Demand Estimation with High Dimensional Data and an Application to Amazon's Book Marketplace[1, 2]

Patrick Bajari, Guido Imbens, Ana Gomez Lemmen Meyer and Wilko Z. Schulz-Mahlendorf

March 27, 2015

DO NOT CITE OR DISTRIBUTE WITHOUT CORRESPONDING AUTHOR'S PERMISSION

Abstract

Much of the recent research in differentiated product demand estimation is based on extensions of the logistic models pioneered by McFadden (1974). However, these models do not scale effectively to the very large data sets that have recently become available in applied marketing. For example, we use daily data from Amazon's US and EU e-book and physical book marketplaces. This data has approximately 44 million titles available for purchase in the US and over 33 million titles in the EU.[3] The data also has hundreds of thousands of hedonic attributes that could be used as regressors. We propose methods to analyze demand more accurately at this larger scale by combining methods from econometrics and Machine Learning. Unlike most work in Machine Learning, we construct estimators that have a clear economic interpretation and that are identified under standard conditions used in the evaluation literature. As an application, we quantify some aspects of consumer behavior and price setting. In particular, we estimate the compensating variation to consumers attributable to the increased selection available from Amazon's online retail catalog. We find that increased selection conservatively generates $2.2 billion in increased compensating variation in the US and $2.1 billion in the EU.[4] We also demonstrate that the demand for physical books is not determined in isolation from the demand for e-books and that there are relevant dynamic features in the demand for e-books. As a consequence, very short run residual demand elasticities will be lower than longer run residual elasticities and may not fully characterize the substitutability of e-books and physical books. We estimate the substitutability of e-books and physical books and find material substitutability between e-books and physical books.

[1] Patrick Bajari is Chief Economist and Vice President for Amazon. He is also a Professor of Economics at the University of Washington and a Research Associate at the National Bureau of Economic Research. Guido Imbens is a Senior Economist for Amazon, Professor of Economics at Stanford Graduate School of Business and a Research Associate at the National Bureau of Economic Research. Ana Gomez Lemmen Meyer is an Economist at Amazon. Wilko Z. Schulz-Mahlendorf is a Senior Economics Manager at Amazon. Corresponding author: Patrick Bajari, bajari@amazon.com

[2] Disclaimer: This paper does not represent the views of the National Bureau of Economic Research, the University of Washington or Stanford University. The authors alone are responsible for all content and any errors or omissions.

[3] The amount of titles constitute distinct titles with an offer listing either by Amazon Retail or third parties through Amazon as of Nov 29, 2014

[4] Compensating variation for additional physical book selection than the one available for purchase offline, i.e. excluding the 50,000 titles with highest quantity sold in 2014.

1. Introduction

In this paper, we consider the problem of estimating demand for differentiated products when both the number of observations and the number of regressors is large. Much of the existing literature in demand estimation builds off of the seminal work of McFadden (1974) and applies discrete choice models. In these models, consumers choose a utility maximizing product from a finite set of mutually exclusive alternatives. Utility is defined as a linear index in a finite vector of product attributes. The error terms are interpreted from a logistic or related distribution and the models are estimated using likelihood, moment based or Bayesian methods. See Train (2009) and Rossi, Allenby and McCulloch (2005) for surveys.

Over the past decade, researchers in marketing and economics have had access to increasingly large data sets. In our application, we include all physical books and e-books sold on Amazon's marketplace. In November 2014, there were 44 million titles offered in the US. [5] Further, we can use additional data sources (e.g. text data and product metadata) to define hundreds of thousands of control variables. Standard discrete choice models do not easily scale, and often fail to produce credible results, when applied to choice sets with millions of alternatives and hundreds of thousands of control variables.

In a parallel development, over the past decade, there has been considerable interest in statistics and computer science in high dimensional data sets with large numbers of observations. This body of methods is sometimes referred to as Machine Learning (ML). When there are a very large number of independent variables (e.g. tens or hundreds of thousands) an obvious problem is multi-collinearity. ML has a number of methods that are designed to avoid the problem of overfitting that will normally arise with a large number of collinear variables. These methods allow for non-linearity in the data generating process and scale well to large numbers of units and regressors. See Hastie, Tibshirani, and Friedman (2009) for a survey.

To the best of our knowledge, ML methods have not yet been developed to estimate the elasticity of demand for differentiated products which also account for the classic problem of price endogeneity. While ML models have been applied extensively to study consumer behavior, these models are largely used for prediction rather than the investigation of causal effects. In many settings in marketing and economics in general, there is a premium on identifying and estimating causal effects rather than correlations.

[5] The amount of titles constitute distinct titles with an offer listing either by Amazon Retail or third parties through Amazon as of Nov 29, 2014

Working Draft – Please do not distribute without corresponding author's permission.

In this paper, we combine insights from ML and econometrics and posit a model of short run consumer demand where the data set is large in the product dimension and where there are a large number of control variables. We apply our approach to a unique data set from Amazon's book marketplace. Our estimation proceeds in two stages. In the first stage, we estimate a separate demand elasticity for each of the millions of product in our data set. Our identification strategy, as described in the text, is based on a regression discontinuity design (RDD). See Imbens and Lemiux (2008), Lee and Lemieux (), and VanDerKlaauw () for recent surveys and Cook (2008) for historical backgroun. Our approach to estimation is similar in spirit to Black (1999). We exploit the fact that prices follow a classic "sawtooth" pattern seen in other retail price series. For each unique product, we run a separate regression where the control variables include product/time fixed effects and price. Since fixed effects regression is a "within" estimator, only the price variation within a given product/time window identifies the price coefficient. If the product/time windows are based on a sufficiently tight time window, the estimator essentially compares demand just before and just after the price change. This is similar to studying the discontinuity in an RDD. At high frequency, the proximate cause for the price changes is typically a decision rule in an automated system. Even if the decision rule is based on demand predictions, at very high frequencies this is a pure supply generated price change and allows us to trace out the demand curve. The regressions that we describe are obviously parallelizable which allows them to scale well in the number of products. We use a computational approach known as Map Reduce in order to process the large number of regressions.

The first stage yields a unique demand elasticity for each product in our data set. The basic theory of regression discontinuity designs tells us that these estimates will be approximately unbiased for the short run causal effects of prices on demand. However, in practice, they will have high variance for many individual products because there is limited information in small windows around price changes.

In the second stage, we reduce the variance of our elasticity estimates by pooling information across products and modeling the elasticities as a function of a potentially very large set of hedonic product attributes. We begin by building a data set of hedonic attributes for books. These attributes include the category and subcategory of each book. We use text processing algorithms and attributes derived from "unstructured" text such as product descriptions and reviews. We also use data from Amazon's recommendation systems, that is, the widget in the webpage for a given product that displays books "Consumers also viewed." We apply clustering algorithms to this data to find groups of similar books.

With the large set of product attributes we then apply tree models (e.g. regression trees, random forests, and boosting). See Breiman and Friedman (1984) and Hastie, Tibshirani, and Friedman (2011) for

Working Draft – Please do not distribute without corresponding author's permission.

a detailed discussions. The dependent variable in these models is a product level elasticity and the independent variables are the hedonic product attributes. We estimate or train these models using cross validation procedures standard in the machine learning literature. An advantage of these models is that they avoid problems with multi-collinearity and scale well to large data sets. Using these models, we obtain more precise product level elasticities compared to the original estimates based on the product-level data alone. Obviously, the marginal effects on elasticities associated with an individual product attribute (e.g., average review rating) do not have a causal interpretation. We use the model instead to predict an product level causal parameter and to reduce variance in our estimates. For the purposes of our application, we do not need the causal effects associated with an individual product attribute. Rather, we only need a predicted value of an product level residual demand elasticity.

We then use our estimated demand elasticities to address a question of economic interest – in the current discussion we seek to quantify the value to consumers associated with massive product selection. Specifically, we estimate the compensating variation attributable to unique selection from Amazon's online marketplace. The method we apply follows Brynjolfsson, Hu and Smith (2003) and Hausman and Leonard (2002). Online retailers such as Amazon are able to offer tens of millions of unique titles to consumers. By comparison, the largest physical retailers offer between 40,000 and 100,000 unique titles.[6] In the case of Amazon, the increase in choice is attributable to at least three causes. First, online retailers use large warehouses in relatively cheap low density locations to hold books in inventory as opposed to stores located in dense, urban, and therefore expensive locations. Search and recommendation systems allow customers to discover titles, augmenting standard manners of browsing for products. Second, many online retailers such as Amazon allow third party sellers to list products on their site. Low barriers to entry allow many small merchants and individuals to list their product to a national and increasingly international customer base. One million and 700 thousand merchants had an offer listing of a physical book through Amazon in November 2014 in the US and EU, respectively. Third, in the e-book space, Amazon recently created the Kindle Direct Publishing platform which allows authors to directly sell e-books to customers. This innovation lowers barriers to entry for authors and also results in increased selection.

