ H_a : \text{plim} \left( W \hat{\theta}_{\text{INT}} - \hat{\theta}_{\text{FE}} \right) \neq 0, \]
where \( W \) is defined as
\[ W \equiv \begin{bmatrix} I_Q & 0_{(Q-1 \times K-1)} \\ f & \end{bmatrix} \]
to produce a sample-weighted estimate of the treatment effect and to return the other parameters.\(^{12}\) In this formulation, \( Q \) is the rank of \( Z_{\text{FE}} \), \( I_Q \) is a \( Q \times Q \) identity matrix, \( K \) is the number of fixed-effect groups, and \( f \) is a \( [1 \times K - 1] \) vector of sample frequencies of fixed effect group membership.

To compute the difference of the estimators, define the matrix
\[ R = [-I_Q, W]. \]
Then, the difference between the estimators is \( R \hat{\delta} = W \hat{\theta}_{\text{INT}} - \hat{\theta}_{\text{FE}} \) and the variance of this difference is estimated according to
\[ \hat{\text{Var}}[W \hat{\theta}_{\text{INT}} - \hat{\theta}_{\text{FE}}] = R \hat{\text{Var}}[\hat{\delta}] R'. \]

The Wald test statistic
\[ H = \left( W \hat{\theta}_{\text{INT}} - \hat{\theta}_{\text{FE}} \right)' \left( N^{-1} \hat{\text{Var}} \left[ W \hat{\theta}_{\text{INT}} - \hat{\theta}_{\text{FE}} \right] \right)^{-1} \left( W \hat{\theta}_{\text{INT}} - \hat{\theta}_{\text{FE}} \right) \]
has an asymptotic \( \chi^2(q) \) distribution under \( H_0 \); \( H_0 \) is rejected at level \( \alpha \) when \( H > \chi^2_{\alpha}(q) \).

\(^{12}\)Recall that, in our case, \( x_i \) is a scalar.
This test compares all coefficients in both models. Other tests can also be conducted using 
\(W\hat{\theta}_{INT} - \hat{\theta}_{FE}\) and \(\text{Var}[W\hat{\theta}_{INT} - \hat{\theta}_{FE}]\) by imposing the necessary restrictions on \(H\). For example, we provide \(t\) tests of the single null hypothesis that the estimate of the treatment effect from a FE model differs from the sample-weighted average in our meta-study.

2.B  **GSSUtest.ado**

As a companion to this chapter, we develop a Stata command called **GSSUtest** that computes the sample-weighted average treatment effect, tests for equality of coefficients with those of a fixed effects model, and computes the percentage change in the parameter of interest. The command is available from the authors and can be executed with the following syntax:

GSSUtest y Tr FEg [varlist] [if] [in] [, options]

where

- **y** is the dependent variable,
- **Tr** is the independent variable of interest (e.g., treatment) and,
- **FEg** is a categorical variable indexing the fixed effect group.

Other predictors can be included in **varlist** and several options including sample weights and clustering are also available. **GSSUtest** automatically uses robust standard errors in its calculations.

2.C  **AER Replications**

2.C.1  **Paper Selection**

We aim to show the broad importance of these fixed effects interactions in capturing the sample-weighted average treatment effect. We do this by replicating high quality papers from a variety of fields. We begin by outlining guidelines for inclusion in our analysis:

- The paper must be in the *American Economic Review*. We enact this qualification in order to limit our universe of analysis both in terms of quantity and quality of papers and to guarantee easy access to the necessary data.

- The paper must be published in the March 2004 issue or later (to March 2009, the issue predating our literature search). The *AER* policy during this period requires that, barring any acceptable restriction, data for these papers be posted to the EconLit website. This leads to the condition that:
Chapter 2. Broken or Fixed Effects?

- The data necessary to replicate the main specification(s) of the paper must be readily available on the EconLit website. We use these data and direct those interested to the EconLit website to obtain these files.

- The main specification(s) of the paper must have a specific effect of interest.

- The main specification(s) of the paper must use some type of fixed effect. We identify papers meeting this qualification by searching the PDF files of the published papers for the terms “fixed effect” (which captures the plural “effects” as well) and for “dumm” (which captures “dummy” or “dummies,” common synonyms for fixed effects).

- We limit ourselves to microeconomic analyses and do not consider papers based on financial economics issues.

- We ignore papers that require special methods to incorporate time series issues.

  We choose to replicate a total of nine papers in our analysis. To order our search, we consider papers in order of citations per year since publication. First, we use the citation counts provided by the ISI Web of Science on July 16, 2009. We limit our search to the American Economic Review and years 2004–2009, as outlined above. Unfortunately, the Web of Science does not provide the volume for the papers contained therein. We create an algorithm that assigns a volume number to a paper based upon its page number; these assignments are verified as papers are considered. The total number of citations are divided by the years since publication. For example, in June 2009, a paper published in June 2004 was published 5 years ago and a paper published in September 2004 was published 4.75 years ago.

  Citation counts are very noisy in the short time after publication that we consider here. Our citations-per-year metric might overweight later papers. Nonetheless, we consider all papers in this period with over 20 citations and 86% of all papers with 15 or more citations. It appears that we consider most of the highly cited papers from this period and do not ignore the most recent papers, as would occur using the gross citation count.

  Papers that we select must be highly-cited and fit the qualifications necessary to be relevant to our inquiry; we replicate papers from 2004, 2005, 2006, and 2008, missing only 2007 and the one quarter of 2009 that predates our search. We examine a breadth of papers that covers several fields, several years, and several units of analysis and thus they serve as a decent representation of the use of fixed effects in the applied econometrics literature.

  Before incorporating interaction terms into the specifications that we consider, we first ensure that we can replicate the results obtained by the authors as given in their respective

---

\footnote{We determine which specifications are the “main” ones by considering the discussion of the effects in the text by the authors and ignore those specifications identified as robustness checks.}

\footnote{In June 2009, 1 citation for a paper published in March 2009 is equal to 4 for a paper published in June 2008 and 20 for a paper published in June 2004.}
papers. We can provide Stata D0 and log files that generate and produce these results. We extend these files by incorporating the interactions as introduced in the paper. In choosing the interactions when there are several fixed effects in the regressions, we choose such that the number of groups is not unruly (U.S. states, for example, may simply produce too many terms to be informative). Our interacted regressions preserve all other features of the replicated specifications (e.g., clustering, robust standard errors, and inclusion of other covariates) unless otherwise noted in the text.

We do not justify that the interactions that we employ are the most salient within the given economic situation. Additionally, we do not suggest that the inclusion of interactions is the first-order extension of the analysis in the papers that we examine. We make no effort to search the subsequent literature to identify other areas of concern in these papers. Lastly, many of these papers employ instrumental variables to confront endogeneity. In these cases, we use the base OLS case to illustrate our point.

2.C.2 Replication Details

We replicate the specifications cited in Table 2.5. Some of these authors include fixed effects interactions or run regressions separately for subgroups; we list these practices in Table 2.6. In Banerjee and Iyer (2005), the authors have eight separate outcomes of interest. In the body of the chapter, we give results only for a sample of these results. In Tables 2.11 through 2.14, we provide the results for all outcome-group combinations.

<table>
<thead>
<tr>
<th>Citation</th>
<th>Table</th>
<th>Column</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banerjee and Iyer (2005)</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Bedard and Deschenes (2006)</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Card, Dobkin and Maestas (2008)</td>
<td>3</td>
<td>6, 8</td>
</tr>
<tr>
<td>Griffith, Harrison and Van Reenen (2006)</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Karlan and Zinman (2008)</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Lochner and Moretti (2004)</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Meghir and Palme (2005)</td>
<td>2</td>
<td>1 (row 1)</td>
</tr>
<tr>
<td>Oreopoulos (2006)</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Pérez-González (2006)</td>
<td>9</td>
<td>1, 6</td>
</tr>
</tbody>
</table>

Notes: In Griffith, Harrison and Van Reenen (2006), we do not cluster at the industry level as the authors do in their paper. We also do not cluster as Oreopoulos (2006) does. In both cases, clustering does not change the results. We are not able to replicate the point estimate that Oreopoulos (2006) provides for his regression of Northern Ireland and Great Britain combined; we use the specification that he provides and base our results on this model.
Table 2.6: Fixed effects interactions and regressions by subgroup conducted in the original papers

<table>
<thead>
<tr>
<th>Citation</th>
<th>Separate regressions</th>
<th>Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banerjee and Iyer (2005)</td>
<td>Entire country, subregion</td>
<td></td>
</tr>
<tr>
<td>Bedard and Deschênes (2006)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Card, Dobkin and Maestas (2008)</td>
<td>Race × education</td>
<td>Age, age-squared</td>
</tr>
<tr>
<td>Griffith, Harrison and Van Reenen (2006)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Karlan and Zinman (2008)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lochner and Moretti (2004)</td>
<td>Race (black, white)</td>
<td></td>
</tr>
<tr>
<td>Meghir and Palme (2005)</td>
<td>Sex</td>
<td>Sex</td>
</tr>
<tr>
<td></td>
<td>Father’s education (low, high)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ability (low, high)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ability × father’s education × sex</td>
<td></td>
</tr>
<tr>
<td>Pérez-González (2006)</td>
<td>Less selective college dummy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Graduate school dummy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive R&amp;D spending dummy</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Separate regressions and interaction terms only listed for specifications based upon the one given in Table 2.5. Pérez-González (2006) does not include the dummy variables that he subsequently interacts with treatment in his base regression; hence, we do not test their interactions here.
Table 2.7: Detailed replication results

