Can Product Heterogeneity Explain Violations of the “Law of One Price”?

Aaron Bodoh-Creed, Jörn Boehnke, and Brent Hickman

1 Introduction

The “Law of one Price” is a basic prediction of economics that implies that all exchanges of homogenous goods in a thick, frictionless market ought to take place at the perfectly competitive price. While in some markets, most notably security exchanges, the “Law of One Price” holds very well, in most consumer product markets the Law of One Price fails to describe the reality faced by everyday consumers. This sentiment is perhaps most pithily summarized by Hal Varian, who wrote “the law of one price is no law at all” (Varian [8]). Most economists view the failure of the law of one price as an opportunity to discover new and potentially important economic frictions rather than a failure of economic theory itself.

Diamond [3] proved in a simple model of consumer search that even small costs of obtaining a price quote can yield large deviations from perfect competition. This model suggests that if consumers need to go through a costly process of visiting shops to obtain price quotes or expend the effort to call various retailers to check on prices, then one ought to expect prices to sharply deviate from the perfectly competitive benchmark. While these early models did not imply price heterogeneity, later game-theoretic models of the competition between price-setting firms yielded mixed strategy equilibria involving price heterogeneity (see Baye et al. [2] for a survey).

The relevance of these “quote cost” frictions have become questionable in light of the ease of price-comparison on the internet. In particular, the rise of price aggregator websites that provide price data from a large sample of retailers might lead one to think that the cost of obtaining a price quote has been reduced to the point of economic insignificance. However, it is equally true that there remains price variation even amongst firms that provide data to price aggregation websites. This new empirical fact has in turn spawned a new set of
models studying the economics of price aggregation websites (Baye and Morgan [1], Baye et al. [2]).

On an empirical note, there is now a large literature that seeks to discover the causes of price variation by testing the predictions of the various models of price dispersion. These analyses sometimes take the form of reduced form tests of the predictions of the models. For example, Sorenson [7] shows that pharmaceutical products that necessitate repeated purchases have lower price variation, which makes sense given consumers have a strong incentive to find a low price for products that are purchased many times. Other papers estimate a structural model and attempt to tease apart the sources of price variation based on the estimates of the structural primitives (Hong and Shum [4]).

However, there is a more elemental explanation for price variation that lingers in the background of both the theoretical and the empirical literatures. It could be that while different units of a particular product appear homogeneous to the econometrician, there may be features of individual units that render them heterogeneous (e.g., damage to the packaging). It could also be that certain retailers provide unobserved services that consumers value (e.g., refunds for dissatisfied customers).

While the fact that product heterogeneity exists is not surprising, it is difficult to assess the significance of this source of price variation - unobserved heterogeneity could explain all of the price variation or none of it. To the best of our knowledge, no previous research project has been able to provide any bound on the importance of product heterogeneity for generating the price variation we observe in real world settings. At best, product heterogeneity has been treated as (part of) a residual that absorbs all price variation not otherwise explained by other forces.

Our first goal is to develop an analysis framework for assessing the importance of product heterogeneity in what one might naively view as a market for a homogeneous product. Our second goal is to argue that heterogeneity is a powerful source of price variation. Using our data and analysis framework, we can predict between 40% and 50% of the observed price variation using a high-dimensional array of product attributes analyzed using regression forest techniques.

The ideal market to assess the role of unobserved heterogeneity would be one in which (1) price variation is present and (2) the econometrician observes all or almost all of the information that the consumer sees. We focus on sales of new Amazon Kindles on eBay’s “Buy It Now” platform from October 2012 through September 2013. This platform allows sellers to list goods for sale at a fixed price, and the would-be buyers can use sophisticated search algorithms to sift through a large number of listings very quickly. By scraping the
listings directly from eBay, our data set includes almost everything buyers observe when shopping.

An individual listing consists of a standardized description of the products written and provided by eBay as well as information provided by the seller (e.g., photos, text descriptions, warranties the seller offers, etc.). Each listing also includes information about the seller’s reliability (feedback score), shipping cost, and the posted sale price. The standardized information provided by eBay does an excellent job of concisely spelling out the technical features of the Amazon Kindle as well as eBay’s definition of a “new product.” We also recorded any additional text information the seller provided about the product, which is listed following the standardized description provided by eBay. Finally, we observed the number of photos the seller posted.