Next, we consider the case where demand for e-books and demand for physical books are not determined in isolation. Both the empirical and methods literatures are concerned with the estimation

[6] Brynjolfsson, Hu, Smith (2003), Table 1.

Working Draft – Please do not distribute without corresponding author's permission.

of cross price elasticities as measures of substitutability of products.[7] When there are tens of millions of unique choices, however, standard approaches to estimating cross elasticities will not scale. In the case of our application to books the decision to purchase an e-book, as opposed to a physical book, is also related to the purchase of a tablet or e-reader and involves adopting a new technology. This implies that a necessary condition to understand substitutability requires a dynamic model, unlike the static models commonly used in demand estimation.

We propose an approach that combines aspects of propensity score estimation and synthetic controls with concepts from the empirical demand estimation literature. The underlying data is a panel of randomly sampled consumers.[8] The dependent variable in the model is the number of physical or e-books purchased. The control variables in the model are lagged dependent variables such as prior purchases or participation in programs such as Amazon Prime. We estimate a diversion ratio between e-books and physical books by casting this as a problem of treatment effects.[9] The treatment of interest is the purchase of an e-book and the outcome of interest is purchases in a future period, ranging from a month to a year, of physical or electronic book purchases. We begin by estimating a model of the propensity to purchase an e-book. We split our data into a finite number of propensity bins (e.g. 10). For each propensity bin, we build a flexible regression model of the predicted number of e-books on a control group of customers that did not purchase an e-book in the time period of interest. Using the estimated coefficients from this regression, we then predict the number of books individuals in the treated group of e-book purchasers, had been expected to purcahse, in the counterfactual world where they had not purchased an e-book. The mean value of the residuals for the treated group is our estimate of the average treatment effect or buying an e-book. In this manner, we can estimate the causal impact of purchasing an e-book on physical books. If the standard selection on observable assumption is satisfied (e.g., Rosenbaum and Rubin, 1983, and Imbens and Rubin, 2015), this is an estimate of the diversion ratio between physical and e-books. Our estimates imply that the decision to purchase an e-book is related to the adoption of a new technology for reading books such as purchasing

[7] See the literature on differentiated demand literature, not limited to but certainly including Berry, Levinsohn, and Pakes (1995).

[8] Note all customer data is anonymized and no identifiable information regarding customers is used in this study. In our analysis, we aggregate up any customer level ordered units and product revenue to blocks of 100 customers. This aggregation may slightly impact precision, but maintains the statistical unbiasedness of our estimates and allows us to preserve strict anonymity of customer data.

[9] See Farrell and Shapiro (2010) for a thorough treatment on the concept of the diversion ratio, which in our case measures the amount of physical book units traded for e-book units as some customers substitute towards consumption of e-books.

Working Draft – Please do not distribute without corresponding author's permission.

a device.[10] Also, we find that the diversion ratio between physical books and e-books is large when viewed over a month or longer time horizon.

The remainder of the paper is organized as follows. Section two introduces our approach for short-run demand estimation in high dimensions, including subsections on the description of data to be used, identification, and estimation approach. The third section provides our compensating variation approach and estimates. The fourth section provides a discussion of our approach to measuring and estimates of e-book and physical book substitutability. The fifth and final section concludes.

2. Short-Run Demand Estimation in High Dimensions

2.1. Data Summary and Institutional Detail

We begin by building a comprehensive panel of all physical books and e-books sold on Amazon between January 1st, 2010 and December 15th, 2014. A unit of observation in our data is an ASIN/format/seller/day.[11] The term ASIN is as acronym for Amazon standard identification number and is a label that identifies a unique product in Amazon's catalog.[12] A format is an indicator for a new physical book, a used physical book or an e-book. The seller variable can take on the values Amazon Retail, Fulfillment by Amazon (FBA) or Merchant Fulfillment Network (MFN). Amazon allows third party sellers to list titles on Amazon's marketplace for a fee. An FBA offer is one in which the inventory is warehoused in an Amazon fulfillment center. Such offers are eligible for Amazon Prime and benefit from Amazon's investment in physical capital. An MFN offer is distinguished by the fact that sellers ship the package themselves. Note that not all books are available on all publication dates, e.g., a customer cannot receive an order for a book prior to publication. As a result, we do not include observations where the book is unavailable as our panel would not be balanced.

In Table 1 we present some summary statistics about the number of ASINs with a buyable offer as of November 29, 2014. A buyable offer is defined exactly as the term would suggest – the book is available for purchase either from Amazon or one of its third party sellers. According to Brynjolfsson, Hu, Smith (2003), the largest offline retailers have at most 40,000-100,000 unique titles available for sale.[13] This means that the selection of physical books from Amazon is 440 times larger in the US and at least 330

[10] Note, consumers can access e-books through Amazon's digital book marketplace without purchasing an Amazon e-reader, for instance by installing the Amazon Kindle application on their PC/Mac, smartphone, or tablet.
[11] Note, the underlying data are time-stamped at the clock time of transaction. We use day and hour interchangeably in some of the ensuing discussion.
[12] Note, ASINs are also language-specific, e.g. Dan Brown's *Da Vinci Code* in German has a distinct ASIN as does the same book in English. On the other hand, the e-book and physical book share an ASIN code.
[13] Brynjolfsson, Hu, Smith (2003), Table 1.

Working Draft – Please do not distribute without corresponding author's permission.

times larger in Europe.[14] Note that a sizable fraction of this selection is in local languages. For example, at least four million titles are offered in French in France. The books with missing language attributes in the US and UK are likely to be in English, adding to the local language titles there, but in Germany and France a substantial proportion of those books may be in languages other than the local language.

Table 1. Selection: Offers of Distinct Physical Books

	Distinct physical books offered	YoY growth rate	Distinct physical books of local language	Distinct physical books with missing language attribute
US	44 M	9.0%	20 M	15.5 M
UK	37.5 M	13.2%	18 M	11 M
Germany	36 M	9.1%	6 M	7.5 M
France	33 M	13.3%	4 M	5.5 M

Notes: ASINs with a buyable offer as of November 29, 2014.

Much of the selection comes from third-party sellers and Kindle Direct Publishing. By aggregating the offerings of a large number of sellers and authors publishing original content, the selection available on Amazon's marketplace is increased. Given the ability to list products nationally (and increasingly internationally), the likelihood of a match between a customer and book in the long tail of selection is increased. Table 2 lists the total number of merchants with an offer on the Amazon book platform.[15] Amazon has a process that makes listing offers simple for sellers and therefore has a lower entry cost than physical retail channels. We find that there are 1.1 million unique merchants in the US and 112-321K unique merchants in the three countries in Europe for which we analyze data here. For many of the books there is a large number of merchants offering the item. On average there are 19 offers per ASIN in the US and at least 8 offers per ASIN in Europe for new physical books with a sale in November 2014, and as many as 33 on average in Germany.

Table 2. Number of Merchants and Mean Number of Offers

[14] This calculation is conditional on the 11/29/2014 data snapshot of ASINs on offer.
[15] Note, we define as a merchant an individual or firm who sells through Amazon as a third party (either as FBA or as MFN).

Working Draft – Please do not distribute without corresponding author's permission.

	US		UK		Germany		France	
Num. merchants with offers	1,116,579		321,463		244,665		112,815	
	US		**UK**		**Germany**		**France**	
Physical book format	New	Used	New	Used	New	Used	New	Used
Mean num. of offers per distinct ASIN with a sale	19	29	17	16	33	9	8	9
Mean num. of MFN offers per distinct ASIN with a sale	18	27	16	15	32	9	7	9

Notes: Numbers as of November 2014. We define as a merchant an individual or firm who sells through Amazon as a third party (FBA or MFN).