<table>
<thead>
<tr>
<th>Citation</th>
<th>Fixed effect</th>
<th>FE est.</th>
<th>FE SE</th>
<th>SWE</th>
<th>SWE SE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bedard and Deschênes (2006)</strong></td>
<td>Age</td>
<td>0.078</td>
<td>0.005</td>
<td>0.078</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>Education</td>
<td>0.078</td>
<td>0.005</td>
<td>0.078</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>Race</td>
<td>0.078</td>
<td>0.005</td>
<td>0.078</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>Region</td>
<td>0.078</td>
<td>0.005</td>
<td>0.078</td>
<td>0.005</td>
</tr>
<tr>
<td><strong>Card et al. (2008)</strong></td>
<td>Ethnicity (saw doctor)</td>
<td>0.013</td>
<td>0.008</td>
<td>0.013</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>0.013</td>
<td>0.008</td>
<td>0.013</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>Region</td>
<td>0.013</td>
<td>0.008</td>
<td>0.013</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>0.013</td>
<td>0.010</td>
<td>0.013</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>Education (whites)</td>
<td>0.006</td>
<td>0.008</td>
<td>0.006</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>Education (non-whites)</td>
<td>0.039</td>
<td>0.013</td>
<td>0.039</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>Ethnicity (hospitalized)</td>
<td>0.012</td>
<td>0.004</td>
<td>0.012</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>0.012</td>
<td>0.004</td>
<td>0.012</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>Region</td>
<td>0.012</td>
<td>0.004</td>
<td>0.012</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>0.012</td>
<td>0.004</td>
<td>0.012</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>Education (whites)</td>
<td>0.013</td>
<td>0.005</td>
<td>0.013</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>Education (non-whites)</td>
<td>0.005</td>
<td>0.007</td>
<td>0.006</td>
<td>0.007</td>
</tr>
<tr>
<td><strong>Griffith et al. (2006)</strong></td>
<td>Industry</td>
<td>0.076</td>
<td>0.014</td>
<td>-0.170</td>
<td>0.104</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>0.076</td>
<td>0.014</td>
<td>0.080</td>
<td>0.014</td>
</tr>
</tbody>
</table>

Notes: Column 1 gives the paper and column 2 gives the fixed effects under consideration. Columns 3 and 4 give the standard FE model estimate of the treatment effect and its standard error. Columns 5 and 6 give the sample-weighted estimate from an interacted model and its standard error. Results for two outcomes of interest are reported for Card et al. (2008); those outcomes are indicators for whether the individual saw a doctor or was hospitalized in the previous year.
Table 2.8: Detailed replication results, continued

<table>
<thead>
<tr>
<th>Citation</th>
<th>Fixed effect</th>
<th>FE est.</th>
<th>FE SE</th>
<th>SWE</th>
<th>SWE SE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lochner and Moretti (2004)</strong></td>
<td>Race (all)</td>
<td>-0.122</td>
<td>0.004</td>
<td>-0.121</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>Age (blacks)</td>
<td>-0.370</td>
<td>0.015</td>
<td>-0.493</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>Year (blacks)</td>
<td>-0.370</td>
<td>0.015</td>
<td>-0.379</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>Age (whites)</td>
<td>-0.099</td>
<td>0.003</td>
<td>-0.130</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>Year (whites)</td>
<td>-0.099</td>
<td>0.003</td>
<td>-0.099</td>
<td>0.003</td>
</tr>
<tr>
<td><strong>Meghir and Palme (2005)</strong></td>
<td>High father’s ed.</td>
<td>0.014</td>
<td>0.009</td>
<td>0.017</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>Sex</td>
<td>0.014</td>
<td>0.009</td>
<td>0.014</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>0.014</td>
<td>0.009</td>
<td>0.014</td>
<td>0.009</td>
</tr>
<tr>
<td><strong>Oreopoulos (2006)</strong></td>
<td>N.Ireland</td>
<td>0.078</td>
<td>0.002</td>
<td>0.081</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>Age (GB)</td>
<td>0.075</td>
<td>0.002</td>
<td>0.076</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>Age (NI)</td>
<td>0.106</td>
<td>0.004</td>
<td>0.104</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>Age (NI &amp; GB)</td>
<td>0.078</td>
<td>0.002</td>
<td>0.079</td>
<td>0.002</td>
</tr>
<tr>
<td><strong>Pérez-González (2006)</strong></td>
<td>High fam. own. (MB)</td>
<td>-0.256</td>
<td>0.086</td>
<td>-0.302</td>
<td>0.079</td>
</tr>
<tr>
<td></td>
<td>High fam. own. (OR)</td>
<td>-0.027</td>
<td>0.009</td>
<td>-0.030</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>Year (MB)</td>
<td>-0.256</td>
<td>0.086</td>
<td>-0.226</td>
<td>0.083</td>
</tr>
<tr>
<td></td>
<td>Year (OR)</td>
<td>-0.027</td>
<td>0.009</td>
<td>-0.027</td>
<td>0.009</td>
</tr>
<tr>
<td><strong>Karlan and Zinman (2008)</strong></td>
<td>Mailing wave</td>
<td>-4.368</td>
<td>1.093</td>
<td>-4.319</td>
<td>1.084</td>
</tr>
<tr>
<td></td>
<td>Risk category</td>
<td>-4.368</td>
<td>1.093</td>
<td>-7.047</td>
<td>1.917</td>
</tr>
</tbody>
</table>

Notes: Column 1 gives the paper and column 2 gives the fixed effects under consideration. Columns 3 and 4 give the standard FE model estimate of the treatment effect and its standard error. Columns 5 and 6 give the sample-weighted estimate from an interacted model and its standard error. The regression coefficients and standard errors from **Lochner and Moretti (2004)** are multiplied by 100, following the reporting of the authors in their paper.
Table 2.9: Detailed replication results, continued

<table>
<thead>
<tr>
<th>Citation</th>
<th>Fixed effect</th>
<th>Joint test of interactions</th>
<th>Test of treat. diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Wald stat.</td>
<td>DF</td>
</tr>
<tr>
<td>Bedard and Deschênes (2006)</td>
<td>Age</td>
<td>11.09</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Education</td>
<td>14.79</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Race</td>
<td>3.07</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Region</td>
<td>5.51</td>
<td>8</td>
</tr>
<tr>
<td>Card et al. (2008)</td>
<td>Ethnicity (saw doctor)</td>
<td>18.71</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>114.37</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Region</td>
<td>5.23</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>18.67</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Education (whites)</td>
<td>13.13</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Education (non-whites)</td>
<td>1.13</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Ethnicity (hospitalized)</td>
<td>21.54</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>18.50</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Region</td>
<td>10.50</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>7.26</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Education (whites)</td>
<td>13.99</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Education (non-whites)</td>
<td>1.23</td>
<td>3</td>
</tr>
<tr>
<td>Griffith et al. (2006)</td>
<td>Industry</td>
<td>52.78</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>19.04</td>
<td>10</td>
</tr>
</tbody>
</table>

Notes: Column 1 gives the paper and column 2 gives the fixed effects under consideration. Column 3 gives the Wald statistic of a joint test of the significance of the interactions, column 4 gives the degrees of freedom for that test, and column 5 gives the p-value. Column 6 gives a t statistic from a test of the difference between the FE and sample-weighted estimates using the test derived in Appendix 2.A.2 and the corresponding p-value. Results for two outcomes of interest are reported for Card et al. (2008); those outcomes are indicators for whether the individual saw a doctor or was hospitalized in the previous year.
### Table 2.10: Detailed replication results, continued

<table>
<thead>
<tr>
<th>Citation</th>
<th>Fixed effect</th>
<th>Joint test of interactions</th>
<th>Test of treat. diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Wald stat.</td>
<td>DF</td>
</tr>
<tr>
<td><strong>Lochner and Moretti</strong></td>
<td>Race (all)</td>
<td>24.22</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Age (blacks)</td>
<td>865.10</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Year (blacks)</td>
<td>41.60</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Age (whites)</td>
<td>1860.06</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Year (whites)</td>
<td>12.03</td>
<td>2</td>
</tr>
<tr>
<td><strong>Meghir and Palme</strong></td>
<td>High father’s ed.</td>
<td>46.73</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Sex</td>
<td>0.40</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>41.96</td>
<td>11</td>
</tr>
<tr>
<td><strong>Oreopoulos</strong></td>
<td>N.Ireland</td>
<td>44.65</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Age (GB)</td>
<td>879.85</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Age (NI)</td>
<td>148468.65</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Age (NI &amp; GB)</td>
<td>173.47</td>
<td>28</td>
</tr>
<tr>
<td><strong>Pérez-González</strong></td>
<td>High fam. own. (MB)</td>
<td>1.48</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>High fam. own. (OR)</td>
<td>0.49</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Year (MB)</td>
<td>39.78</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Year (OR)</td>
<td>34.88</td>
<td>18</td>
</tr>
<tr>
<td><strong>Karlan and Zinman</strong></td>
<td>Mailer wave</td>
<td>2.21</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Risk category</td>
<td>8.26</td>
<td>2</td>
</tr>
</tbody>
</table>

Notes: Column 1 gives the paper and column 2 gives the fixed effects under consideration. Column 3 gives the Wald statistic of a joint test of the significance of the interactions, column 4 gives the degrees of freedom for that test, and column 5 gives the p-value. Column 6 gives a t statistic from a test of the difference between the FE and sample-weighted estimates using the test derived in Appendix 2.A.2 and the corresponding p-value.
Table 2.11: Detailed replication results for Banerjee and Iyer (2005)