The natural language text varied widely from listing to listing. Some listings were terse:

- Item is brand new. This is the latest model. This is not the HD one. Shall ship via USPS. No Returns.

Other listings were extremely verbose and (for example) repeated the information contained in eBay’s standardized description at length, offered warranty terms and conditions, advertised other goods that seller has for sale, or provided detailed descriptions of the condition of the packaging.

As a first step, we document that there is indeed significant price variation within each day of our sample. We find that the standard deviation of the price is equal to 21.2% of the mean. This ought to strike the reader as somewhat surprising given that one might think that there ought to be little heterogeneity in a market for a new electronics product, especially given eBay’s explicit insurance and the strong reputation of many of the sellers.

Prior studies have had difficulty explaining price variation in terms of underlying product characteristics. Part of our ability to predict the price variation is due to the rich set of observables we have, and part of our success is due to the powerful machine learning techniques we employ. Our second goal is to determine the extent to which each of these factors determined our success.

As a benchmark, we limit our data to a subset of explanatory variables that we think most resembles the data sets used by work studying price variation amongst brick-and-mortar stores. These include the shipping cost, whether returns are allowed, and the eBay seller score. We use this data to predict the degree of price variation using a flexible OLS model. We find that the OLS model is able to explain 12% of the price variation, which is in line with the prior literature.
We then turn to our full data set. Our biggest challenge is to process the complex natural language postings of the sellers. We use a bag-of-words approach to assess the informational content of the descriptions of the products. We chose a vector of 193 words that commonly appear in the listings and we judged as relevant to product heterogeneity. We then defined a dummy variable for each of these words that is equal to 1 if the word was used in the text of a listing and 0 otherwise. Since the dimensionality of the regressors is enormous, we used principal component analysis to reduce these 193 regressors to a set of 20 components. We also included variables that describe the number of pictures that were included in the listing, the total number of words in the listing, and the fraction of the listing that was capitalized. On top of this, we included all of the data used in the limited analysis described above.

Finally, we apply both a flexible OLS model and a random forest model to the our full data set. The OLS model is able to explain 24% of the price variation, while our random forest model explains 45% of the price variation.\footnote{We also explored using boosted trees, but found that the predictive power was essentially the same as our random forest model.} The success of the OLS estimator when using our rich set of regressors suggests that the data set we have collected (unsurprisingly) helps us detect underlying product heterogeneity. The fact that our regression forest detects almost twice as much price heterogeneity suggests that our data reflects product heterogeneity in complex, nonlinear ways. Our results are summarized in the following table:

<table>
<thead>
<tr>
<th></th>
<th>Limited Data Set</th>
<th>Full Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordinary Least Squares</td>
<td>0.12</td>
<td>0.24</td>
</tr>
<tr>
<td>Regression Forest</td>
<td>0.16</td>
<td>0.45</td>
</tr>
</tbody>
</table>

We interpret the fact that we can predict 45% of the price variation from the data we have on the listings as a lower bound on the amount of price variation that can be explained by product heterogeneity. This high degree of heterogeneity is somewhat surprising given that we are focusing on new Amazon Kindles. Many of the obvious sources of product heterogeneity are ruled out in our setting. For example, one might have thought that the products might be bundled with accessories, but this turns out to be rare in our data and, when present, the accessories are of relatively low value. One might have also assumed that issues such as seller reliability might be inducing price heterogeneity, but eBay provides a strong warranty against seller misbehavior. Moreover, seller reputation was already included in a predictor in our limited data set.

Finally, one might believe that the data we have includes indicators for other sources of cost heterogeneity. For example, the regressors could reflect differing seller reservation
values that drive price heterogeneity (Reinganum [6], MacMinn [5]). We provide a simple model of the problem facing sellers choosing a list price that suggests that unless the sellers are unrealistically impatient or very heterogeneous, heterogeneous marginal costs ought to generate very little price dispersion.

We conclude that even in markets wherein one might have expected the law of one price to hold that (1) the law fails badly and (2) underlying product heterogeneity can explain a great deal of the price variation. Given the relative homogeneity of the products we study, we suspect that product heterogeneity is at least as important in other markets. We believe our analysis is a first step in tracing out the relative importance of product heterogeneity and the other drivers of product heterogeneity explored in the prior literature.

References