The large number of sellers allows Amazon to provide additional unique selection. Table 3 shows that more than 30 percent of the new books offered through Amazon were only offered by a third party. This number is almost 100% for used books, which are rarely offered through Amazon.

Table 3. Unique Selection: Offered Physical Books (percentage)

	US		**UK**		**Germany**		**France**	
Offered by:	**New**	**Used**	**New**	**Used**	**New**	**Used**	**New**	**Used**
Only by Amazon	13	0	15	0	17	0	16	0
Only by 3P	47	100	39	100	32	99	31	99
Both	40	0	45	0	51	1	53	1
Total	100	100	100	100	100	100	100	100

Notes: Numbers as of Nov 29, 2014.

In Table 4, we present data on the selection available through Kindle Direct Publishing. KDP lowers the cost to authors of publishing a book. The table below shows that KDP significantly increases the amount of unique selection available through e-books. The number of titles available through KDP exceeds the number of titles available from other publishers. We note that KDP also increases the selection available in local languages. In both Germany and France, KDP represents a significant share of the total number of titles. By lowering the costs to publish a book, selection is expanded and more authors are able to distribute their books to readers.

Table 4. Selection: Offers of Distinct e-books

Working Draft – Please do not distribute without corresponding author's permission.

	Total e-books	KDP	Non-KDP	KDP-Local language	Non-KDP Local Language
US	3.5 M	2 M	1.5 M	1.5 M	1 M
UK	3 M	2 M	1 M	1.5 M	775 K
Germany	3 M	2 M	1 M	90 K	200 K
France	3 M	2 M	1 M	50 K	125 K

Notes: Numbers as of Nov 30, 2014 for US and December 18, 2014 for EU countries.

2.2. Hedonic Data Attributes

In addition to quantity, price, format, seller, and language, our data contains a large number of hedonic attributes that can be used to capture differing elasticities across books. These attributes include:

- **Category/subcategory/browse node:** The browse node corresponds to the menus at the left hand side of a typical Amazon detail page that allows customers to navigate Amazon's website to search for products. For example, the browse node corresponding to Econometrics is books/business and money/economics/econometrics. Categories and subcategories are internal classifications of product groupings.
- **Unstructured text:** We can extract keywords from the detail pages (the webpages associated with a unique book title) such as the product descriptions and product reviews. These extractions are done with text processing algorithms from Machine Learning (e.g. bag of words).
- **Similarities:** Amazon has a recommendation system for books. This can be seen on our detail pages under the tab "Customers also viewed." There are up to 100 recommended products for each title. This recommendation system is an internal estimate of which products are most substitutable based on a Machine Learning model trained on highly anonymized and highly obfuscated browse data.[16]

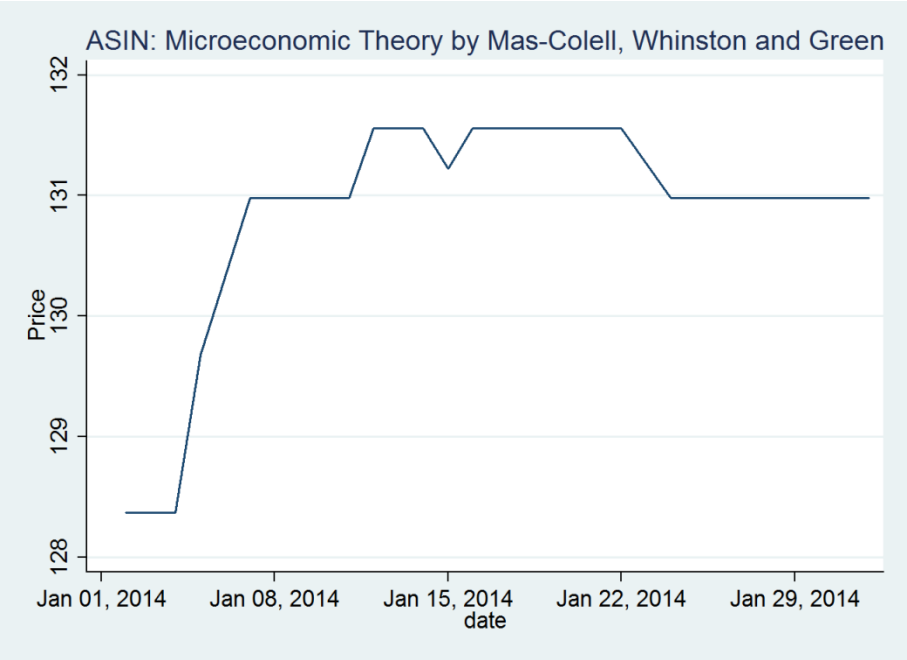
2.3. Identification

We use an identification strategy closely related to Regression Discontinuity Design (RDD). The key idea behind the identification is quite simple. It is well known that retail price series follow a "sawtooth" pattern. That is, prices tend to change in discrete time and the price changes tend to be in discrete increments. For example, the figure below plots the prices for *Microeconomic Theory* by Mas-Collel, Whinston, and Green.

[16] In no portion of the similarities algorithm process do data scientists view individually identifying customer information or customer specific browse behavior. Internal users of similarities data do not observe individually identifying customer information or customer browse behavior. The group generating similarities data also consumes highly anonymized and obfuscated data.

Working Draft – Please do not distribute without corresponding author's permission.

Figure 1. Price (US$) for Microeconomic Theory (MWG) in US Marketplace in January, 2014



Suppose that we regress quantity on price and wish to interpret the estimates as a very short run demand curve. One strategy we can use for identification is regression discontinuity. Recall that a key assumption to check is the continuity condition (see Imbens and Lemieux (2008)). Recall from Imbens and Lemieux (2008) that a discontinuity of the covariate that underlies the assignment at the discontinuity point is suggestive of violations of the "no manipulation assumption." In our application, if customers could somehow affect the price to which they are exposed, our RD design may be invalidated. At high frequency, the proximate cause for changes in prices is a decision made by an automated system. At sufficiently high frequency, it is unclear how there could be a discrete jump in demand that occurs at the same instant that a price change occurs. Therefore, the continuity condition would be satisfied, as long as the time period over which we compare demand is short enough. Also, since we are using a pure supply change, we trace out the residual demand curve.

As in all of applied econometrics, our identification conditions are an approximation and will not be perfectly satisfied. We perceive two mechanisms that could generate a violation of the continuity condition. The first is if a customer filters search results for products by price. Note that filters for search results are not available for an individual title but rather a collection of titles vended by search results. If a customer is searching for a very specific title (e.g. *Microeconomic Theory* by Mas-Collel, Whinston, and Green) we would not expect this filtering to affect demand as the filters in question

Working Draft – Please do not distribute without corresponding author's permission.

cannot be applied to specific titles. To the extent that customers are searching for a generic title (e.g. "Thesaurus" or "Dictionary"), these high frequency changes could change customer demand and hence violate the continuity condition. A second mechanism is that consumers can sort results for an individual ASIN by price on the offer listing page (e.g. if they click the link "25 New from $95.00" displayed under the title). We are collecting data to measure the frequency that these events occur. Our suspicion is that the events are somewhat uncommon and hence standard formulas for omitted variables will suggest that the bias is not large. We should also interpret our estimates for residual demand elasticities as conservative since failures of our identifying assumptions will in general generate a bias towards zero.

2.4. Estimation Approach: First Stage

In very large data sets, the standard method proposed by Hahn, Todd, and Van der Klaauw (2001) is not a practical approach to estimation, particularly if we wish to capture how the residual demand varies with hedonic attributes of the ASIN. We build on the insight of Black (1999) and use a within estimator as a practical empirical estimator based on an RDD identification strategy. However, rather than pooling observations across *i* in the panel, we run the following regression separately for each product *i:*

$$\log(q_{it}) = \alpha_{it} + x'_{it}\beta_i + \eta_i \log(p_{it}) + \varepsilon_{it}$$

The time period *t* is defined as an hour or day depending on the specification used. The coefficient α_{it} is an ASIN/time effect and x'_{it} is a vector of additional control variables (e.g. promotional events). [17] This regression looks slightly different from a conventional regression discontinuity analysis. Typically these methods are applied with a binary treatment. However, if we estimate the above regression using data from time interval narrow enough that there is only a single price change, then in the absence of the other covariates the estimated coefficient is simple the difference in average log quantities by the two values of the price, scaled by the difference in the log of the prices.