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Fixed effect</th>
<th>FE est.</th>
<th>FE SE</th>
<th>SWE</th>
<th>SWE SE</th>
<th>% Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prop. Fertilized</td>
<td>Soil — red</td>
<td>10.71</td>
<td>3.33</td>
<td>12.03</td>
<td>3.47</td>
<td>12.4</td>
</tr>
<tr>
<td></td>
<td>Soil — black</td>
<td>10.71</td>
<td>3.33</td>
<td>10.78</td>
<td>3.46</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>Soil — all</td>
<td>10.71</td>
<td>3.33</td>
<td>10.67</td>
<td>3.36</td>
<td>-0.4</td>
</tr>
<tr>
<td></td>
<td>Coastal</td>
<td>10.71</td>
<td>3.33</td>
<td>10.73</td>
<td>3.33</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>10.71</td>
<td>3.33</td>
<td>10.76</td>
<td>3.34</td>
<td>0.5</td>
</tr>
<tr>
<td>Log yield</td>
<td>Soil — red</td>
<td>0.16</td>
<td>0.07</td>
<td>0.17</td>
<td>0.07</td>
<td>10.3</td>
</tr>
<tr>
<td></td>
<td>Soil — black</td>
<td>0.16</td>
<td>0.07</td>
<td>0.16</td>
<td>0.07</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>Soil — all</td>
<td>0.16</td>
<td>0.07</td>
<td>0.17</td>
<td>0.07</td>
<td>5.3</td>
</tr>
<tr>
<td></td>
<td>Coastal</td>
<td>0.16</td>
<td>0.07</td>
<td>0.16</td>
<td>0.07</td>
<td>-0.8</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>0.16</td>
<td>0.07</td>
<td>0.16</td>
<td>0.07</td>
<td>0.0</td>
</tr>
<tr>
<td>Log rice yield</td>
<td>Soil — red</td>
<td>0.17</td>
<td>0.08</td>
<td>0.16</td>
<td>0.08</td>
<td>-3.6</td>
</tr>
<tr>
<td></td>
<td>Soil — black</td>
<td>0.17</td>
<td>0.08</td>
<td>0.18</td>
<td>0.08</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>Soil — all</td>
<td>0.17</td>
<td>0.08</td>
<td>0.18</td>
<td>0.08</td>
<td>5.8</td>
</tr>
<tr>
<td></td>
<td>Coastal</td>
<td>0.17</td>
<td>0.08</td>
<td>0.17</td>
<td>0.08</td>
<td>-0.5</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>0.17</td>
<td>0.08</td>
<td>0.17</td>
<td>0.08</td>
<td>-0.2</td>
</tr>
<tr>
<td>Log wheat yield</td>
<td>Soil — red</td>
<td>0.23</td>
<td>0.07</td>
<td>0.24</td>
<td>0.07</td>
<td>6.4</td>
</tr>
<tr>
<td></td>
<td>Soil — black</td>
<td>0.23</td>
<td>0.07</td>
<td>0.24</td>
<td>0.07</td>
<td>3.4</td>
</tr>
<tr>
<td></td>
<td>Soil — all</td>
<td>0.23</td>
<td>0.07</td>
<td>0.24</td>
<td>0.07</td>
<td>6.8</td>
</tr>
<tr>
<td></td>
<td>Coastal</td>
<td>0.23</td>
<td>0.07</td>
<td>0.21</td>
<td>0.07</td>
<td>-6.7</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>0.23</td>
<td>0.07</td>
<td>0.23</td>
<td>0.07</td>
<td>-0.1</td>
</tr>
</tbody>
</table>

Notes: Column 1 gives the paper and column 2 gives the fixed effects under consideration. Columns 3 and 4 give the standard FE model estimate of the treatment effect and its standard error. Columns 5 and 6 give the sample-weighted estimate from an interacted model and its standard error. The final column gives the percent difference between the FE and SWE estimates.
Table 2.12: Detailed replication results for Banerjee and Iyer (2005), continued

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Fixed effect</th>
<th>FE est.</th>
<th>FE SE</th>
<th>SWE</th>
<th>SWE SE</th>
<th>% Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prop. Cereals</td>
<td>Soil — red</td>
<td>0.06</td>
<td>0.03</td>
<td>0.05</td>
<td>0.03</td>
<td>-17.1</td>
</tr>
<tr>
<td></td>
<td>Soil — black</td>
<td>0.06</td>
<td>0.03</td>
<td>0.06</td>
<td>0.03</td>
<td>-0.2</td>
</tr>
<tr>
<td></td>
<td>Soil — all</td>
<td>0.06</td>
<td>0.03</td>
<td>0.06</td>
<td>0.03</td>
<td>6.6</td>
</tr>
<tr>
<td></td>
<td>Coastal</td>
<td>0.06</td>
<td>0.03</td>
<td>0.06</td>
<td>0.03</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>0.06</td>
<td>0.03</td>
<td>0.06</td>
<td>0.03</td>
<td>0.1</td>
</tr>
<tr>
<td>Prop. HYV rice</td>
<td>Soil — red</td>
<td>0.08</td>
<td>0.04</td>
<td>0.08</td>
<td>0.05</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>Soil — black</td>
<td>0.08</td>
<td>0.04</td>
<td>0.08</td>
<td>0.04</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>Soil — all</td>
<td>0.08</td>
<td>0.04</td>
<td>0.08</td>
<td>0.04</td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td>Coastal</td>
<td>0.08</td>
<td>0.04</td>
<td>0.08</td>
<td>0.04</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>0.08</td>
<td>0.04</td>
<td>0.08</td>
<td>0.04</td>
<td>-0.2</td>
</tr>
<tr>
<td>Prop. HYV wheat</td>
<td>Soil — red</td>
<td>0.09</td>
<td>0.05</td>
<td>0.07</td>
<td>0.05</td>
<td>-20.5</td>
</tr>
<tr>
<td></td>
<td>Soil — black</td>
<td>0.09</td>
<td>0.05</td>
<td>0.07</td>
<td>0.05</td>
<td>-18.3</td>
</tr>
<tr>
<td></td>
<td>Soil — all</td>
<td>0.09</td>
<td>0.05</td>
<td>0.09</td>
<td>0.05</td>
<td>3.3</td>
</tr>
<tr>
<td></td>
<td>Coastal</td>
<td>0.09</td>
<td>0.05</td>
<td>0.09</td>
<td>0.04</td>
<td>-1.5</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>0.09</td>
<td>0.05</td>
<td>0.09</td>
<td>0.05</td>
<td>0.6</td>
</tr>
<tr>
<td>Prop. Irrigated</td>
<td>Soil — red</td>
<td>0.07</td>
<td>0.03</td>
<td>0.08</td>
<td>0.03</td>
<td>19.5</td>
</tr>
<tr>
<td></td>
<td>Soil — black</td>
<td>0.07</td>
<td>0.03</td>
<td>0.07</td>
<td>0.04</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>Soil — all</td>
<td>0.07</td>
<td>0.03</td>
<td>0.07</td>
<td>0.03</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>Coastal</td>
<td>0.07</td>
<td>0.03</td>
<td>0.06</td>
<td>0.03</td>
<td>-1.1</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>0.07</td>
<td>0.03</td>
<td>0.07</td>
<td>0.03</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Notes: Column 1 gives the paper and column 2 gives the fixed effects under consideration. Columns 3 and 4 give the standard FE model estimate of the treatment effect and its standard error. Columns 5 and 6 give the sample-weighted estimate from an interacted model and its standard error. The final column gives the percent difference between the FE and SWE estimates.
Table 2.13: Detailed replication results for Banerjee and Iyer (2005), continued

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Fixed effect</th>
<th>Joint Test of interactions</th>
<th>Test of treat. diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Wald stat.</td>
<td>DF</td>
</tr>
<tr>
<td>Prop. Fertilized</td>
<td>Soil — red</td>
<td>4.52</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Soil — black</td>
<td>0.06</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Soil — all</td>
<td>0.04</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Coastal</td>
<td>0.28</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>124.52</td>
<td>31</td>
</tr>
<tr>
<td>Log yield</td>
<td>Soil — red</td>
<td>2.06</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Soil — black</td>
<td>0.14</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Soil — all</td>
<td>3.48</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Coastal</td>
<td>1.16</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>274.22</td>
<td>31</td>
</tr>
<tr>
<td>Log rice yield</td>
<td>Soil — red</td>
<td>0.40</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Soil — black</td>
<td>1.19</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Soil — all</td>
<td>6.29</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Coastal</td>
<td>1.31</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>171.87</td>
<td>31</td>
</tr>
<tr>
<td>Log wheat yield</td>
<td>Soil — red</td>
<td>1.04</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Soil — black</td>
<td>0.47</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Soil — all</td>
<td>6.97</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Coastal</td>
<td>3.05</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>117.86</td>
<td>31</td>
</tr>
</tbody>
</table>

Notes: Column 1 gives the paper and column 2 gives the fixed effects under consideration. Column 3 gives the Wald statistic of a joint test of the significance of the interactions, column 4 gives the degrees of freedom for that test, and column 5 gives the p-value. Column 6 gives a t statistic from a test of the difference between the FE and sample-weighted estimates using the test derived in Appendix 2.A.2 and the corresponding p-value.
Table 2.14: Detailed replication results for Banerjee and Iyer (2005), continued

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Fixed effect</th>
<th>Joint Test of interactions</th>
<th>Test of treat. diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Wald stat.</td>
<td>DF</td>
</tr>
<tr>
<td>Prop. Cereals</td>
<td>Soil — red</td>
<td>3.09</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Soil — black</td>
<td>0.00</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Soil — all</td>
<td>4.97</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Coastal</td>
<td>0.05</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>78.04</td>
<td>22</td>
</tr>
<tr>
<td>Prop. HYV rice</td>
<td>Soil — red</td>
<td>0.01</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Soil — black</td>
<td>0.04</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Soil — all</td>
<td>1.05</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Coastal</td>
<td>0.12</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>108.78</td>
<td>22</td>
</tr>
<tr>
<td>Prop. HYV wheat</td>
<td>Soil — red</td>
<td>6.31</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Soil — black</td>
<td>8.02</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Soil — all</td>
<td>2.64</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Coastal</td>
<td>7.58</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>179.01</td>
<td>22</td>
</tr>
<tr>
<td>Prop. Irrigated</td>
<td>Soil — red</td>
<td>3.07</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Soil — black</td>
<td>0.75</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Soil — all</td>
<td>0.35</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Coastal</td>
<td>1.43</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>84.84</td>
<td>26</td>
</tr>
</tbody>
</table>

Notes: Column 1 gives the paper and column 2 gives the fixed effects under consideration. Column 3 gives the Wald statistic of a joint test of the significance of the interactions, column 4 gives the degrees of freedom for that test, and column 5 gives the \( p \)-value. Column 6 gives a \( t \) statistic from a test of the difference between the FE and sample-weighted estimates using the test derived in Appendix 2.A.2 and the corresponding \( p \)-value.
Chapter 3

LATE for School: Instrumental Variables and the Returns to Education

with Michael B. Urbancic
Department of Economics
University of California, Berkeley

3.1 Introduction

It is common in the applied literature to compare estimates from OLS and instrumental variables (IV) models and to use similarity of these results as proof of exogeneity. The logic behind this reasoning, however, is flawed. We offer graphical and closed-form examples illustrating this mistake. Additionally, we consider the implications of this insight for the Hausman test in the IV context. While our results do not contain new econometric theory, we provide evidence from several key papers in the returns to education literature that demonstrate a disconnect between existing theory and its application. Our aim is to elucidate this common mistake in the empirical literature and to offer guidelines for the applied researcher.