For products in the "long tail," there may not be sufficient price variation to identify the coefficient η_i. At the same time, we expect the elasticitiies for these products to potentially be different from those for products with more sales. For these products, we therefore pool observations based on the 25 closest similarities from Amazon's recommendation system. We use a two-way fixed effects model for both ASINs and time in order to identify the coefficients. This obviously relies on stronger conditions for

[17] Note, the α_{it} potentially represents a vector of time attributes. Certainly, if the underlying data are minutely, the specification would not include minutely fixed effects.

Working Draft – Please do not distribute without corresponding author's permission.

identification. However, this appears to be the most flexible specification for tail products that suffer from multi-collinearity.

2.5. Estimation Approach: Second Stage

The basic theory of regression discontintuity designs tells us that $\hat{\eta}_i$ is an unbiased estimator of the price coefficient. However, given that we are estimating a parameter for each ASIN, the variance will likely be high. In order to reduce variance, we project $\hat{\eta}_i$ onto a vector of hedonic attributes. This is similar to the modeling approach in differentiated product demand estimation where the demand parameters are functions of hedonic attributes.

Abstractly, we can write this relationship as:

$$E\hat{\eta}_i = g(z_{it}; \theta)$$

The problem we face in practice is that $z_{it}$is a high dimensional vector, on the order of thousands of attributes. In order to fit this relationship, we estimate tree models and leverage Elastic Map Reduce on Amazon Web Services for scalable computation. Regression trees and their extensions, including random forests, bagging, and boosting, are flexible methods for estimating conditional expectations that allow for the presence of many predictors or covariates, possibly more than there are observations. The classic reference is the book by Breiman and Friedman (1984). A more recent discussion in the context of other machine learning methods is in Hastie, Tibshirani and Friedman (2011).

In our setting, we are interested in predicting an outcome y given some predictors x, on the basis of an independent and identically distributed random sample. The idea is to approximate the conditional expectation of y given x with a piecewise constant function, or an average of piecewise constant functions. We sequentially split the sample into subsamples so that within the subsamples the variation in the outcome is small compared to the variation between the subsamples. The sample splits are determined by one of the predictors exceeding some threshold. The algorithms search over the full set of predictors and all possible values of the threshold to find the optimal one, in terms of in-sample predictive power. The algorithm continues the sample splitting until the improvement in in-sample predictive power from any additional split no longer exceeds a threshold. The threshold is typically set optimally using out-of-sample cross-validation.

To be specific, suppose we have two predictors, x_1 and x_2 and N observations. Starting with the full sample, we can split it into two subsamples on the basis of x_j exceeding a threshold x_0. Given that split, we have two subsamples with N0 and N1 observations, and average outcomes in the two subsamples equal to $\bar{y}_0$ and $\bar{y}_1$. The in-sample goodness-of-fit measure given that split is:

Working Draft – Please do not distribute without corresponding author's permission.

$$Q(j, x_0) = \sum_{\{i:x_j<x_0\}} (y_i - \bar{y}_0)^2 + \sum_{\{i:x_j\geq x_0\}} (y_i - \bar{y}_1)^2$$

Whereas prior to the split it was:

$$Q_0 = \sum_i (y_i - \bar{y} \quad)^2$$

where $\bar{y}$ is the overall average outcome. The optimal split is the one that minimizes Q(j,x0) over the choice of predictor, j=1,2, and the choice of threshold. Given a split, the process repeats itself, now considering splitting each of the two subsamples (''leaves''), based on any of the predictors, and at any threshold. At each step the in-sample goodness-of-fit measure weakly improves by taking out the variation in the average outcomes by subsamples.

The process of splitting continues until the improvement in the objective function no longer exceeds a threshold *a*. The value of the penalty parameter *a* is determined by out-of-sample cross-validation, typically ten-fold cross-validation. The major advantages of the tree regression method include that it is flexible in the sense that it allows for higher order interactions between the predictors, that it can deal with many predictors, and that it can deal with many observations.

3. Compensating Variation

We use our estimated elasticities to measure the increase in consumer welfare attributable to increased selection from e-commerce. In order to make our results comparable to prior research, we follow the approach of Brynjolfsson, Hu, and Smith (2003) who build on Hausman and Leonard (2002). The demand function takes the following form:

$$q_{it} = A_{it} p_{it}^{\eta_i} y^{\delta}$$

Where q_{it} is the consumer demand, p_{it} is the price of book *i* at time *t*, A_{it} is a book/time specific constant that enters into demand, $\eta_i$is the price elasticity of demand for book *i*, *y* is income and δ is the income elasticity. Assuming that income effects are negligible up to a first order approximation, which seems reasonable for books, the compensating variation from the introduction of a new product is:

$$CV = -\frac{p_{it} q_{it}}{1 + \eta_i}$$

Where q_{it} is the observed quantity of quantity *i* at time *t*.

Working Draft – Please do not distribute without corresponding author's permission.

According to Brynjolfsson, Hu and Smith (2003), physical retailers have approximately 40,000 – 100,000 unique books available for purchase in the largest offline location. We do not have data that allows us to directly measure which ASINs would be available for purchase at a physical retailer. However, a conservative assumption for the purposes of our analysis is that the 50,000 ASINs on Amazon with the highest quantity sold in 2014 would be available for purchase offline.

Given that we have estimates of the residual demand elasticities and we record product level revenues we can use the formula from the previous section to estimate the compensating variation from removing the ASINs outside of the top 50,000 ASINs on Amazon from consumers' choice sets.

Table 5 shows our estimates disaggregated by country and type of seller. We find that increased selection generates $4.3 billion in increased compensating variation, $2.2 billion in the US and $2.1 billion in the UK, Germany, and France. The increase in consumer welfare is derived by selection sold not only through Amazon Retail but also through books sold by third parties (FBA and MFN).

Table 5. Compensating Variation for Physical Books ($US)

Seller	Total	US	UK	Germany	France
Amazon Retail	2,983,705,651.54	1,567,671,897.59	557,126,510	592,985,877.71	265,921,366.60
FBA	65,474,578.33	40,349,278.41	16,375,708	7,576,500.08	1,173,091.60
MFN	1,253,921,057.99	580,496,691.93	325,148,506	247,168,207.41	101,107,652.64
Total	4,303,101,287.86	2,188,517,867.93	898,650,724	847,730,585.20	368,202,110.83

Notes: Amounts originally computed in local currency and converted to dollars using the exchange rates as of Dec 30, 2014 from www.bloomberg.com.

We recognize that our estimates of the compensating variation have a number of limitations. One factor that may cause us to overestimate the compensating variation is the difference between the long and short run. If unique titles were removed from Amazon's catalog, we would expect those titles to be carried by other offline or online retailers. There is material competition in the books market. Among other things, both Google and Apple have extensive selection of e-books and platforms such as eBay, Google Shopping, and other online and physical retailers have a deep catalog of physical books. We would expect the supply of books through these channels to increase if for some reason Amazon exogenously decided to stop carrying a given title. Similarly, in the next section, we discuss dynamics evident in e-book consumption which may impact our compensating variation calculation.

Working Draft – Please do not distribute without corresponding author's permission.

Conversely, our estimates have some inherent conservatism. First, we do not measure the producer surplus from access to Amazon's marketplace by third party merchants. These merchants have access to the information technology embedded in product detail pages that facilitate product discovery and product reviews. These technologies should reduce informational asymmetries between buyers and merchants and increase efficiency as a consequence. Finally, with regard to e-books, KDP allows authors to publish their books with a low entry cost and a relatively high royalty rate by industry standards. This has allowed a more efficient match between readers and particularly emergent authors by reducing entry costs associated with the publication of a book.

4. Substitutability and the Dynamics of Demand

In the empirical literature on differentiated product demand estimation, most empirical research has focused on estimating short run elasticities using discrete choice models. It is well known, however, that long-run elasticities can differ substantially from short-run elasticities. See Aguirregabiria and Nevo (2013) for a survey of recent work on dynamic demand. In fact, the choice of whether to examine consumer behavior over a day versus a week can considerably change elasticity estimates in retail. For example, in the context of consumer packaged goods Hendel and Nevo (2006) find smaller cross-price elasticities in the short-run than in the long-run due to stockpiling behavior of customers.