The IV approach estimates the local average treatment effect (LATE), rather than the average treatment effect (ATE); “local” refers to the subpopulation that changes its behavior in response to an exogenous change in the value of the instrument. This group is known as the compliers and its composition is instrument-dependent and unobservable. In the absence of endogeneity, under the necessary assumptions, both OLS and IV estimate the ATE. It has become common practice to compare IV estimates to those from OLS and to use equality of these estimates as evidence of exogeneity. Unfortunately, this comparison is misguided.
Chapter 3. LATE for School: Instrumental Variables and the Returns to Education

As we show, econometric theory states that, if there is no endogeneity, then the LATE equals the ATE. Comparisons of IV to OLS estimates rest upon the converse of this assertion, a proposition which is not necessarily true. The converse is true if the researcher assumes homogeneous treatment responses across unobservable characteristics, but this assumption is rarely stated or justified. The focus of this chapter is to expose and illustrate this common fallacy and provide guidelines for the proper interpretation of IV estimates.

We consider this logical error both theoretically and empirically. We first supply a simple example illustrating how equality between the LATE and ATE may arise under endogeneity. Regression discontinuity provides a framework familiar to economists that we use to clearly illustrate the identification of a local effect. As a final source of intuition, we offer a graphical examination of the relationship between LATE and ATE and how equality might arise. In order to formalize our argument, we use the results of [Vytlacil 2002] and [Heckman and Vytlacil 2005] to demonstrate the error in comparing the ATE and LATE in the contexts of point estimation and in the estimation of distributional treatment effects. This formal derivation illustrates that heterogeneity of treatment responses conditional upon the observed covariates is the source of the false converse.

We then show that the use of the Hausman test to consider exogeneity is mistaken. Authors typically provide Hausman statistics as a test for a null hypothesis of exogeneity, but this null hypothesis is not and cannot be tested. Instead, equality between the LATE and ATE is tested, which could arise under exogeneity or due to the happenstance of LATE from an endogenous subpopulation equaling the population average treatment effect. We show, however, that if the null hypothesis of equality can be rejected, this can provide evidence of endogeneity.

Turning to an empirical example, we show that the distinction between LATE and ATE is often overlooked by researchers. Specifically, we use the instrumental variables approach to estimate both the private and social returns to education. The literature on this subject provides a unique opportunity to examine the relationship between the LATE and the ATE. First, the topic has been examined by several important papers that guide our inquiry; the influential works of [Acemoglu and Angrist 2000] and [Lochner and Moretti 2004] serve as touchstones that allow us to illustrate the LATE in practice. Second, large data sets like the U.S. Census are readily available and quite familiar to the applied economist. Lastly, unique to the education literature, three different instruments have been found that may identify the effect of education in the presence of endogeneity. These qualities combine to provide an excellent empirical illustration of the logical flaw that we expose.

In Section 3.2 we begin by offering a simple example illustrating the difference in identifying the LATE and the ATE. An explanation in the regression discontinuity context provides a formulation familiar to the applied economist. A graphical example provides a third intuitive illustration. Then, in Section 3.3 we use the definitions of the LATE and ATE to expose this mistake analytically in estimating single parameters and distributional effects. We show that, if there is no endogeneity, then the LATE equals the ATE, but the
converse is false; if the LATE equals the ATE, that does not imply that endogeneity is absent. The interpretation of the Hausman test in the IV framework follows this false converse; we discuss this point in Section \ref{sec:3.4} In Section \ref{sec:3.5}, we discuss how applied economists approach the problem of endogeneity, with specific focus on the application of IV to the returns to education literature. We use this literature to illustrate the issue empirically in Section \ref{sec:3.6} Lastly, we conclude by offering guidance for the practitioner in Section \ref{sec:3.7}.

### 3.2 Simple Examples of LATE

In the absence of endogeneity, OLS models deliver the average treatment effect (ATE). With valid instruments, IV models provide the local average treatment effect (LATE) for those individuals for whom the instrument is binding (e.g., those that change their educational attainment in response to changes in compulsory schooling laws, holding their observed characteristics fixed), given the necessary monotonicity assumption \cite{Imbens1994}.\footnote{Specifically, the monotonicity assumption states that, if $D_i(z)$ indicates the treatment that individual $i$ chooses in response to instrument $z$, then either $D_i(z) \geq D_i(z') \forall i$ or $D_i(z) \leq D_i(z') \forall i \forall z'$. Note that we also assume correct specification in both the OLS and IV cases and that the regularity conditions necessary for each model are met. We discuss the assumptions needed for an instrument to be valid in Footnote \ref{fn:instrument_validity}.}

Note that, because the identifying population is instrument-dependent, the LATE itself is identified only for this unobservable and unidentifiable subset of the population. Some authors attempt to characterize this hidden subpopulation \cite[e.g.,][]{Kling2001}. This characterization is based upon observable characteristics, while the heterogeneity that we consider here is over unobserved covariates. Consider the returns to education case. Knowing the composition of the compliers relative to race or other demographic factors does not suggest anything about the responses of the members of this group to treatment relative to unobserved factors like ability or intelligence. It is heterogeneity over the latter set of characteristics that complicates comparisons of LATE and ATE\footnote{A separate literature addresses the relevance of this subpopulation for policy analysis \cite[see, e.g.,][]{Heckman1997}.}. Hence, if the returns to education is a function of ability, a likely case, then the error that we highlight arises. Deriving the demographic make-up of the compliers does not ameliorate this issue. If endogeneity is not present or there is a homogeneous treatment effect, then the LATE will equal the ATE; if treatment choices are endogenous and the responses are heterogeneous, then the LATE may or may not equal the ATE.

#### 3.2.1 A LATE Example: Regression Discontinuity

As has already been mentioned, the “local” interpretation of the IV estimator arises because this parameter is only identified for an unknown subgroup of the sample. This interpretation
may not be entirely intuitive. The regression discontinuity (RD) framework provides a clearer example of a local effect. Here, the probability of receiving treatment changes discontinuously at some threshold along a value scale. While the threshold is determined exogenously, the location of an individual on the value scale may be endogenous. Under either homogeneity of treatment responses or a selection-on-observables assumption, RD identifies the ATE. RD can be used if the researcher assumes that selection in the neighborhood of the threshold is minimal. This belief relies upon assumed random variation in the precise location of an individual along the value scale; individuals near the threshold have roughly similar probabilities of landing above or below the cutoff, suggesting that individuals just above and just below are similar in terms of observed and unobserved characteristics.

Notice the use of “neighborhood,” “near,” and “just” above or below—RD estimates in the presence of endogeneity identify the treatment effect only for those individuals near the threshold\(^3\). Hence, RD identifies a local average treatment effect, a LATE \cite{Hahn2001}. The local nature is clearer here than in the IV context because the researcher limits the sample to a window around the threshold and this entire subsample identifies the LATE; here, the subgroup for which the LATE is identified is clear and known. But the subgroup defining the LATE in the IV framework (the compliers) is unknown and unobservable. This demonstrates that the LATE generated by RD has a clearer interpretation and better informs its applicability to policy questions relative to a LATE produced by IV.

### 3.2.2 A Simple Example

Consider the following simple illustration of the relationship between the LATE and the ATE. Imagine that a population of interest is divided into three groups based upon their treatment responses, each composed of otherwise identical individuals. One group has a high treatment response, one a low response, and one an average response. Individuals endogenously select into treatment based upon their personal benefits and costs of doing so. An individual’s group membership is unobserved by the researcher. Since the LATE is instrument-specific and defined for some (unknown) population, in this example it can take a high, low, or average value depending upon the instrument. If the researcher happens to estimate the LATE for the average response group, then the LATE will equal the ATE. This result arises by happenstance. Obviously, equality here does not prove the absence of endogeneity. This simple situation shows that the equality of the LATE and ATE does not imply exogeneity of treatment.

\(^3\)There is no theoretical basis for defining “near” precisely, but \cite{Lee2008} emphasizes that the “auxiliary prediction” of balance on the observed covariates should hold for those on either side of the threshold.
3.2.3 A Graphical Comparison of LATE and ATE

As a prelude to a more formal exposition below, we can relate LATE and ATE in a simplified graphical context. Suppose that there is a one-dimensional unobserved covariate that elicits a heterogeneous response to treatment; the relationship between this covariate and treatment is the solid curve in Figure 3.1. Assume further that individuals in the population are distributed uniformly across this dimension and self-select treatment levels. Then, the average treatment effect (ATE) for the population is equal to the slope of the line connecting the endpoints of the response curve; this is line $A$ in the figure. The marginal treatment effect (MTE) for an individual with a given value of the unobserved covariate is the slope of the response curve at the corresponding level. In the figure, we see that the MTE for an individual with level $\tilde{u}$ of the unobserved covariate is equal to the ATE, as line $M$ has the same slope as $A$.

Suppose that an instrument is identified by individuals with unobserved covariate levels between points $u_L$ and $u_H$ and that compliers are evenly distributed in this range. Here, the LATE is the slope of the line connecting the values of the curve at the extreme ends of this range. Hence, the LATE for this instrument is the slope of line $L$, which is equal to the ATE. As we can see, there are a number of ways that the LATE can equal the ATE despite the endogeneity and heterogeneity of treatment responses.

Figure 3.1: The response curve to treatment with examples of a MTE and LATE equal to the ATE
Chapter 3. LATE for School: Instrumental Variables and the Returns to Education

3.3 A Formal Comparison of LATE and ATE

In this section, we refine the distinction between the LATE and ATE and formalize the arguments outlined above. We offer definitions and relationships between the ATE, LATE, and MTE. We then present a more precise, numerical example of the preceding simplified version to demonstrate that equality of IV and OLS estimates does not imply a lack of endogeneity. A second example elucidates the point in the context of estimating distributional treatment effects.

Imbens and Angrist (1994) present conditions under which instrumental variables identify the LATE. Their assumptions relate the interpretation of instrumental variables estimates to the experimental framework and consider heterogeneous treatment effects. In a recent paper, Vytlacil (2002) demonstrates that these assumptions are logically equivalent to those of a latent index model. In this model, treatment is triggered by an unobservable index surpassing a threshold. In particular, given the assumptions that identify the LATE, “there always exists a selection model that rationalizes the observed and counterfactual data” Vytlacil (2002). Further, Heckman and Vytlacil (2005) demonstrate that the LATE and the ATE can be written as weighted averages of the marginal treatment effects (MTE) over different (endogenous) subsets of the sample. Using these results and an index model framework, we clarify the perils of comparing ATE and LATE.

For simplicity, we consider the situation of a binary treatment. Define the potential outcomes under the treatment, $Y_1$, and control regimes, $Y_0$, as functions of observable covariates $X$ and unobservable factors in the treatment, $U_1$, and control, $U_0$, states, given by

$$Y_1 = \mu_1(X, U_1) \quad \text{and} \quad Y_0 = \mu_0(X, U_0).$$

An individual will choose to accept treatment if $Y_1 \geq Y_0$. Since $U_1$ and $U_0$ are unobservable, we cannot predict this choice directly. Instead, we create a new model of the value of treatment, $D^*$, which is a function of the observable covariates and instruments $Z$, but not of any unobserved factors. Let $D$ equal 1 if the individual accepts treatment and 0 if he does not. This formulation implies the following latent index model:

$$D^* = \mu_D(Z) - U_D$$

$$D = \begin{cases} 
1 & \text{if } D^* \geq 0 \\
0 & \text{otherwise.} 
\end{cases}$$

$U_D$ can be a function of $U_0$ and $U_1$, e.g., $U_D = U_0 - U_1$. This variable can be viewed as an individual’s unobserved value (or utility) from not taking treatment. Given the equivalence result of Vytlacil (2002), endogeneity is characterized as individual choice informed by from his knowledge of the unobserved utility $U_D$.

Without loss of generality, $U_D$ is normalized to be distributed Uniform[0,1]. Given
that $\mu_D$ is an arbitrary and unknown function on $[0, 1]$ and as a result of our normalization of $U_D$, $\mu_D$ is an individual’s $p$-score—the probability of taking up treatment conditional on $Z$ and $X$. To see this, note that

$$\Pr(D = 1 | X, Z) = \Pr(D^* \geq 0 | X, Z) = \Pr(\mu_D(Z) - U_D \geq 0 | X, Z) = \Pr(U_D \leq \mu_D(Z) | X, Z) = \mu_D(Z),$$

where the last equivalence follows from the uniform assumption on $U_D$ and the assumption that $\mu_D(Z)$ is conditional on $X$.

This formulation shows that treatment status $D$ depends upon both the observable instruments $Z$ and the unobservable element $U_D$. Instead, suppose that there were no unobservable aspects of an individual’s choice in this model. Then, again without loss of generality, $U_D$ would be a degenerate random variable equal to 0. The probability line above would yield a $p$-score that is simply 0 or 1 depending upon the value of $\mu_D(Z)$, a function of only observed covariates. If there is no unobserved component in the problem—i.e., there is no endogeneity—treatment assignment can be perfectly predicted based upon the value of observed covariates.

Write the observed outcome as $Y = Dy_1 + (1 - D)y_0$. Assuming that $Z$ is a set of valid instrumental variables and under the necessary regularity conditions, Vytlacil (2002) shows the model above is equivalent to the assumptions that Imbens and Angrist (1994) supply to identity the LATE.\(^4\)

Treatment effects can now be analyzed within this structure. Let $\Delta = y_1 - y_0$. The assumptions, as stated in Heckman and Vytlacil (2005), are:

1. $\mu_D(Z)$ is a non-trivial function of $Z$ conditional on $X$ (inclusion restriction of instrument $Z$).
2. $(U_0, U_1, U_D)$ are independent of $Z$ conditional on $X$ (exclusion restriction of instrument $Z$).
3. $U_D$ has an absolutely continuous distribution with respect to Lebesgue measure.
4. $\mathbb{E}[Y_1]$ and $\mathbb{E}[Y_0]$ are finite.
5. $0 < \Pr(D = 1 | X) < 1$ (there are both treatment and control groups).
6. $X_1 = X_0$ a.e. (the treatment and control groups have common support over exogenous parameters).
Chapter 3. LATE for School: Instrumental Variables and the Returns to Education

ATE (conditional on \( X \)) is defined by:

\[
\Delta^{ATE}(x) \equiv \mathbb{E}[\Delta | X = x].
\]

This causal effect could be identified if treatment could be randomly assigned among individuals with \( X = x \), under full compliance, and in the absence of general equilibrium effects. Under these assumptions, OLS estimates the ATE.

The LATE under instrument \( z \) is defined as the average treatment effect for individuals who did not undertake treatment under the \( z' \) regime, but do take treatment under \( z \), conditional upon \( X \); members of this group change their behavior in response to an (exogenous) change in the instrument and are known as “compliers” \cite{Angrist1996}. Let \( z, z' \) be two realizations of the instrument. Given this formulation, the LATE can be written as

\[
\Delta^{LATE}(x, \mu_D(z), \mu_D(z')) \equiv \mathbb{E}[\Delta | X = x, D_z = 1, D_{z'} = 0] = \mathbb{E}[\Delta | X = x, \mu_D(z') < U_D \leq \mu_D(z)].
\]

The second line uses the following logic: if an individual did not take treatment under \( z' \), then his value of the alternative to treatment, \( U_D \), must be greater than the value of taking that treatment conditional on the instrument \( z' \), which is \( \mu_D(z') \); the reverse logic implies that \( U_D \leq \mu_D(z) \), yielding the given expression.

This notation emphasizes that the LATE is an expectation taken over a distribution of unobserved characteristics \( U_D \) defined by the instruments \( z \) and \( z' \). This subpopulation is the set of compliers, defined as \( \{U_D : \mu_D(z') < U_D \leq \mu_D(z)\} \). Since \( U_D \) is unobserved, the set of compliers is also unobserved. Lastly, a different pair of instruments would change the set of compliers and, therefore, the LATE. Now it becomes clear that the characterization of the compliers using demographic factors does not highlight the heterogeneity present in the LATE; the heterogeneity here stems from unobserved factors.

We can define the MTE by taking the limit of the LATE as \( \mu_D(z) \rightarrow u_D \) (under the regularity conditions imposed above), giving

\[
\Delta^{MTE}(x, u_D) \equiv \lim_{\mu_D(z) \rightarrow u_D} \Delta^{LATE}(x, u_D, \mu_D(z)) = \mathbb{E}[\Delta | X = x, U_D = u].
\]

This is simply the average treatment effect for individuals with observable characteristics \( x \)

\footnote{When we typically think of the ATE or LATE, the parameter is typically unconditional, \textit{i.e.}, integrated over the set of \( X \) covariates. Hence, our usual conception for the ATE is

\[
ATE = \int_{x \in X} \Delta^{ATE}(x) dF(x)
\]

We maintain the definition above for simplicity and concreteness.}
and unobservable characteristics $u_D$. If $\Delta$ is a value measure, then the MTE is an individual’s willingness-to-pay for treatment given his observed and unobserved characteristics.\footnote{Some authors argue that the MTE is the desired parameter of interest because it permits empirical examinations of the relationship between marginal benefits and costs. It is precisely these values that economic theory uses to design optimal policies. \cite{HeckmanV} and \cite{HeckmanV2} pursue these considerations, but are beyond the scope of our exposition.}

At last we are able to express the relationship between the three treatment effects. Specifically, the ATE and LATE can be written

$$\Delta^{ATE}(x) = \int_0^1 \Delta^{MTE}(x, u_D) \, du_D, \quad \text{and}$$

$$\Delta^{LATE}(x, \mu_D(z), \mu_D(z')) = \frac{1}{\mu_D(z) - \mu_D(z')} \int_{\mu_D(z')}^{\mu_D(z)} \Delta^{MTE}(x, u_D) \, du_D. \quad (3.2)$$

These equations elucidate several important points. First, the LATE is instrument-dependent, while the ATE is not. The limits of integration of the LATE (i.e., the set of compliers) and a multiplicative factor are both functions of the instruments, thus making the parameter itself dependent upon the instrument.

Suppose that there is no endogeneity in this framework. Then, $\Delta^{MTE}(x, u_D) = \Delta^{MTE}(x)$. Hence,

$$\Delta^{ATE}(x) = \int_0^1 \Delta^{MTE}(x, u_D) \, du_D = \Delta^{MTE}(x) \int_0^1 du_D = \Delta^{MTE}(x), \quad \text{and}$$

$$\Delta^{LATE}(x, \mu_D(z), \mu_D(z')) = \frac{1}{\mu_D(z) - \mu_D(z')} \int_{\mu_D(z')}^{\mu_D(z)} \Delta^{MTE}(x, u_D) \, du_D$$

$$= \Delta^{MTE}(x) \frac{1}{\mu_D(z) - \mu_D(z')} \int_{\mu_D(z')}^{\mu_D(z)} du_D = \Delta^{MTE}(x).$$

This result would also hold if there were homogeneous responses to treatment conditional upon $x$: i.e.,

$$\Delta^{MTE}(x, u_D) = \Delta^{MTE}(x) \forall u_D.$$

Here again we see that the relevant heterogeneity is over unobservables.
A second observation is that, if there is no endogeneity, then
$$\Delta^{ATE}(x) = \Delta^{LATE}(x) = \Delta^{MTE}(x).$$

We reiterate, however, that the converse is not true—specifically, if the LATE equals the ATE, it *does not follow* that there is no endogeneity. It is conceivable that Equation [3.1] equals Equation [3.2] while the latter is still integrated over an endogenous subpopulation. Examples of this occurrence follow below.

As an aside, the following intuition may help explain the relationship between the LATE and the ATE. If there is no endogeneity bias present, then, setting the LATE equal to the ATE, we see that every individual in the sample can be considered a complier; i.e., $$\mu_D(z') = 0$$ and $$\mu_D(z) = 1.$$ Intuitively, if there is no endogeneity in the model, then the best instrument of treatment is treatment itself. Since treatment is exogenous, it is independent of unobserved characteristics and thus it is a valid instrument. Changing the instrument, which is treatment itself, obviously changes the individual’s treatment and thus every individual is tautologically a complier.7 These results demonstrate that exogenously-determined treatment is a valid instrument and that all individuals are compliers. This, of course, greatly simplifies the results above.

This formalization aims to clarify the distinction that we make between the LATE and the ATE. After providing a few examples to illustrate this result concretely, we move to an examination of the empirical literature on the returns to education. The theoretical results outlined above illustrate the reasons why the three common instruments employed in the education literature—compulsory schooling laws, minimum employment ages, and quarter-of-birth—should not be expected to yield the same parameter estimates if educational attainment is an endogenous choice and responses to education are heterogeneous over unobserved characteristics.