The difference between short run and longer run elasticities is particularly important in the case of goods related to new technologies. For example, in the case of e-books, consumers frequently read these books on devices such as tablets or e-readers. As a result, the decision to purchase an e-book is linked to the demand for devices. Purchasing a device in turn is both a durable good purchase and an adoption of a new technology. Models of durable good consumption are typically forward looking. Also, learning how to use a device requires customers to incur what we might think of as a fixed cost to operate a new technology. Standard models of demand do not account for this dependency and hence will generate biased estimates of elasticities.

Even if consumers do not read their books on a device, there is still a sunk cost associated with changing the format used for reading. Customers may wish to compare offers from competing e-book sellers and will need to install an app for reading e-books. The presence of this sunk cost makes the demand dynamic and implies that short run elasticities may underestimate the degree of substitutability between physical books and e-books.

An additional difficulty with estimating the degree of substitutability is the dimensionality of the data. The standard measure in econometrics of the substitutability between two products is a cross-price

Working Draft – Please do not distribute without corresponding author's permission.

elasticity. However, with many millions of products, the number of cross-price elasticities would be very large and imprecisely estimated without auxiliary assumptions on the structure of substitution.

4.1. Causal Lift Model

We propose a method to estimate how purchasing an e-book changes the likelihood that a customer will purchase a physical book in subsequent time periods from a month to year.[18] This is the diversion ratio between e-books in the current time period and physical books over the next month. Also, we can apply this method to estimating how a device purchase or a download of a Kindle app changes the propensity to purchase e-books as well.[19] This will allow us to test the hypothesis that device purchases and e-book purchases may be jointly determined.

The method that we use builds off of two ideas in the treatment effects literature. The first is propensity score estimation as in Rosenbaum and Rubin (1983). The second is the method of synthetic controls, notably Abadie, Diamond, and Hainmueller (2010).

Our hypothesis test can be framed as a classic problem of identifying the treatment effect of a dummy variable on a continuous outcome. Our data is a cross section of a random set of blocks of 100 customers where a block is indexed by *i=1,…,N* in some time period t. The dependent variable q_t is the number of physical books or e-books purchased at a time period strictly greater than *t* and is an aggregate over quantities purchased by individual customers in a block. We use purchases ranging from one month to one year after *t*. We use a vector of lagged dependent and exogenous variables x_{it}. In the private sector, many firms use RFM (recency, frequency and money) as controls in lifetime value models. We have found that similar control variables are also useful in predicting future customer spending. We use lagged customer block level revenue and units for all of Amazon and by many hundreds of product lines as controls, where lags are monthly aggregates. In addition, we use participation in Amazon programs such as Prime, Amazon Student, etc. as control variables. Note that we do not use any identifiable customer information other than prior purchase history on the Amazon platform, which is itself an aggregate of customer data within a block of 100 customers.

We begin by building a propensity model as proposed by Rubin. Given the large number of control variables and observations, logistic models may have difficulty converging. Therefore, we use a linear probability model. We run the regression:

$$d_{it} = x'_{it}\beta + \varepsilon_{it}$$

[18] Our precise notion of an e-book purchase is a spike in e-book purchases, namely that customers are considered to be in the treatment group if deviations of their e-book purchases from prior average are in the top 25th percentile of such deviations. Again, by "customer," we refer to clusters of 100 obfuscated customers.

[19] The Kindle Application can be used on PCs/Macs, smartphones, and tablets and does not necessitate the purchase of an Amazon Kindle e-reader.

Working Draft – Please do not distribute without corresponding author's permission.

where d_{it} is an indicator variable for the treatment, e.g. an e-book spike or device purchase/application download. Using the fitted value of d_{it} we partition the data into 10 propensity bins *b=1,…,10* with an equal number of customer blocks in each bin. We let *b(i)* denote the bin associated with customer block *i*.

In order to estimate the average treatment effect on the treated, we use the intuition of Abadie, Diamond, and Hainmueller (2010). For each block of treated customers, we estimate the counterfactual book expenditure using a block of control group customers with a similar purchase history. In particular, for each bin *b* we estimate the following equation:

$$q_{it} = x'_{it}\beta_b + \varepsilon_{it} \text{ for } i \text{ in bin } b$$

The purchases of books exhibit a heavy tail. In order to reduce the variance of our estimates we Winsorize q_{it} at the 99th percentile. We also map the values of x'_{it} to their associated percentiles in order to avoid well known problems using regressors that may follow power laws. As is well known, trimming and Winsorizing have the advantage of reducing variance but at the cost of generating a downward bias in estimates. Hence our estimates of the treatment effect will be conservative.

We train the above regression on observations in the control group separately for each propensity bin *b*. Using the fitted value for the coefficients $\hat{\beta}_b$, for each customer block *i* in the treated group we predict the counterfactual number of books as:

$$\hat{q}_{it} = x'_{it}\hat{\beta}_b$$

Our estimate of the average change in demand attributable to treatment is then:

$$\frac{1}{I}\sum_{i=1}^{I} q_{it} - \hat{q}_{it}$$

The identification of the model depends crucially on whether omitted variables are correlated with the treatment. In practice, we attempt to gain some insight into the plausibility of this assumption by testing for pre-trends.

4.2. Data and Results

We constructed a dataset consisting of a 50 percent random sample of all Amazon customers who purchased a physical book or e-book through the Amazon.com website during May 2012.[20] To facilitate meaningful comparisons between different consumer groups, we linked future purchases to previous

[20] Or in the case of EU, Amazon.co.uk, Amazon.de, or Amazon.fr.

Working Draft – Please do not distribute without corresponding author's permission.

characteristics for these buyers, before and after the May 2012 date on which they purchased a book (referred to as "the event date" below).[21]

The breadth and depth of this dataset is uncommon in academic and non-academic work. This information allowed us to confidently match groups of new Kindle users to similar groups of non-Kindle users, eliminating confounding influences in the comparisons below. Note, in the ensuing discussion, when we refer to "Kindle users," we are referring to Amazon.com customers who have either purchased an Amazon e-reader or customers who have installed an Amazon Kindle application on their PC/Mac, smartphone, or tablet. Customers can obtain Amazon.com e-books without purchasing an Amazon.com e-reader or device.

US Estimates

Simple Statistics

Before presenting regression results, Table 6 shows simple purchase comparisons between the groups described above in log scale. Although suggestive, these results reflect correlations only, since other confounding factors may affect purchase differences. We address these confounding factors with our regression analysis in the next section.

In Table 6, we see that new Kindle users purchased 0.66 fewer physical books in log scale than non-Kindle users in the 30 days after their event date. Differences over longer time periods are similar. Due to the log transformation, we can interpret these differences as approximate percent changes.[22] In other words, the analysis says that new Kindle users purchased around 66 percent fewer physical books over the first 30 days. Differences are even larger for longer time periods and for revenue. The bottom half of the table shows that new Kindle users purchase substantially more e-books than non-Kindle users. The difference for e-books is not surprising, but the difference for physical books suggests that consumers are indeed substituting substantially between e-books and physical books by diminishing their physical book consumption when they increase their e-book consumption.

Table 6. Differences in Physical Books and e-books
Purchased between New Kindle and Non-Kindle Users in Log Scale

[21] Customer level data is aggregated to blocks of 100 customers to preserve anonymity.
[22] Log scale differences overstate percent changes as the changes become larger. However, this is the standard way to quantify log scale differences, and the qualitative conclusion is not affected.

Working Draft – Please do not distribute without corresponding author's permission.

	Days after Kindle Engagement Window					Days after Kindle Engagement Window				
	30	90	180	270	365	30	90	180	270	365
	PHYSICAL BOOKS - REVENUE					*PHYSICAL BOOKS - UNITS*				
Kindle vs. Non-Kindle	-1.84	-1.90	-1.93	-1.92	-1.93	-0.66	-0.75	-0.83	-0.89	-0.92
95% Confidence Interval	[-1.84 to -1.83]	[-1.91 to -1.89]	[-1.94 to -1.92]	[-1.94 to -1.91]	[-1.94 to -1.92]	[-0.66 to -0.66]	[-0.75 to -0.75]	[-0.84 to -0.83]	[-0.89 to -0.88]	[-0.93 to -0.92]
	E-BOOKS - REVENUE					*E-BOOKS - UNITS*				
Kindle vs. Non-Kindle	1.17	1.40	1.53	1.54	1.57	0.96	1.16	1.27	1.29	1.32
95% Confidence Interval	[1.16 to 1.18]	[1.39 to 1.41]	[1.52 to 1.54]	[1.53 to 1.55]	[1.56 to 1.58]	[0.96 to 0.97]	[1.15 to 1.17]	[1.26 to 1.28]	[1.28 to 1.30]	[1.31 to 1.33]

Note: This table compares purchasing behavior between Kindle and non-Kindle users over various lengths of time following the Kindle engagement window. The point estimates represent the natural log of physical book and e-book revenue and units purchased by new Kindle users minus the same measures for non-Kindle users.