**Example 1: Comparing LATE and ATE**

Consider an estimation problem where the LATE is identified following the assumptions above and the MTE has the following form:

$$\Delta^{MTE}(x, u_D) = \begin{cases} 
2\% & \text{if } u_D \leq \frac{1}{3}, \\
1\% & \text{if } u_D \in \left(\frac{1}{3}, \frac{2}{3}\right), \\
0\% & \text{if } u_D \geq \frac{2}{3}.
\end{cases}$$

Also assume that individuals are uniformly distributed over $$u_D \in [0, 1].$$ Treatment here could be, as an example, a job training program that assists those individuals with a low

---

7Here we are assuming that treatment received is exogenous and can be measured. As is well known, treatment may be randomly assigned, but compliance or take-up may be endogenous. In this case, treatment is not truly exogenous.
Chapter 3. LATE for School: Instrumental Variables and the Returns to Education

opportunity cost of treatment $u_D$ more than those with high-valued alternatives. Now, suppose that the chosen instrument generates compliers consisting of those individuals with $u_D \in \left(\frac{1}{3}, \frac{2}{3}\right)$. It then follows that

$$\Delta^{ATE}(x) = \int_0^1 \Delta^{MTE}(x, u_D) du_D = 1\% \quad \text{and}$$

$$\Delta^{LATE}(x) = \frac{1}{\frac{2}{3} - \frac{1}{3}} \int_{\frac{1}{3}}^{\frac{2}{3}} \Delta^{MTE}(x, u_D) du_D = 1\%.$$ 

The LATE equals the ATE, yet the effect of treatment is not exogenous—it is defined by the unobserved factor $u_D$. Hence, relying upon the equality between the LATE and ATE to conclude that endogeneity is absent from this model is incorrect. This leap is commonly made throughout the applied econometrics literature, despite it being erroneous.

Example 2: Distributional Treatment Effects

The result in the previous example arises because the treatment effect varies with unobserved factors. Here we offer an analogous example that estimates not averages of outcomes, but rather distributions of outcomes. We discuss the proper interpretation imparted to these estimates.

The literature on non-parametric estimation of treatment effects has recently explored the identification of distributions of potential outcomes (see, e.g., Imbens and Rubin [1997]; Abadie [2002]; Abadie, Angrist and Imbens [2002]). Abadie (2002) demonstrates that the same assumptions employed by Imbens and Angrist (1994) to identify the LATE can be generalized to identify functions of potential outcomes. Specifically, by using an instrumental variables approach, he is able to identify the entire cumulative distribution function (CDF) of potential outcomes under treatment for the set of compliers.

Consider the following example where individuals self-select into a treatment regime by choosing a level of education, for example. Assume that the CDF of potential outcomes (e.g., income) both prior to treatment and under the control regime is

$$F_0(y|u_D) = \frac{1}{100} \times y \quad \forall u_D \in [0, 100],$$

where $y$ is income measured in thousands of dollars. In words, the income distribution is uniform and bounded between $0$ and $100,000$. Additionally, assume that the marginal
effect of treatment is proportional to the individual’s pre-treatment income, \( y_0 \):

\[
\Delta^{\text{MTE}}(y_0, u_D) = \begin{cases} 
3y_0 & \text{if } u_D \leq \frac{1}{3}, \\
1.5y_0 & \text{if } u_D \in (\frac{1}{3}, \frac{2}{3}), \\
y_0 & \text{if } u_D \geq \frac{2}{3}.
\end{cases}
\]

Lastly, assume the existence of an instrument that identifies potential outcomes for the subset of the population with \( u_D \in (\frac{1}{3}, \frac{2}{3}) \).

Now, we can apply the instrumental variables method of Abadie (2002) to estimate the distribution of income of the compliers under the treatment regime. In his paper, he divides the sample based upon the value of the instrument for each individual and calculates the empirical CDF of the outcomes for each value of the instrument. Then, he develops a bootstrapping strategy to test the hypothesis of equivalence of these distributions. In our example, we would get an estimate of the CDF for the compliers, which is:

\[
F_1(y_1 \mid u_D \in (\frac{1}{3}, \frac{2}{3})) = F_0(1.5y_0)
\]

\[
= \Pr\left(\frac{3}{2}y_0 \leq y\right) = \Pr\left(y_0 \leq \frac{2}{3}y\right)
\]

\[
= \frac{1}{100} \times \frac{2y}{3} = \frac{y}{150}.
\]

The first equality follows because \( F_0 \) is the distribution of income regardless of the unobserved covariates. Here we see an estimate that is analogous to the LATE, though it estimates an entire CDF, rather than an average. We can create an estimate of the CDF that is analogous to the ATE:

\[
F_1(y) = \frac{1}{3} \frac{y}{100} + \frac{1}{3} \frac{y}{150} + \frac{1}{3} \frac{y}{300} = \frac{y}{150}.
\]

This problem illustrates precisely the same issue raised in the prior example. Now, the CDF of the potential outcomes for the compliers (a “local” CDF, perhaps) is the same as the function for all the treated individuals, which ignores the presence of endogeneity. Again, in this case, we must resist the temptation to believe that this equality falsifies claims of endogeneity. This example shows how the logic of our inquiry applies to not only average treatment effects, but also to functions of outcomes in the presence of endogeneity more generally.

---

8 For expositional purposes, he describes the procedure using a binary instrument. See his paper for additional details.
3.4 Hausman Specification Tests

In applied econometrics, it is common to compare two estimators with the same asymptotic limit as a specification test of the estimation strategy. In a classic paper, Hausman (1978) outlines the following procedure. First, choose two consistent, asymptotically-normal estimates of a parameter (vector) $\beta$; the first, $\hat{\beta}_0$, must be efficient (in that it achieves the Cramér-Rao Lower Bound) under the null hypothesis of correct specification, but, under the alternative hypothesis, may be biased. The second estimate, $\hat{\beta}_1$, must be consistent under both the null and alternative hypotheses, but may not be efficient. Then, $\hat{\beta}_0$ has a zero covariance (matrix) with the difference between the estimates, $\hat{q} = \hat{\beta}_1 - \hat{\beta}_0$. Now, the familiar Hausman test of $H_0 : \hat{q} = 0$ can be created.

Many authors in applied econometrics have used the Hausman test to determine the presence of endogeneity by comparing an OLS estimate, which is the efficient estimate under the null hypothesis of exogeneity, to an IV estimate, which is the estimate that is robust to the proposed alternative hypothesis of endogeneity. If $\beta$ is truly exogenous, then the OLS and IV estimates have the same asymptotic limit; this is because the LATE equals the ATE under exogeneity. This implies that, if the parameters are exogenous, then you cannot reject the hypothesis that the difference between the OLS and IV estimates is 0. The contrapositive is also true; namely, if you can reject the hypothesis that the LATE equals the ATE, then endogeneity is present. But the converse is not true; if you cannot reject the null hypothesis, then it is not necessarily the case that the parameters are exogenous. Indeed, this could be another situation, like Example 1 above, in which the LATE equals the ATE despite the presence of endogeneity.

Note that this is a matter entirely separate from the type II error of erroneously failing to reject the null hypothesis. Even if a test of power 1 could be created, failing to reject the null hypothesis does not imply exogeneity. This is because the null hypothesis being tested is not exogeneity, but rather the equality of the LATE and the ATE. This equality could arise under three circumstances:

1. There is no endogeneity bias (the conclusion that the researcher would like to draw from failure to reject parameter equality).
2. There is insufficient power to distinguish between the two parameters, which are, in fact different (i.e., a type II error is made).
3. The LATE over an endogenous set of compliers happens to equal the ATE for the entire population (a chance equality between two distinct parameters).

The researcher would like to make the simple claim of (1), accepting (2) as a necessary consequence of statistical estimation. But this assertion ignores the possibility of (3), an
entirely separate concern that confounds the analysis. In summary, if the Hausman test leads you to reject the null hypothesis that the LATE equals the ATE, then endogeneity is present. But, if you cannot reject equality of these two parameters, then you cannot conclude that the covariates are exogenous. This latter conclusion is another manifestation of the confusion incumbent in comparing the LATE and the ATE.

### 3.5 Causal Effects and the Returns to Education

The returns to education literature considered here attempts to determine the effect of education on wages and incarceration. Ideally, this calculation would compare the outcomes under both the treatment and its absence—the treatment being an additional year of education. Alas, we cannot observe both of these outcomes for each student and the outcome that is observed is based upon endogenous treatment choices. As posited above, differences in mandatory schooling policies may be exogenous to an individual’s attainment choice. The use of this condition as a treatment, however, would require identifying precisely those students for whom the treatment was binding (i.e., those for which the laws changed their behavior). While we cannot explicitly identify these individuals, this unknown population is used by the instrumental variables approach to identify the LATE (see Section 3.2).

Acemoglu and Angrist (2000) and Lochner and Moretti (2004) employ compulsory schooling age as an instrument for educational choice. The mandatory age for schooling has fluctuated in the U.S. across states and time and this variation can be used as an exogenous factor influencing educational attainment decisions. A similar instrument is minimum employment age laws Acemoglu and Angrist (2000). Rather than compelling students to attend school for a certain period of time, these laws foreclose the most promising alternative to school—work—and therefore reduce the opportunity cost of school attendance. These laws, too, vary across states and time.

Angrist and Krueger (1991) use a third instrument for educational attainment. Compulsory schooling laws often require students to attend school until they reach a certain age, rather than complete a certain number of years of education. The same laws require that students begin schooling at a certain age. As a result, students with a particular birth date are systematically older or younger relative to their peers in that grade. Because education requirements are in terms of age, rather than in terms of years of enrollment, older students can drop out with lower educational attainment. Indeed, Angrist and Krueger find that older students in a given grade level were less likely to complete the grade.12

10 Though it is true that you cannot reject the hypothesis of exogeneity.
11 See, e.g., Lleras-Muney (2005) for additional applications of this instrument.
12 Angrist and Krueger find a “small but persistent pattern” in educational attainment by quarter of birth. This seasonal pattern is only present through the twelfth grade and was not evident in levels of higher
Our empirical analysis compares the estimates obtained using this trio of instruments. We state immediately that we do not expect all these instruments to yield identical parameter estimates. Different “local” estimates arise from different instruments if students have heterogeneous returns to education. By examining the results generated by different instruments on our data, we show that the parameter estimates are indeed instrument-dependent and, while one instrument may produce estimates similar to those arising from an OLS model, that does not imply the absence of endogeneity.