Table 7 shows the same comparison between consumers with an e-book purchasing spike and consumers without an e-book purchasing spike, in log scale. Results are similar 30 and 90 days after their event day: consumers with a spike purchased fewer physical books and substantially more e-books. These differences fade and even reverse for physical books over the 270 and 365 day timeframes. The relative increase in physical book purchasing by these consumers likely reflects that they are ramping up purchasing in general, not just for e-books. These simple comparisons may reflect confounding differences such as ramping up behavior between consumers rather than pure substitution, a concern that we address with our regression analysis below. Overall, the numbers in both tables suggest that consumers substitute between physical books and e-books, even over fairly short terms (one month).

Table 7. Differences in Physical Books and e-books

Purchased between Consumers with and without an e-book Spike in Log Scale

	Days after Ebook Spike Window					Days after Ebook Spike Window				
	30	90	180	270	365	30	90	180	270	365
	PHYSICAL BOOKS - REVENUE					*PHYSICAL BOOKS - UNITS*				
Spike vs. No Spike	-0.55	-0.28	-0.06	0.07	0.13	-0.18	-0.05	0.05	0.12	0.16
95% Confidence Interval	[-0.58 to -0.53]	[-0.31 to -0.25]	[-0.09 to -0.03]	[0.05 to 0.10]	[0.11 to 0.16]	[-0.19 to -0.16]	[-0.07 to -0.04]	[0.04 to 0.07]	[0.11 to 0.14]	[0.14 to 0.18]
	E-BOOKS - REVENUE					*E-BOOKS - UNITS*				
Spike vs. No Spike	1.95	2.57	2.91	3.02	3.12	0.98	1.43	1.71	1.85	1.96
95% Confidence Interval	[1.93 to 1.97]	[2.55 to 2.59]	[2.89 to 2.92]	[3.01 to 3.04]	[3.10 to 3.14]	[0.96 to 0.99]	[1.41 to 1.44]	[1.70 to 1.73]	[1.83 to 1.86]	[1.94 to 1.98]

Note: This table compares purchasing behavior between consumers who showed a spike in e-book purchasing and those who did not over various lengths of time following the spike purchase window. The point estimates represent the natural log of physical book and e-book revenue and units purchased by those with a spike minus the same measures for those without a spike.

Regression Results

Working Draft – Please do not distribute without corresponding author's permission.

Table 8 considers the impact of a first purchase of a Kindle or installation of a Kindle application. These estimates approximate the percent difference in revenue and units purchased for new Kindle users compared with similar non-Kindle users over various time periods after each consumer's event date. For example, the -0.87 value in the top left corner of the table says that new Kindle users spent approximately 87 percent less on physical books than similar non-Kindle users over the 30 days post-event date. Similarly, the -0.32 value for units over 30 days says that new Kindle users purchased 32 percent fewer physical books than non-Kindle users over the same period. We observe consistent changes for all time frames, and the 95-percent confidence intervals for these estimates are very narrow. These results strongly suggest that recent Kindle adopters reduce physical book consumption relative to similar non-Kindle users.

Table 8. Differences in Physical Books and e-books

Purchased between Similar New Kindle and Non-Kindle Users in Log Scale

	Days after Kindle Engagement Window					Days after Kindle Engagement Window				
	30	90	180	270	365	30	90	180	270	365
	PHYSICAL BOOKS - REVENUE					*PHYSICAL BOOKS - UNITS*				
Kindle vs. Non-Kindle	-0.87	-0.84	-0.81	-0.77	-0.76	-0.32	-0.33	-0.33	-0.33	-0.33
95% Confidence Interval	[-0.88 to -0.86]	[-0.85 to -0.84]	[-0.82 to -0.80]	[-0.78 to -0.76]	[-0.77 to -0.75]	[-0.32 to -0.32]	[-0.33 to -0.32]	[-0.33 to -0.32]	[-0.33 to -0.32]	[-0.33 to -0.32]
	E-BOOKS - REVENUE					*E-BOOKS - UNITS*				
Kindle vs. Non-Kindle	0.73	0.86	0.92	0.86	0.87	0.51	0.62	0.68	0.64	0.65
95% Confidence Interval	[0.72 to 0.74]	[0.85 to 0.87]	[0.91 to 0.93]	[0.85 to 0.87]	[0.86 to 0.88]	[0.51 to 0.52]	[0.62 to 0.63]	[0.67 to 0.69]	[0.63 to 0.65]	[0.65 to 0.66]

Note: This table compares purchasing behavior between similar Kindle and non-Kindle users over various lengths of time following the Kindle engagement window, using the regression methodology outlined in the text. The point estimates reflect differences between the natural log of predicted and actual physical book and e-book revenue and units purchased by new Kindle users.

All of the estimates for e-books in Table 8 are greater than 50 percent and highly statistically significantly different from zero. These numbers indicate that new Kindle users bought more e-books than similar non-Kindle users. Based on all the evidence in Table 8, we can state with reasonable certainty that recent Kindle adopters shifted away from physical books to e-books over the next 30, 90, 180, 270, and 365 days.

Table 9 shows the same analysis, comparing consumers who demonstrated an e-book purchase spike in the 30 days prior to their event date with those who did not. Again, we see negative estimates for physical books and positive estimates for e-books, suggesting that consumers substitute from physical books to e-books, even over fairly short time frames. The differences are somewhat smaller though still

Working Draft – Please do not distribute without corresponding author's permission.

meaningful, ranging from 1 to 10 percent decreases in physical book purchases and 22 to 59 percent increases in e-book purchases.

Table 9. Differences in Physical Books and e-books

Purchased between Similar Consumers with and without an e-book Spike in Log Scale

	Days after Ebook Spike Window					Days after Ebook Spike Window				
	30	90	180	270	365	30	90	180	270	365
	PHYSICAL BOOKS - REVENUE					*PHYSICAL BOOKS - UNITS*				
Spike vs. No Spike	-0.01	-0.07	-0.08	-0.10	-0.09	-0.03	-0.05	-0.07	-0.08	-0.08
95% Confidence Interval	[-0.03 to 0.01]	[-0.09 to -0.05]	[-0.10 to -0.07]	[-0.11 to -0.08]	[-0.11 to -0.08]	[-0.04 to -0.02]	[-0.06 to -0.04]	[-0.08 to -0.06]	[-0.09 to -0.07]	[-0.09 to -0.07]
	E-BOOKS - REVENUE					*E-BOOKS - UNITS*				
Spike vs. No Spike	0.52	0.59	0.57	0.51	0.49	0.22	0.31	0.34	0.33	0.34
95% Confidence Interval	[0.50 to 0.54]	[0.57 to 0.61]	[0.55 to 0.58]	[0.49 to 0.53]	[0.47 to 0.50]	[0.21 to 0.23]	[0.30 to 0.32]	[0.33 to 0.35]	[0.32 to 0.34]	[0.32 to 0.35]

Note: This table compares purchasing behavior between similar consumers with and without an e-book purchasing spike over various lengths of time following the spike purchase window, using the regression methodology outlined in the text. The point estimates are differences between the natural log of predicted and actual physical book and e-book revenue and units purchased by consumers with a spike.

EU Estimates

Simple Statistics

Before presenting regression results, Table 10 shows simple purchase comparisons between the groups described above in log scale. Although suggestive, these results reflect correlations only, since other confounding factors may affect purchase differences. We address these confounding factors with our regression analysis in the next section.

In Table 10, we see that new Kindle users purchased 0.72 fewer physical books in log scale than non-Kindle users in the 30 days after their event date. Differences over longer time periods are similar. Due to the log transformation, we can interpret the smaller differences as approximate percent changes.[23] In other words, the analysis says that new Kindle users purchased around 72 percent fewer physical books over the first 30 days. Differences are even larger for longer time periods and for revenue. The bottom half of the table shows that new Kindle users purchase substantially more e-books than non-Kindle users. The difference for e-books is not surprising, but the difference for physical books suggests that consumers are indeed substituting between e-books and physical books by diminishing their physical book consumption when they increase their e-book consumption.