3.5.1 LATE v. ATE in the Returns to Education Literature

Here we demonstrate how the logical flaw that we highlight manifests itself in the empirical literature. Specifically, these quotes suggest that, if the IV and OLS estimates are equal, then there is no endogeneity. Yet, as we have seen, this assertion is false. For example, Lochner and Moretti (2004) state that

[t]hese [instrumental variable] estimates are stable across specifications and nearly identical to the corresponding OLS estimates... This indicates that the endogeneity bias is not quantitatively important after controlling for age, time, state of residence and state of birth.

In analyzing the private returns to education and hypothesizing about education externalities, Acemoglu and Angrist (2000) claim that

it is noteworthy that the IV estimates using quarter of birth are very close to the OLS estimates for the same period... Thus, estimates of external returns that treat individual schooling as exogenous and endogenous should give similar results, at least for [that period].

These pronouncements follow the converse of the results shown in Section 3.3, however, and are not necessarily true. Lastly, Angrist and Krueger (1991) state that

[u]sing season of birth as an instrument for education in an earnings equation, we find a remarkable similarity between the OLS and the TSLS [two-stage least education. This result suggests that an individual’s birth date affects educational attainment only through compulsory schooling laws, a necessary prerequisite for an appropriate instrument. This approach is criticized by Bound, Jaeger and Baker (1995) due to weak instruments and incorrect confidence interval coverage. Imbens and Rosenbaum (2005) reply offering an improved estimator for instrumental variables standard errors and deride the use of asymptotic normal approximations to this distribution. Their replication found statistically-similar results to Angrist and Krueger though with wider confidence intervals boasting correct coverage. Under the assumption that quarter-of-birth does not affect imprisonment other than via education, we apply this instrument to the Lochner and Moretti data. The Bound, Jaeger and Baker (1995) critique is less of a concern here, since we are not confronted by a weak instruments issue (see Appendix 3.B).
squares] estimates of the monetary return to education... This evidence casts doubt on the importance of omitted variables bias in OLS estimates of the return to education, at least for years of schooling around the compulsory schooling level.

This quote provides a more balanced statement of the results, namely that equality of IV and OLS estimates does not necessarily rule out endogeneity bias, but it does not preclude the possibility that the bias is significant.

### 3.6 An Empirical Illustration of LATE

Clearly, an individual’s choice of educational attainment is based upon unobservable characteristics; an instrumental variables approach attempts to alleviate this endogeneity problem. We are able to identify three different LATEs because we employ three different sets of instruments.\(^{13}\) This trio helps to illustrate empirically the difference between the LATE and the ATE.

While \[\text{Lochner and Moretti}\] only use compulsory schooling laws as an instrument, we augment their data set by adding quarter of birth in order to implement the approach of \[\text{Angrist and Krueger}\] as well. \[\text{Acemoglu and Angrist}\] use these instruments along with minimum employment age laws and we employ the same trio.\(^{14}\) Since the LATE is instrument-dependent, we do not anticipate the parameter estimates being the same across the instruments because we are estimating the parameters over different subpopulations in each instance. The use of several instruments underscores that the LATE and ATE are different parameters.

#### 3.6.1 IV estimates

The education literature considered here uses U.S. Census data on black and white men from 1960, 1970, and 1980. Note that, while these files identify incarcerated respondents, they do not provide information on the crimes committed. Additionally, the Census data only identify individuals in prison at the time of the Census and not individuals that have ever been to prison or have committed crimes. To use incarceration as a proxy for crime per se, \[\text{Lochner and Moretti}\] assume that education does not affect the probabilities of arrest or incarceration or sentence length.\(^{15}\) Since younger men are more likely to be incarcerated than older men at a given time, there is a significant age trend in these data \[\text{Gibbons, Suárez Serrato and Urbancic}\] (2009). Our approach employs a base set of explanatory variables in

\(^{13}\) It is possible, though perhaps unlikely, that all three instruments are identified by the same unobservable subpopulation, thus providing estimates of the same LATE.

\(^{14}\) See Appendix 3.A for detailed information regarding the data that we use.

\(^{15}\) \[\text{Lochner and Moretti}\] (2004) offer a detailed examination of these limitations.
each specification. Fixed effects for age (categorized into 14 dummies spanning three-year intervals: 20–22, 23–25, . . . ), state of birth, state of residence, and year are included.\footnote{These specifications correspond to those used in \cite{Lochner2002} but are broadly consistent with the approach of \cite{Acemoglu2004} as well. We use the same approach for studying both wages and crime to unite these strands and to illustrate the desired methodological points, rather than to perform strict replications. Since we add additional instruments to our analysis, we do not use the data provided openly by either pair of authors; instead, we acquire the raw data from their source. See Appendix \ref{appendix:3} for a thorough data description. It should be noted that the fixed effects used here may not generate stable parameter estimates, as discussed in \cite{Gibbons2009}.}

The results of the IV regressions (plus the OLS baseline) are given in Table \ref{table:3.1}. For the incarceration regressions, the parameter estimates are the change in percentage points of an individual’s propensity to be incarcerated at the time of the Census for each additional year of education. In the wage regressions, the estimates are the change in an individual’s log weekly wage for an additional year of education; put another way, it is the proportion increase in an individual’s wage per year of schooling. In the incarceration regressions for whites, we find that the LATE for the compulsory schooling age instrument is similar to the ATE given by OLS. But examination of additional instruments shows that the LATE differs from the ATE. Based upon the arguments of Section \ref{section:3.2}, it is impossible to conclude that “endogeneity bias is not quantitatively important,” as \cite{Lochner2002} do. The parameter estimates for whites vary by a factor of nine; one estimate for blacks is positive (though insignificant). These divergent results demonstrate that the LATE is instrument-dependent and, while one LATE estimate is near the ATE, these results suggest that endogeneity bias is an important factor in the returns to education.

The results here suggest that, for both blacks and whites, there is a subpopulation for which education does reduce criminal propensity, though there is also a subpopulation for which this effect is insignificant. All the instruments in the wage regressions yield significant, positive effects of education on wages, suggesting that there is a subpopulation for both blacks and whites for which education increases their wages. Even for these regressions that produce unanimous verdicts on significance and direction, we cannot claim that this result holds for all subpopulations, since we do not know which individuals are compliers in any of these regressions.

In Section \ref{section:3.4} we discuss the application of the Hausman test to compare IV estimates to those of OLS. To recall, if we can reject the hypothesis that the OLS estimate equals the IV estimate, then we can reject exogeneity of educational attainment. But, if we cannot reject the null hypothesis of equality, we nonetheless are not able accept exogeneity. We may find the results suggestive, especially when this equality holds for several instruments with respect to the OLS estimate. The Hausman test statistics (HS) are presented in Table \ref{table:3.1}. In all but one instance, we cannot reject the null hypothesis of equality for the incarceration regressions. This might lead us to believe that education is exogenous to incarceration, though we cannot provide a level of statistical significance for this assertion. Contrarily,
Table 3.1: IV estimates

(a) Effect of education on imprisonment

<table>
<thead>
<tr>
<th></th>
<th>Whites</th>
<th></th>
<th>Blacks</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>HS</td>
<td>Estimate</td>
<td>HS</td>
</tr>
<tr>
<td>OLS estimates</td>
<td>-0.095**</td>
<td>(0.002)</td>
<td>-0.364**</td>
<td>(0.014)</td>
</tr>
<tr>
<td>Compulsory schooling age</td>
<td>-0.087*</td>
<td>(0.044)</td>
<td>-0.589**</td>
<td>1.53</td>
</tr>
<tr>
<td>Minimum employment age</td>
<td>-0.033</td>
<td>(0.044)</td>
<td>0.198</td>
<td>6.21*</td>
</tr>
<tr>
<td>Quarter-of-birth</td>
<td>-0.275**</td>
<td>(0.100)</td>
<td>-0.064</td>
<td>0.29</td>
</tr>
<tr>
<td>All instruments</td>
<td>-0.083**</td>
<td>(0.036)</td>
<td>-0.268</td>
<td>0.40</td>
</tr>
</tbody>
</table>

(b) Effect of education on wages

<table>
<thead>
<tr>
<th></th>
<th>Whites</th>
<th></th>
<th>Blacks</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>HS</td>
<td>Estimate</td>
<td>HS</td>
</tr>
<tr>
<td>OLS estimates</td>
<td>0.052**</td>
<td>(0.001)</td>
<td>0.053**</td>
<td>(0.001)</td>
</tr>
<tr>
<td>Compulsory schooling age</td>
<td>0.139**</td>
<td>(0.014)</td>
<td>0.135**</td>
<td>30.62**</td>
</tr>
<tr>
<td>Minimum employment age</td>
<td>0.374**</td>
<td>(0.076)</td>
<td>0.047</td>
<td>0.04</td>
</tr>
<tr>
<td>Quarter-of-birth</td>
<td>0.212**</td>
<td>(0.016)</td>
<td>0.072**</td>
<td>0.57</td>
</tr>
<tr>
<td>All instruments</td>
<td>0.160**</td>
<td>(0.013)</td>
<td>0.111**</td>
<td>22.27**</td>
</tr>
</tbody>
</table>

Notes: Parameters for the incarceration regressions are in percentage point terms. For all regressions, standard errors are clustered by state-year and appear in parentheses. Sample includes individuals 60 years old and younger. All specifications contain dummies for age category (20–22, 23–25, . . .), year, state of birth, and state of residence. Regressions for blacks include a dummy for individuals in the south who turned 14 in 1958 or later to account for the impact of Brown v. Board of Education. Hausman test statistics (HS) test for equality between OLS and IV estimates. For the parameter estimates and Hausman tests, one star indicates significance at the 5% level; two stars indicate significance at the 1% level.

we can reject equality for many of the estimates in the wage equations. This suggests that education and income are endogenously related. By applying and comparing the results of
three different instruments on these data, we are able to clearly illustrate the significance of the local aspect of the LATE.