Table 10. Differences in Physical Books and e-books

[23] Log scale differences overstate percent changes as the changes become larger. However, this is the standard way to quantify log scale differences, and the qualitative conclusion is not affected.

Working Draft – Please do not distribute without corresponding author's permission.

Purchased between New Kindle and Non-Kindle Users in Log Scale, EU

	Days after Kindle Engagement Window					Days after Kindle Engagement Window				
	30	90	180	270	365	30	90	180	270	365
	PHYSICAL BOOKS - REVENUE					*PHYSICAL BOOKS - UNITS*				
Kindle vs. Non-Kindle	-1.85	-1.78	-1.68	-1.54	-1.50	-0.72	-0.77	-0.81	-0.81	-0.82
95% Confidence Interval	[-1.86 to -1.84]	[-1.79 to -1.77]	[-1.69 to -1.66]	[-1.56 to -1.53]	[-1.51 to -1.49]	[-0.72 to -0.71]	[-0.78 to -0.77]	[-0.82 to -0.80]	[-0.82 to -0.80]	[-0.83 to -0.82]
	E-BOOKS - REVENUE					*E-BOOKS - UNITS*				
Kindle vs. Non-Kindle	1.61	2.06	2.33	2.40	2.48	1.33	1.69	1.90	1.97	2.05
95% Confidence Interval	[1.60 to 1.62]	[2.05 to 2.07]	[2.32 to 2.34]	[2.39 to 2.41]	[2.46 to 2.49]	[1.32 to 1.33]	[1.68 to 1.69]	[1.90 to 1.91]	[1.97 to 1.98]	[2.04 to 2.06]

Note: This table compares purchasing behavior between Kindle and non-Kindle users over various lengths of time following the Kindle engagement window. The point estimates represent the natural log of physical book and e-book revenue and units purchased by new Kindle users minus the same measures for non-Kindle users.

Table 11 shows the same comparison between consumers with an e-book purchasing spike and consumers without an e-book purchasing spike, in log scale. Results are similar 30 and 90 days after their event day: consumers with a spike purchased fewer physical books and substantially more e-books. These differences fade for physical books, particularly over the 270 and 365 day timeframes. Overall, the numbers in both tables suggest that consumers substitute between physical books and e-books, even over fairly short terms (one month). However, these simple comparisons may reflect confounding differences between consumers rather than pure substitution, a concern that we address with our regression analysis below.

Table 11. Differences in Physical Books and e-books

Purchased between Consumers with and without an e-book Spike in Log Scale, EU

	Days after Ebook Spike Window					Days after Ebook Spike Window				
	30	90	180	270	365	30	90	180	270	365
	PHYSICAL BOOKS - REVENUE					*PHYSICAL BOOKS - UNITS*				
Spike vs. No Spike	-0.85	-0.48	-0.23	-0.05	0.01	-0.32	-0.18	-0.08	0.00	0.04
95% Confidence Interval	[-0.87 to -0.83]	[-0.51 to -0.46]	[-0.25 to -0.21]	[-0.07 to -0.03]	[-0.01 to 0.03]	[-0.33 to -0.31]	[-0.19 to -0.17]	[-0.10 to -0.07]	[-0.01 to 0.02]	[0.03 to 0.05]
	E-BOOKS - REVENUE					*E-BOOKS - UNITS*				
Spike vs. No Spike	2.10	2.86	3.30	3.47	3.60	1.29	1.87	2.24	2.41	2.54
95% Confidence Interval	[2.08 to 2.12]	[2.84 to 2.88]	[3.29 to 3.32]	[3.45 to 3.49]	[3.59 to 3.62]	[1.28 to 1.30]	[1.86 to 1.89]	[2.22 to 2.25]	[2.39 to 2.42]	[2.52 to 2.55]

Note: This table compares purchasing behavior between consumers who showed a spike in e-book purchasing and those who did not over various lengths of time following the spike purchase window. The point estimates represent the natural log of physical book and e-book revenue and units purchased by those with a spike minus the same measures for those without a spike.

Working Draft – Please do not distribute without corresponding author's permission.

Regression Results

To limit the influence of individual characteristics and other factors on our results, we used the regression procedure described above to compare purchases for similar individuals in each group. The results are presented in Tables 12 and 13 below. Table 12 shows the average deviation from expected purchases for new Kindle users. These estimates approximate the percent difference in revenue and units purchased for new Kindle users compared with similar non-Kindle users over various time periods after each consumer's event date. For example, the -0.77 value in the top left corner of the table says that new Kindle users spent approximately 77 percent less on physical books than similar non-Kindle users over the 30 days post-event date. Similarly, the -0.32 value for units over 30 days says that new Kindle users purchased 32 percent fewer physical books than non-Kindle users over the same period. We observe consistent changes for all time frames, and the 95-percent confidence intervals for these estimates are very narrow. These results strongly suggest that recent Kindle adopters reduce physical book consumption relative to similar non-Kindle users.

Table 12. Differences in Physical Books and e-books

Purchased between Similar New Kindle and Non-Kindle Users in Log Scale, EU

	Days after Kindle Engagement Window					Days after Kindle Engagement Window				
	30	90	180	270	365	30	90	180	270	365
	PHYSICAL BOOKS - REVENUE					*PHYSICAL BOOKS - UNITS*				
Kindle vs. Non-Kindle	-0.77	-0.73	-0.66	-0.59	-0.57	-0.32	-0.33	-0.32	-0.30	-0.30
95% Confidence Interval	[-0.78 to -0.76]	[-0.74 to -0.72]	[-0.67 to -0.65]	[-0.60 to -0.58]	[-0.58 to -0.56]	[-0.32 to -0.32]	[-0.33 to -0.32]	[-0.33 to -0.32]	[-0.31 to -0.30]	[-0.31 to -0.30]
	E-BOOKS - REVENUE					*E-BOOKS - UNITS*				
Kindle vs. Non-Kindle	0.76	0.94	1.02	0.93	0.92	0.66	0.80	0.87	0.76	0.77
95% Confidence Interval	[0.75 to 0.77]	[0.93 to 0.95]	[1.01 to 1.03]	[0.92 to 0.94]	[0.91 to 0.93]	[0.65 to 0.66]	[0.80 to 0.81]	[0.86 to 0.88]	[0.76 to 0.77]	[0.76 to 0.78]

Note: This table compares purchasing behavior between similar Kindle and non-Kindle users over various lengths of time following the Kindle engagement window, using the regression methodology outlined in the text. The point estimates reflect differences between the natural log of predicted and actual physical book and e-book revenue and units purchased by new Kindle users.

All of the estimates for e-books in Table 12 are greater than 60 percent and highly statistically significantly different from zero. These numbers indicate that new Kindle users bought more e-books than similar non-Kindle users. Based on all the evidence in Table 12, we can state with reasonable certainty that recent Kindle adopters shifted away from physical books to e-books over the next 30, 90, 180, 270, and 365 days.

Table 13 shows the same analysis, comparing consumers who demonstrated an e-book purchase spike in the 30 days prior to their event date with those who did not. Again, we see negative estimates

Working Draft – Please do not distribute without corresponding author's permission.

for physical books and positive estimates for e-books, suggesting that consumers substitute from physical books to e-books, even over fairly short time frames. The differences are somewhat smaller though still meaningful, ranging from 9 to 14 percent decreases in physical book purchases and 28 to 50 percent increases in e-book purchases.

Table 13. Differences in Physical Books and e-books

Purchased between Similar Consumers with and without an e-book Spike in Log Scale, EU

	Days after Ebook Spike Window					Days after Ebook Spike Window				
	30	90	180	270	365	30	90	180	270	365
	PHYSICAL BOOKS - REVENUE					*PHYSICAL BOOKS - UNITS*				
Spike vs. No Spike	-0.14	-0.14	-0.14	-0.14	-0.14	-0.09	-0.11	-0.12	-0.12	-0.13
95% Confidence Interval	[-0.16 to -0.12]	[-0.16 to -0.13]	[-0.16 to -0.13]	[-0.15 to -0.12]	[-0.15 to -0.12]	[-0.10 to -0.08]	[-0.12 to -0.10]	[-0.13 to -0.11]	[-0.13 to -0.11]	[-0.14 to -0.12]
	E-BOOKS - REVENUE					*E-BOOKS - UNITS*				
Spike vs. No Spike	0.48	0.50	0.48	0.44	0.42	0.28	0.35	0.37	0.36	0.36
95% Confidence Interval	[0.46 to 0.49]	[0.49 to 0.52]	[0.46 to 0.49]	[0.42 to 0.45]	[0.40 to 0.43]	[0.26 to 0.29]	[0.33 to 0.36]	[0.35 to 0.38]	[0.35 to 0.38]	[0.34 to 0.37]

Note: This table compares purchasing behavior between similar consumers with and without an e-book purchasing spike over various lengths of time following the spike purchase window, using the regression methodology outlined in the text. The point estimates are differences between the natural log of predicted and actual physical book and e-book revenue and units purchased by consumers with a spike.

Lastly, we note that over the past decade electronic media consumption has shown tremendous long-run growth worldwide at the expense of physical media. PricewaterhouseCoopers (2013) states that annual e-book publishing revenues have risen from around $1 billion in 2008 to over $8 billion in 2012, while physical book revenues have fallen from $100 billion to $93 million over the same period. This high-level view of the two types of goods suggests that customers are actively substituting between physical media and electronic media. Our substitutability estimates are consistent with these broad patterns.

5. Conclusion

In this paper, we address the problem of differentiated product demand estimation when both the number of products and the number of control variables is large. We posit a new estimation approach that combines elements from the treatment effects literature, demand estimation literature (empirical IO/applied marketing), and the literature on Machine Learning methods. We apply this novel approach to estimate product level residual demand elasticities using an extremely rich and entirely new data source containing information on physical book and e-book transactions on the US and EU Amazon marketplaces. We then use the estimated demand elasticities to address an economic question of interest – how to measure the value to consumers associated with the massive selection present in

Working Draft – Please do not distribute without corresponding author's permission.

many modern marketplaces (especially digital marketplaces). Our measurements suggest that the impacts are meaningful in dollar terms. We find that increased selection generates $4.3 billion in increased compensating variation, $2.2 billion in the US and $2.1 billion in the UK, Germany, and France. We do note however that in our particular application to physical and e-books, interesting dynamics in the consumption of e-books as well as substitutability between physical books and e-books require additional consideration. To address these particulars, a standard approach would be to estimate cross-price elasticities among physical and digital books. However, in our empirical setting, standard approaches to estimating cross-price elasticities are not obviously tractable. We therefore posit an approach to estimating short run and longer run substitutability among e-books and physical books. We find material substitutability between e-books and physical books, as well as the fact that the demand for e-books is determined jointly with the demand for devices such as e-readers or tablets.

References:

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller, (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association* 105, No. 490, pp 493-505.

Ackerberg, Daniel and Marc Rysman, (2005). Unobserved Product Differentiation in Discrete Choice Models: Estimating Price Elasticities and Welfare Effects. *RAND Journal of Economics* 36*,* pp 771-788*.*

Aguirregabiria, Victor and Aviv Nevo, (2013). Recent Developments in Empirical IO: Dynamic Demand and Dynamic Games. *Advances in Economics and Econometrics: Theory and Applications.Tenth World Congress of the Econometric Society.*

Berry, Steven, James Levinsohn, and Ariel Pakes, (1995). Automobile Prices in Market Equilibrium. *Econometrica: Journal of the Econometric Society,* Vol. 63, No. 4, pp. 841-890.

Breiman, Leo, Jerome Friedman, Charles J. Stone, and R.A. Olshen. *Classification and Regression Trees*. Boca Raton, FL, USA: Chapman & Hall/CRC, 1984.

Black, Sandra E, (1999). Do better schools matter? Parental valuation of elementary education. *Quarterly Journal of Economics,* Vol. 114, No. 2, pp. 577-599.

Brynjolfsson, Eric, Yu (Jeffrey) Hu, and Michael D. Smith, (2003). Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science*, Vol. 49, No. 11, pp. 1580-1596.

Working Draft – Please do not distribute without corresponding author's permission.

Caplin, Andrew and Barry Nalebuff, (1991). Aggregation and imperfect competition: On the existence of equilibrium. *Econometrica: Journal of the Econometric Society*, Vol. 59, No. 1, pp. 25-59.

Chevalier, Judith and Austan Goolsbee, (2003). Measuring prices and price competition online: Amazon. com and BarnesandNoble.com. *Quantitative Marketing and Economics*, 1, pp. 203-222.

Chevalier, Judith and, Dina Mayzlin, (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, Vol . 42, pp. 345-354.

Farrell, Joseph, and Carl Shapiro, (2010). Antitrust Evaluation of Horizontal Mergers: An Economic Alternative to Market Definition. *The B.E. Journal of Theoretical Economics*, 10(1): Article 9, pp. 1-41.

Glenn Ellison and Sara Fisher Ellison, (2014), Match Quality, Geography, and Search: The Internet Market for Used Books. *MIT Economics Working Paper*.

Gaudin, Germain and Alexander White, (2014). On the Antitrust Economics of the Electronics Book Industry. *Mimeo*.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw, (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica: Journal of the Econometric Society,* Vol. 69, No. 1, pp. 201-209.

Hastie, Trevor, Robert Tibshirani, Jerome Friedman. *The Elements of Statistical Learning*. Second Edition. New York, NY, USA: Springer, 2009.

Hausman, Jerry A., (1981). Exact consumer's surplus and deadweight loss. *American Economic Review*, 91(4), pp. 662-676.

Hausman, Jerry A., and Gregory K. Leonard, (2002). The competitive effects of a new product introduction: A case study. *The Journal of Industrial Economics* 50.3, pp. 237-263.

Hendel, Igal, and Aviv Nevo, (2006). Measuring the implications of sales and consumer inventory behavior. *Econometrica: Journal of the Econometric Society,* Vol. 74, No. 6, pp. 1637-1673.

Hu, Yu (Jeffery) and Michael D. Smith, (2013). The Impact of E-book Distribution on Print Sales: Analysis of a Natural Experiment. *Mimeo*.

Imbens, Guido W. and Thomas Lemieux, (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics* 142.2, pp. 615-635.

Li, Hui, (2014). Intertemporal Price Discrimination with Complementary Products: E-Books and E-Readers. *University of Pennsylvania working paper*.

Working Draft – Please do not distribute without corresponding author's permission.

Johnson, Justin P, (2013). The Agency Model and Wholesale Models in Electronic Content Markets. *Mimeo.*

Johnson, Justin P., (2014). The Agency Model and MFN Clauses. *Mimeo.*

Kim, Jin-Hyuk and Tin Cheuk Leung, (2013). Quantifying the Impacts of Digital Rights Management and E-Book Pricing on the E-Book Reader Market. *NET Institute Working Paper*, #13-03.

Kirkwood, John B., (2014). Collusion to Control a Powerful Customer: Amazon, E-Books, and Antitrust Policy. *Seattle University Law Review,* Vol. 38: 161, pp. 161-185.

McFadden, Daniel, "Conditional Logit Analysis of Qualitative Choice Behavior", in *Frontiers of Econometrics*, ed. By P. Zarembka, New York: Academic Press, 1974, pp. 105-142.

PricewaterhouseCoopers, (2013). "Global Entertainment and Media Outlook 2013-2017," http://www.pwc.com/gx/en/global-entertainment-media-outlook/index.jhtml.

Reimers, Imke and Joel Waldfogel, (2014). Throwing the Books at Them: Amazon's Puzzling Long Run Pricing Strategy. *Mimeo*.

Rosenbaum, Paul R., and Donald B. Rubin, (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70.1, pp. 41-55.

Rossi, Peter E., Greg M. Allenby, and Robert Edward McCulloch. *Bayesian statistics and marketing*. West Sussex, England: J. Wiley & Sons, 2005.

De los Santos, Babur and Matthijs R. Wildenbeest, (2014). E-book Pricing and Vertical Restraints. *Mimeo.*

Train, Kenneth E. *Discrete Choice Methods with Simulation*. New York, NY, USA: Cambridge University Press, 2009.

Working Draft – Please do not distribute without corresponding author's permission.