### 3.6.2 Validity of the Three Instruments

Two major conditions must be fulfilled for an instrument to be valid. First, the instrument must be correlated with the endogenous regressor. This is known as the inclusionary restriction. Here, the instruments must change the educational attainment of some students, holding their other characteristics constant. Additionally, this relationship must be sufficiently strong to produce consistent estimates [Bound, Jaeger and Baker (1995)]. Appendix 3.B tests for weak instruments.

Second, the exclusionary restriction requires that the instrument be uncorrelated with the error term in the second stage regression. This ensures that the instrument only operates on the outcome via the endogenous variable. This, too, is necessary for the instrumental variables estimate to be consistent. For this study, the enactment and amendment of the compulsory schooling and minimum employment laws are assumed to be independent of crime and wages other than through educational attainment. But, if policy makers during the period of investigation believed the results of the literature (i.e., that more education reduces crime and increases wages) and increased compulsory schooling laws or minimum employment ages in response to increased crime or depressed wages, then the instruments would not be valid. Despite its importance, this condition is untestable.

An additional condition is required to permit the LATE interpretation of the IV estimates—the monotonicity assumption (see Footnote 1). This assumption implies that the instrument alters the decision to participate in treatment in a monotonic fashion for every individual. If, for example, as the value of the instrument increases, people on the whole are more likely to select into treatment, then every individual must be more likely to enter treatment for any increase in the instrument. In our instance, as, say, the compulsory schooling age increases, every individual must be more likely to (weakly) extend their educational career. Though this assumption is untestable, it likely holds in these situations.

### 3.7 Conclusion

In this chapter, we distinguish between the ATE that arises from a correctly-specified OLS model and the LATE generated by a correctly-specified IV model. If there are homogeneous treatment effects or no endogeneity bias, then the LATE will equal the ATE. Several authors in the applied literature, however, assert the converse; if the LATE equals the ATE, then endogeneity bias is minimal. A simple counterexample and a graphical exposition show that this logic is flawed. Further, taking advantage of recent developments uniting the structural and non-parametric literatures on the identification of treatment effects, we present
a theoretically-precise illustration. We demonstrate the relationship between the LATE and the ATE empirically by estimating the returns to education.

Based upon these results, we offer some suggestions to the practitioner:

1. If a researcher finds that the estimates from OLS and IV are similar, he cannot make any definite claims regarding the presence of endogeneity or its magnitude. If there happens to be no endogeneity, then these parameters should be equal. But, in the presence of endogeneity, the IV estimate may happen to be close to the (biased) result stemming from OLS estimation.

2. The researcher can use the Hausman test to consider equality of the parameters. This test is often marshaled as one of exogeneity; unfortunately, it cannot provide such proof. The Hausman test can produce satisfactory evidence of endogeneity when equality between the IV and OLS estimates can be rejected.

3. The issues that we discuss here do not arise if the researcher is willing to assume homogeneity in treatment responses as a function of unobserved characteristics. While this may be a reasonable assumption, it should be presented clearly and defended.

4. We were able to illustrate the “local” nature of IV estimation by employing three different instruments for educational attainment. We hope that these instruments each capture different subpopulations. Different estimates may highlight heterogeneous treatment responses under endogeneity, while similar results may make homogeneity a more plausible assumption (see, e.g., Angrist, Lavy and Schlosser (2005)).

We have identified the perils in comparing estimates that are identified by different subgroups of a sample. The purpose has not been to present new econometric results, but rather to highlight a common mistake in the applied literature. By considering the relationship between LATE and ATE, we hope to help applied researchers better understand and interpret their results.

3.A Data

Lochner and Moretti (2004) use data from the 1960–1980 Censuses in their study. Because we extend their analysis, specifically by the introduction of two additional instruments, we cannot use the version of their data that they make publicly available. Instead, we recollect the data from the Integrated Public Use Microdata Series (IPUMS) at the University of Minnesota. We add quarter of birth for all the Census observations and identify the minimum employment age requirement applicable during their youth. The compulsory schooling age and minimum employment age data are only provided from 1914 forward. Hence, Lochner

17 Compulsory schooling and minimum employment age data are found in Acemoglu and Angrist (2000).

89
and Moretti limit their observations to individuals between the ages of 20 and 60 in the 1960–1980 Censuses. Due to computational limitations, Lochner and Moretti use a 90% sample of their Census file in their paper. Since the seed of this sample was not retained, we are not able to perfectly replicate their sample in our data. Our study takes advantage of increased computing power to examine the entire sample.

We follow Lochner and Moretti in analyzing the regression models separately for blacks and whites. This division controls for the inequalities in education experienced by the two races. The regressions for blacks include a dummy variable indicating if the respondent was a black man born in the South who turned 14 in 1958 or later. This incorporates changes in education quality resulting from the Supreme Court decision Brown v. Board of Education.

3.B Testing for Weak Instruments

As an initial test for weak instruments, we regress educational attainment and wages on each set of instrument dummies and the base set of explanatory variables (i.e., perform the first stage of a two-stage instrumental variables procedure). An initial inclusionary restriction test is that the coefficients on the dummies are significant and positive. This implies that the dummies explain educational attainment and operate as expected. This is an ad hoc heuristic, however.

A more reliable test for weak instruments examines the $F$-statistic on the joint hypothesis that the coefficients on all the instruments are zero. Stock and Yogo (2002) develop a distribution for this test and our results are listed in Table 3.2. We can reject the weak instruments hypothesis for most of the instruments. The quarter-of-birth instrument overall yields the strongest relationship to education.

As an additional safeguard against troubles arising from weak instruments, we use the robust instrumental variables procedure of Mikusheva and Poi (2006). These results are presented in Table 3.3 and yield roughly the same estimates and significance as standard IV estimates and do not qualitatively contradict the results of standard IV estimation. This further suggests that our IV estimates are reliable.

---

18Subsequent Censuses do not make incarceration information publicly available, thereby preventing their use here.

19Personal communication with the author.

20We define the south as those states whose Congressmen supported the Southern Manifesto in protest of the Brown v. Board decision, namely Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Texas, and Virginia.

21Another standard definition of weak instruments is having first-stage $F$-statistics below 10; our instruments broadly pass this criterion.
Table 3.2: First-stage $F$-tests for weak instruments

(a) Incarceration regressions

<table>
<thead>
<tr>
<th></th>
<th>Whites</th>
<th>Blacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compulsory schooling age</td>
<td>53.70</td>
<td>31.94</td>
</tr>
<tr>
<td>Minimum employment age</td>
<td>51.96</td>
<td>29.50</td>
</tr>
<tr>
<td>Quarter-of-birth</td>
<td>94.20</td>
<td>33.78</td>
</tr>
<tr>
<td>All instruments combined</td>
<td>52.91</td>
<td>31.17</td>
</tr>
</tbody>
</table>

(b) Wage regressions

<table>
<thead>
<tr>
<th></th>
<th>Whites</th>
<th>Blacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compulsory schooling age</td>
<td>51.47</td>
<td>34.75</td>
</tr>
<tr>
<td>Minimum employment age</td>
<td>4.54</td>
<td>22.88</td>
</tr>
<tr>
<td>Quarter-of-birth</td>
<td>81.88</td>
<td>30.91</td>
</tr>
<tr>
<td>All instruments combined</td>
<td>41.98</td>
<td>32.95</td>
</tr>
</tbody>
</table>

Notes: [Stock and Yogo (2002)](#) provide critical values for a weak instruments test based upon the first-stage $F$-statistics. For each instrument separately, the 5% critical value is 13.91; for the joint application of the instruments, this statistic is 20.53.
### Table 3.3: Robust IV estimates

(a) Effect of education on incarceration

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Confidence interval</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WHITES</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSL</td>
<td>-0.091*</td>
<td>(-0.164, -0.019)</td>
<td>0.014</td>
</tr>
<tr>
<td>MEA</td>
<td>-0.032</td>
<td>(-0.109, 0.046)</td>
<td>0.423</td>
</tr>
<tr>
<td>QOB</td>
<td>-0.277**</td>
<td>(-0.475, -0.083)</td>
<td>0.005</td>
</tr>
<tr>
<td>All</td>
<td>-0.088**</td>
<td>(-0.148, -0.028)</td>
<td>0.004</td>
</tr>
<tr>
<td><strong>BLACKS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSL</td>
<td>-0.618**</td>
<td>(-1.011, -0.229)</td>
<td>0.002</td>
</tr>
<tr>
<td>MEA</td>
<td>0.159</td>
<td>(-0.308, 0.634)</td>
<td>0.506</td>
</tr>
<tr>
<td>QOB</td>
<td>-0.066</td>
<td>(-1.098, 0.990)</td>
<td>0.897</td>
</tr>
<tr>
<td>All</td>
<td>-0.293</td>
<td>(-0.619, 0.033)</td>
<td>0.078</td>
</tr>
</tbody>
</table>

(b) Effect of education on wages

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Confidence interval</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WHITES</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSL</td>
<td>0.143**</td>
<td>(0.135, 0.152)</td>
<td>0.000</td>
</tr>
<tr>
<td>MEA</td>
<td>0.459**</td>
<td>(0.393, 0.546)</td>
<td>0.000</td>
</tr>
<tr>
<td>QOB</td>
<td>0.348**</td>
<td>(0.301, 0.408)</td>
<td>0.000</td>
</tr>
<tr>
<td>All</td>
<td>0.184**</td>
<td>(0.175, 0.194)</td>
<td>0.000</td>
</tr>
<tr>
<td><strong>BLACKS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSL</td>
<td>0.135**</td>
<td>(0.116, 0.156)</td>
<td>0.000</td>
</tr>
<tr>
<td>MEA</td>
<td>0.047**</td>
<td>(0.015, 0.078)</td>
<td>0.000</td>
</tr>
<tr>
<td>QOB</td>
<td>0.095**</td>
<td>(0.015, 0.184)</td>
<td>0.000</td>
</tr>
<tr>
<td>All</td>
<td>0.118**</td>
<td>(0.101, 0.136)</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Notes: Parameters for the incarceration regressions are in percentage terms. Confidence intervals of 95% are given. Sample includes individuals 60 years old and younger. One star indicates significance at the 5% level; two stars indicate significance at the 1% level.


